uOttawa

L'Université canadienne
Canada's university

Analysis of Healthcare Coverage using Data Mining Techniques

by

Mohammad Hossein Tekieh

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Master of Science

in

Electronic Business Technologies

Thesis directed by:

Dr. Bijan Raahemi

University of Ottawa
Ottawa, Ontario

January 2012

In the name of God,

the compassionate, the merciful

ABSTRACT

This study explores healthcare coverage disparity using a quantitative analysis on a large dataset from the United States. One of the objectives is to build supervised models including decision tree and neural network to study the efficient factors in healthcare coverage. We also discover groups of people with health coverage problems and inconsistencies by employing unsupervised modeling including K-Means clustering algorithm.

Our modeling is based on the dataset retrieved from Medical Expenditure Panel Survey with 98,175 records in the original dataset. After pre-processing the data, including binning, cleaning, dealing with missing values, and balancing, it contains 26,932 records and 23 variables. We build 50 classification models in IBM SPSS Modeler employing decision tree and neural networks. The accuracy of the models varies between 76% and 81%. The models can predict the healthcare coverage for a new sample based on its significant attributes. We demonstrate that the decision tree models provide higher accuracy that the models based on neural networks. Also, having extensively analyzed the results, we discover the most efficient factors in healthcare coverage to be: access to care, age, poverty level of family, and race/ethnicity.

ACKNOWLEDGEMENT

I would like to thank my supervisor, Dr. Bijan Raahemi, for all the support during this research study. I am also appreciative of my friends in the KDD lab which assisted me in different parts of the work.

Moreover, many thanks to my parents for encouraging and supporting me throughout my study in university, and my life.

Last but not least, I wish to thank my wife, Farideh, for her patience, understanding, and support, especially during this research study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One

1 INTRODUCTION

Healthcare coverage disparity is a critical issue since there are groups of people who do not have any appropriate health insurance coverage. Researchers believe lack of healthcare coverage will cause main problems in people's lives, such as poor health and early death [1]. United States of America – one of the most powerful and effective countries in the world – is suffering from a critical healthcare coverage disparity. In the recent presidential election in the United States which was held in year 2008, one of the main issues discussed between the candidates was healthcare coverage. About 15% of the US population didn't have healthcare coverage for even 1 day in that year [2] and this number is still increasing [3]. In the 2011 "Occupy Wall Street" movement in the US, healthcare coverage disparity is also available in people's arguments. One of the unofficial demands of this movement is that they believe the private insurers are trying to make more money from the health system, instead of providing more facilities for doctors, patients, and etc. [4]. Finding the factors with significant impact on healthcare coverage will contribute to improve the health services strategies to plan more effective and target different segments of patients appropriately.

In literature review, some efforts have been done to explore the impact of different variables on healthcare coverage by employing both statistical techniques and data mining techniques [5] [2]. But, still there are some gaps in this area that avoid achieving more operable solutions for healthcare coverage inconsistencies. One of the main gaps in the literature is obtaining a smaller set of factors which have more impacts on healthcare coverage. Another

gap is considering people's whole year healthcare coverage status as their health insurance status, not only their coverage status in the reference time.

Exploring huge amount of data and discovering knowledge in databases need powerful tools such as data mining techniques. Berry and Linoff [6] have defined data mining as "the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules". Statistical techniques try to validate hypotheses and statements using a data collection, but on the other hand, data mining techniques discover knowledge from a massive amount of data [7]. In this research, we will apply data mining techniques including both supervised and unsupervised learning on a real medical expenditure dataset to select the variables that have the most effect on healthcare coverage in order to fill the gaps we observed in the healthcare coverage literature. Therefore, we have suggested this title for our study which represents our work in a general point of view: "Analysis of healthcare coverage using data mining techniques".

1.1 Research Questions

Table 1.1 lists the observations, objectives, and research questions of this study briefly. Regarding the research questions, key concepts of this study and the sampling database are presented in continue

1.1.1 Key Concepts

The key concepts or constructs of this study regarding the research questions are as following:

1. Healthcare Coverage
2. Medical Expenditure Dataset Factors

Table 1.1: The research questions

Observations	Healthcare coverage disparity is a critical issue in the United States.
Thesis	We noticed that more generic factors (such as socio-demographic and financial factors) are usually in the top variable importance rankings.
Enthymeme	Lack of healthcare coverage in the US can get reduced using an optimal set of factors.
Problem Statement	Selecting the variables that have the most effect on healthcare coverage in order to establish more operable solutions for healthcare coverage disparity.
Objectives	Building accurate healthcare coverage prediction models based on the whole year coverage status. Discovering the efficient factors in healthcare coverage. Presenting the main groups of uninsured people in the US.
Research Questions	Which classification model has higher accuracy to predict healthcare coverage status using medical expenditure dataset factors? What is the efficient set of factors in predicting healthcare coverage? What are the main characteristics of uninsured groups of people? What is the healthcare coverage threshold to consider a person as insured?

3. Healthcare Coverage Model

Table 1.2 shows the standard, specific, and operational definitions of these key concepts. The general definition of each concept is in the standard definition column, its definition in this study's area has come in the specific definition column, and the operational definition describes how this concept is going to be measured.

1.1.2 Sampling Database

To accomplish this study, we need an appropriate dataset to build the models and analyze the

Table 1.2: Key concepts definitions

Concept / Construct	Standard definition	Specific definition	Operational definition
Healthcare Coverage	Insurance against the risk of incurring medical expenses among individuals	Legal private and public insurances against the risk of incurring medical expenses among individuals	Level of healthcare coverage
Medical Expenditure Dataset Factors	Factors recorded and reported in medical expenditure databases and statistics	Socio-demographic, health expenditure, utilization, health status, and health insurance coverage variables	Effectiveness of subset of independent medical expenditure dataset variables
Healthcare Coverage Model	Model which shows the healthcare coverage pattern	Model in which healthcare coverage classes and groups will be assigned to a sample	Accuracy rate of assigning samples to healthcare coverage classes and groups

results. We will use a real valid dataset in the healthcare area of the United States with the focus on medical costs and expenditures called the Medical Expenditure Panel Survey (MEPS). This survey will be provided each year by Agency for Healthcare Research and Quality (AHRQ) which is a subset of the US Department of Health and Human Services [8].

1.2 Methodology

Our research is based on a quantitative study on a real valid database. This database is a repeated cross-sectional sampling on the US population. Initially, we preprocess this data and prepare it for this study. Data mining approaches were employed on this data to build supervised and unsupervised models, including decision tree and neural network. We compared the accuracy of models in predicting healthcare coverage based on different

attributes. The complete explanation of the research methodology and the characteristics of this research are available in Chapter 3 of this document.

1.3 Motivation

The contribution of this study gives more detailed information about one of the most important issues of the public health, healthcare coverage, which has significant impacts on our lives. Due to the observed disparities in healthcare coverage, we believe that the attribute rankings and healthcare coverage prediction that we have done in this study can help health researchers, insurance companies, and even policy makers to focus on the uninsured people effectively. If this population can be targeted properly, the healthcare coverage disparities and inconsistencies can also be reduced. Current disappointments about the number of uninsured people in the US indicate that healthcare coverage is still a demanding issue that needs more attention and improvement. Having better health quality and maintaining it can advance the lifestyle of human beings, and rise joy and hope in people's daily life.

We build several types of data mining models in this study. Building, exploring, and analyzing these models lead us toward new results based on different types of modeling. Successful outcomes strengthen the position and reliability of data mining techniques which can motivate machine learning researchers to develop more accurate and faster algorithms. In addition, achieving good results will also strengthen some related works and arguments in the field of healthcare coverage. Note that no similar analysis exists in the literature which employs Medical Expenditure Panel Survey (MEPS) datasets to build healthcare coverage model.

1.4 Structure of the Thesis

This thesis is structured into six chapters. Literature review and related works are covered in the second chapter. In Chapter 3, the research methodology of this study is explained in details. Introducing the raw data, selecting the attributes, and preprocessing the data, are explained in Chapter 4. In Chapter 5, we describe the healthcare coverage models (employing classification and clustering algorithms), and analyze the results in details. Finally, the conclusion of this study, limitations, and future works are presented in Chapter 6.

Chapter Two

2 BACKGROUND

In this chapter, we review the most relevant literature showing the current state-of-the-art works in the field of healthcare coverage. Healthcare coverage models and applications of data mining in healthcare area in the past studies were the main demanding in this review. More than 50 related books, journal articles, conference papers, reports, documentations, and valid webpages are covered in this section trying to answer these questions:

1. What is the meaning of healthcare coverage disparity?
2. What are the recent statistics in the United States and Canada?
3. What are the differences of sources of payment?
4. What are the applications of data mining in health area?
5. Which data mining techniques were employed to build healthcare coverage models in the past studies, and what were their results?

2.1 Healthcare Coverage Disparity

In many countries, people with better health are the ones that have more funds and gained more education. On the other hand, when you don't have health coverage, you won't go for care or delay it and this will cause more problems and finally might end to early death. As a matter of fact, we should consider this assumption that better health results from more expensive care. One of the ways of measuring healthcare coverage disparity is counting the number of uninsured people in a country; people who don't have even a basic health insurance to protect them from costly medical expenditures.

As a noticeable instance, the United States is suffering from a critical healthcare coverage disparity which is getting wider. In this section, we will describe the healthcare coverage disparity of the United States with the words of statistics. In addition, the reasons and possible solutions are also expressed briefly.

2.1.1 Statistics of Healthcare Coverage in the US

To explore the current situation of healthcare coverage disparity in the United States, we should go back several decades and observe the trend of health coverage from that time to now. In 1980s, the percentage of uninsured (without coverage) people increased in the US. Since 1990, the percentage of uninsured nonelderly (under age 65 years) people has remained stable, but the number has increased to 43.3 million in 2007 [9]. This is the reason that healthcare coverage was one of the major issues in the 2008 US presidential election [2]. According to Medical Expenditure Panel Survey (MEPS) [8], in 2008, about 15% of the US population didn't have even one day healthcare coverage in that year. Also, about 15% of the US population had only partial coverage (less than 12 months coverage in a complete year), and the rest (about 70% of the US population) had full-year coverage. By March 2010, the percentage of uninsured people in overall reached to 16.3% of the whole population which means around 50 million people without coverage in the whole past year [10]. Figure 2.1 displays the trends of number of insured people and the uninsured rate from 1987 to 2010.

Private sector was successful in attracting customers between 1959 and 1968. The percentage of nonelderly (under age 65 years) people with private coverage increased to 79% in this period. It didn't change during 1980s, but since 1990, it started to drop down. Finally in 2007, the percentage of nonelderly people with private coverage in the US reached to 67% [9]. Job-based coverage constructs the major segment of the private coverage in the US. The

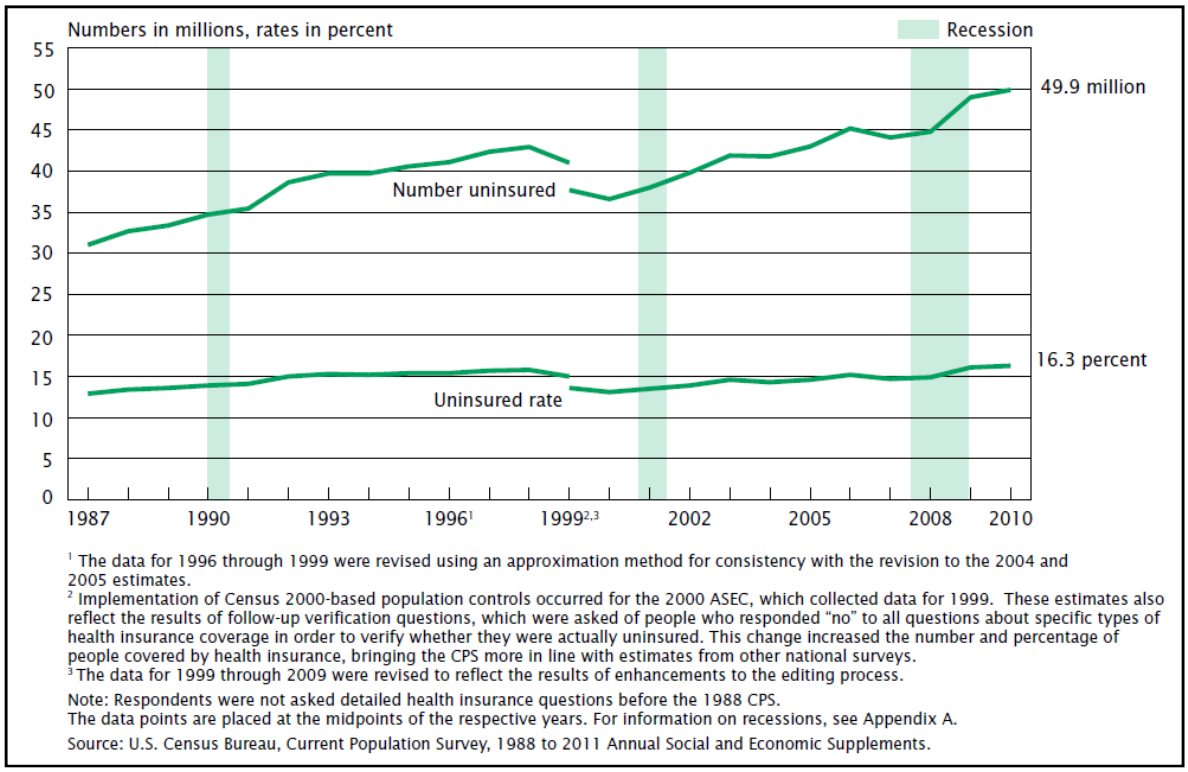


Figure 2.1: Number uninsured and uninsured rate: United States, 1987 to 2010 [10]

percentage of people with employer coverage in 2002 was 64.2%, while in 1987 was 70.1% [11]. Especially since 2000, due to increase in costs of healthcare benefits and also weak economy, the job-based health coverage has reduced [12]. Figure 2.2 draws the trends of children and adults covered by a private health plan from 1999 to 2007.

Many workers in the US will be offered a health plan with their job by their employer which will be discussed in the source of payment section. But, not all workers achieve job-based coverage in the United States, especially the ones in poor families [13]. In addition, employment coverage for the employee and his dependents is also not stable and very depends on the size of firm, type of job, position importance, and etc. Based on these parameters, the employer will offer a specific health insurance plan to the employee and even may not offer any. Figure 2.3 displays that in larger size firms which offer higher hourly

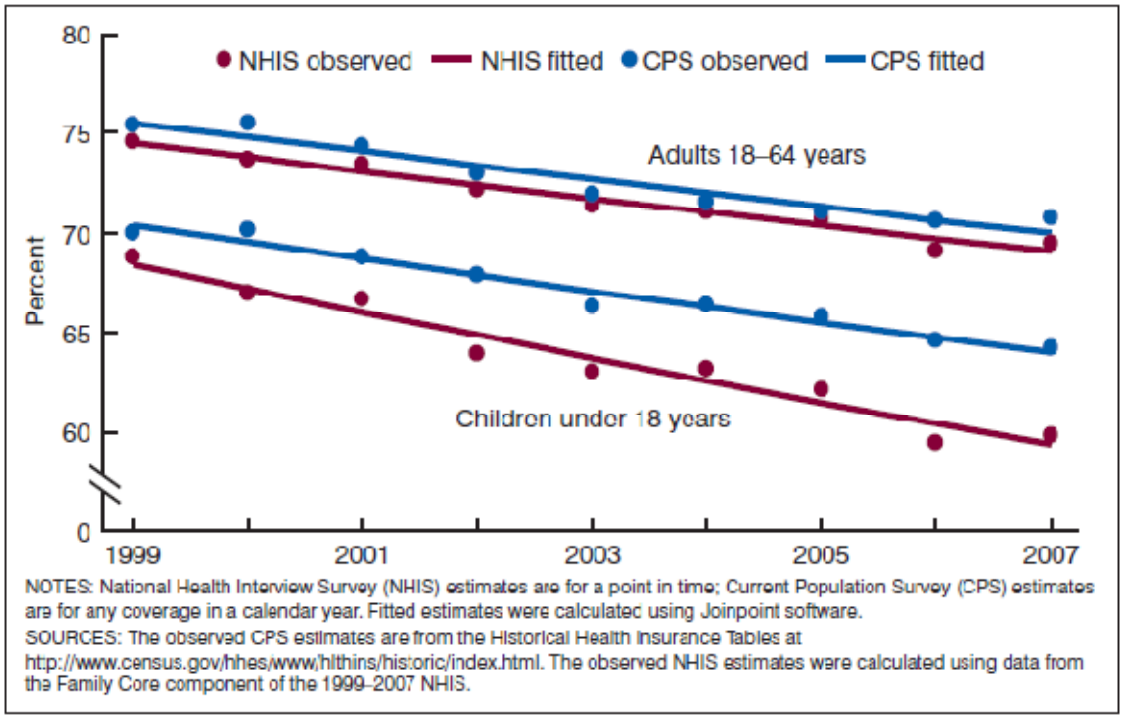
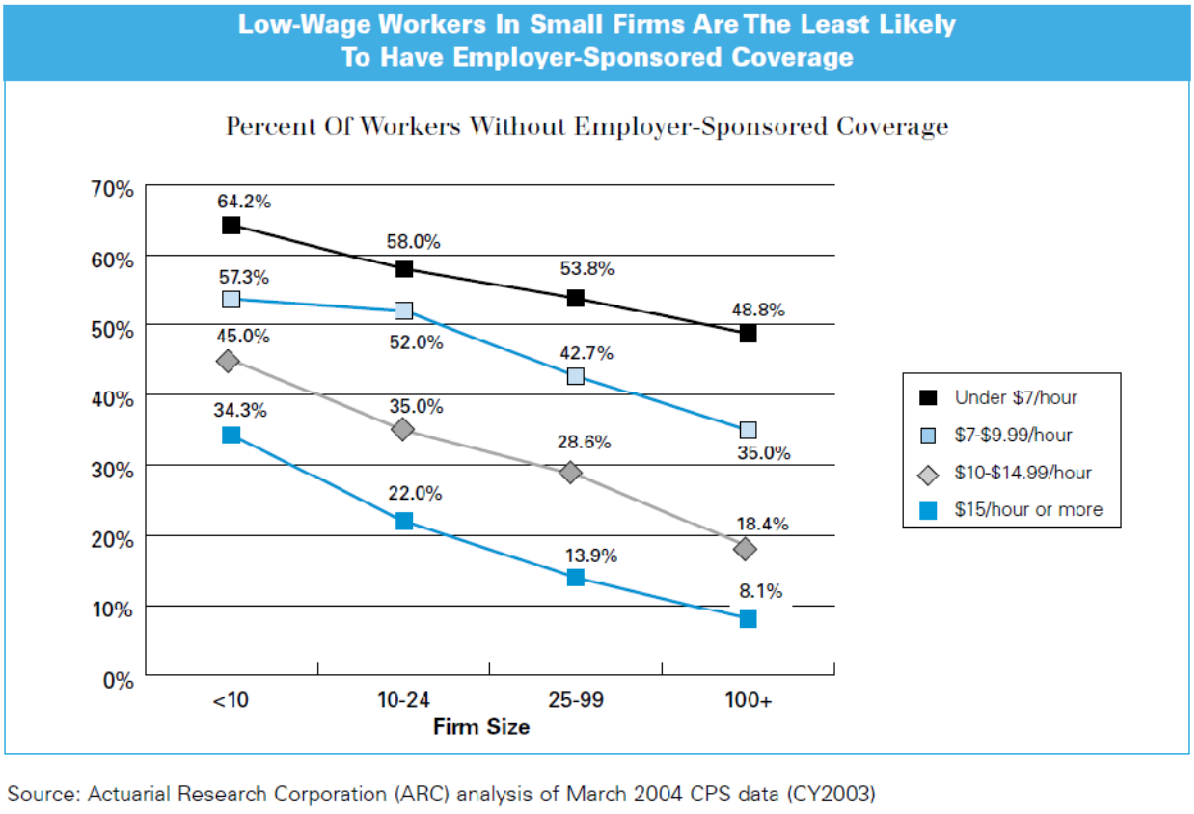


Figure 2.2: Percentage of persons under age 65 years with private coverage, by age group and survey: United States, 1999-2007 [9]



Source: Actuarial Research Corporation (ARC) analysis of March 2004 CPS data (CY2003)

Figure 2.3: Percents of workers without employer-sponsored coverage [14]

wages there are less number of uninsured workers. Note that young adults with low income usually don't get any coverage plan offer from their boss [12].

2.1.2 Universal Coverage as a Possible Solution

Lack of healthcare coverage reduces access to care and weakens health status [15] and will widen the existing gap. A solution to reduce the problem of inequalities in healthcare services is providing publicly funded healthcare to give services based on people's need, not their ability to pay [16]. However, this will cause a high load of health expenditures on the shoulder of government, but the universal coverage has been effective to reduce socioeconomic inconsistencies in health [16].

For instance, all provinces and territories of Canada provided universal coverage for hospital and physician services since 1970s. Canadian Institute for Health Information [17] has estimated the total amount of public sector health expenditures in Canada to reach C$135.1 billion in 2010 which is 70.5% of total health spending. This amount has increased by 4% and 1.4% in 2010, respectively before and after eliminating the effects of the inflation rate and population growth. In addition, forecast for public share of total health expenditure per person in 2010 was expected to be C$3,958 in Canada (total health expenditure per person is C$5,614). United States is the highest spending per person in healthcare which makes health insurance very essential. In 2008, public share of total health expenditure per person in Canada and United States was $2,863 and $3,505, respectively (total health expenditure per person in Canada and United States was $4,079 and $7,538, respectively) [17].

Figure 2.4 displays the public share of total health expenditure among 26 selected countries in 2008. Note that the data of Denmark, Japan, Turkey, and Australia is for 2007, Luxembourg and Portugal is for 2006, and Netherlands for 2002.

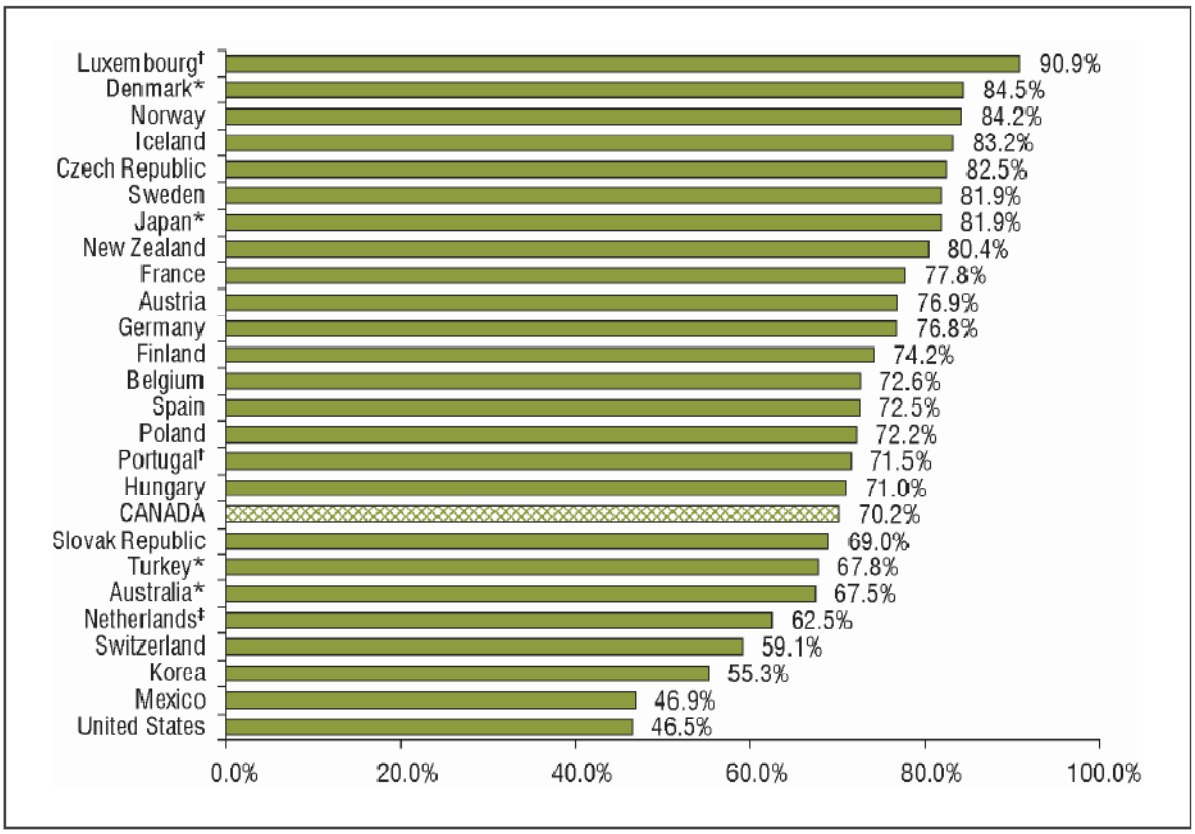


Figure 2.4: Public share of total health expenditure among 26 countries in 2008 [18]

In the United States, although still some problems exist in healthcare coverage [19], by providing Medicaid and Medicare insurances for some specific groups of the population, the disparities in healthcare utilization reduced [20]. Figure 2.5 shows the trends of Medicaid insured rate by age and survey.

Finally, it seems that Canada has performed better than the US in providing health services to all segments of people that are in need, especially lower income people. This conclusion has been achieved by comparing the physician and hospital use in Ontario (of Canada) and the US [21] [22] and also comparing cancer survival in Canadian and the US cities [23].

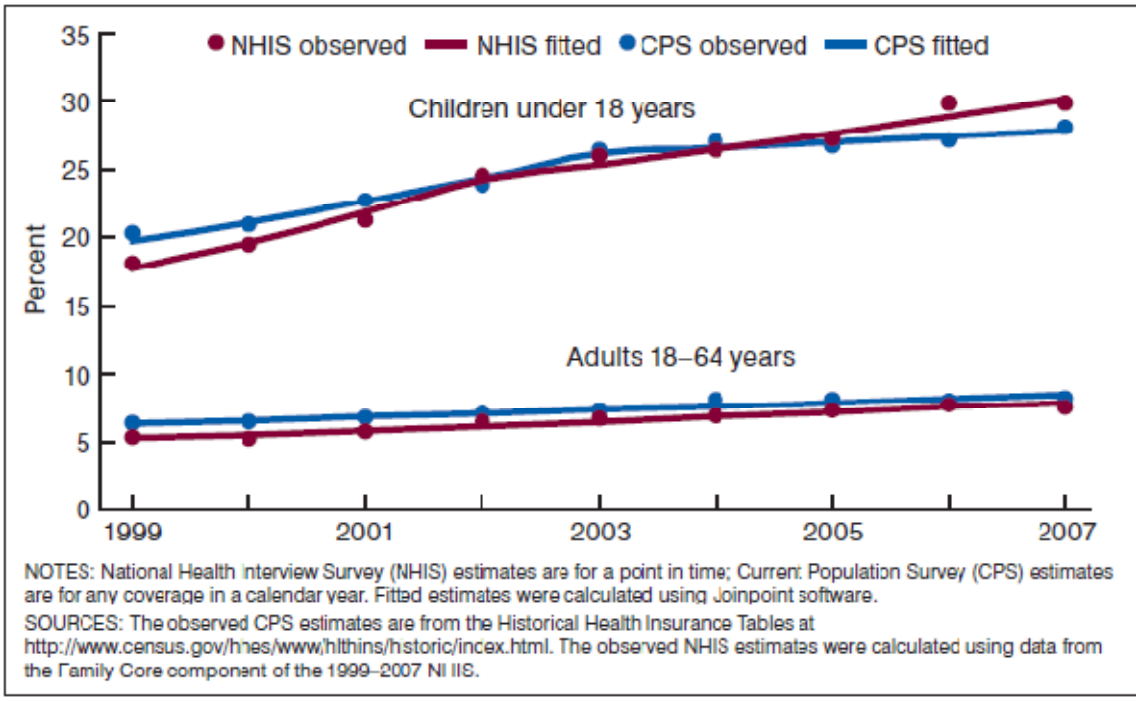


Figure 2.5: Percentage of persons under age 65 years with Medicaid coverage, by age group and survey: United States, 1999-2007 [9]

Based on Cohen *et al.*'s research [9], recent trends in coverage derived from NHIS (National Health Interview Survey) and CPS (Current Population Survey) are similar. NHIS is a continuous in-person household survey of the civilian non-institutionalized population of the US. CPS is held by the US Census Bureau for the Bureau of Labor Statistics. Information on health insurance coverage is collected by CPS as part of the Annual Social and Economic Supplement (ASEC) [24]. Although coverage trends of NHIS and CPS are similar, estimates of health insurance coverage based on the CPS are broadly cited [9]. In general, the difference in survey estimates of health coverage may be for recall period, focus of the survey, data editing, and etc. [24]. The small difference in health coverage estimates between CPS and NHIS is that CPS labels those who had no coverage for the whole calendar year as uninsured, but NHIS reports the person's status in the reference time [9].

2.2 Sources of Payment for Healthcare Coverage

In the United States, sources of payment for healthcare coverage are in two categories: public or private. The third category is the uninsured people who pay for their medical expenditures from their own pocket. Different types of health insurance coverage are included in each of these categories. In this section, we have provided the definition of each of these health plans briefly using an article [9] released by National Health Statistics Reports.

2.2.1 Private Coverage

Private coverage plans are usually divided to two types: employer-sponsored (or job-based) and individual (or direct purchase or non-group). The definition and explanation of these types are as following:

1. *Employer-sponsored:* A private insurance originally obtained through a present or former employer or union; this also includes private insurance obtained through the workplace, self-employment, or a professional association.

   Job-based coverage forms the majority segment of the private coverage in the US. Moreover, having a health insurance, in addition to the wages, in a job offer is a benefit for a worker [12]. This type of coverage is more favored in the US and seems to have some advantages compared to other popular ones.

   Glied and Borzi [12] explain the advantages of employment-based insurance compared to non-group and public insurances. Employment-based insurances are usually provided in a group health plans and therefore they are more cost-effective than non-group insurances. Also, the non-group (individual) insurances are more limited in their services than the group plan [25]. On the other hand, comparing with government insurances, employment-based insurances are more customizable based

on healthcare market, economic, or workplace factors and more easy-going to adjust the desired benefits due to not going through an annoying and public administrative process.

According to the 2003 Health Confidence Survey [26], the majority of people who were covered by an employment-based health insurance are satisfied with their coverage. Moreover, people were asked about the type of health insurance they are more likely to have between government health insurance, employment-based insurance, and individual insurance. Although a growing number selected government health insurance, but the majority selected employment-based insurance.

2. *Individual:* Directly purchased plans, as well as plans obtained through school or other means.

2.2.2 Public Coverage

Public coverage plans are usually divided to four types: CHIP (or SCHIP), Medicaid, Medicare, and Military. Data statistics show that nearly all elderly people aged above 65 years are covered by Medicare. Medicaid and CHIP together cover almost 40% of the poor and 25% of the near poor among people aged under 65 [27]. The definition and explanation of these types are as following:

1. *Children's Health Insurance Program (CHIP):* It provides additional federal matching funds for states to provide health care coverage to low-income, uninsured children aged 18 years and under who are ineligible for Medicaid. In a few states, CHIP has been expanded to cover select portions of the adult population. Within federal guidelines, each state determines the design of its CHIP program, eligibility groups, benefits packages, and payment levels for coverage.

2. *Medicaid:* It was authorized as a jointly funded cooperative venture between the federal and state governments to assist states in the provision of adequate medical care to eligible needy persons. Within broad federal guidelines, each state establishes its own eligibility standards; determines the type, amount, duration, and scope of services; sets the rate of payment for services; and administers its own program. Therefore, Medicaid is for poor parents and children, also for severe and long-term care for disabled people and the eligible elderly. There are also two programs that people in those programs are also eligible for Medicaid:
    - *Temporary Assistance for Needy Families (TANF):* This program was replaced with the Aid to Families with Dependent Children (AFDC) program in 1996. According to this law, children and parents who meet AFDC eligibility standards in effect in their state in July 1996 qualify for Medicaid. AFDC was a grant program to enable states to provide cash welfare payments for needy children who had been deprived of parental support or care because their father or mother was absent from the home, incapacitated, deceased, or unemployed.
    - *Supplemental Security Income (SSI):* It replaced former federal-state programs in the 50 states and the District of Columbia. SSI is administered by the Social Security Administration and provides income support to persons aged 65 years and older, blind or disabled adults, and blind or disabled children. Eligibility requirements and federal payment standards are nationally uniform. SSI recipients are eligible for Medicaid.
3. *Medicare:* It is a nationwide health insurance program providing health insurance protection to people aged 65 years and over, people entitled to Social Security

disability payments for 2 years or more (with limited exceptions for people with specific diagnoses), and people with end-stage renal disease, regardless of income. From its inception it has included two separate but coordinated programs: hospital insurance (Part A) and supplementary medical insurance (Part B). Medicare Advantage (previously Medicare–2007Choice) (Part C) is an expanded set of options for the delivery of health care under Medicare. The Medicare Prescription Drug, Improvement, and Modernization Act (MMA) established a voluntary drug benefit for Medicare beneficiaries and created a new Medicare Part D. In brief, Medicare is for the elderly and for disabled of any age.

4. *Military:* Includes multiple programs serving active duty personnel and families, retirees and their families, and eligible veterans. TRICARE (formerly CHAMPUS) covers active duty service members, retirees, activated Guards/Reserves, and their family members, providing them with government-subsidized medical and dental care. Eligibility for Veterans Administration (VA) health care benefits depends solely on active military service in the Army, Navy, Air Force, Marines, or Coast Guard. Enrolled veterans are assigned to one of eight priority levels based on their service-connected disabilities, income levels, and other factors. The Secretary of Veterans Affairs decides each year whether the VA's medical budget is adequate to serve veterans in all priority groups who seek care. CHAMP-VA (Civilian Health and Medical Program of the Department of Veteran Affairs) provides medical care for spouses and dependent children of disabled or deceased disabled veterans who meet the eligibility requirements of the VA.

2.2.3 Uninsured

A person is defined as uninsured if he or she does not have any private health insurance, Medicare, Medicaid, CHIP, state-sponsored or other government-sponsored health plan, or military plan. A person is also defined as uninsured if he or she has only Indian Health Service coverage or has only a private plan that pays for one type of service, such as accidents or dental care.

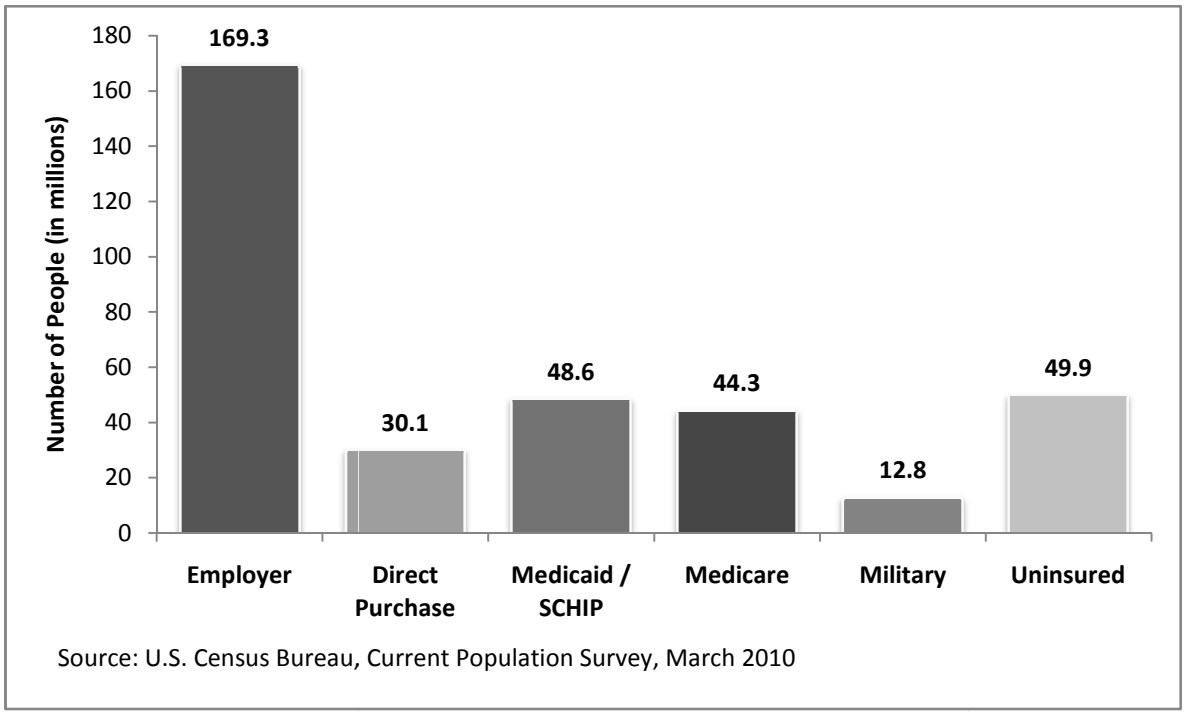


Figure 2.6: Health insurance coverage and type of coverage all person, 2010 [3]

Figure 2.6 shows the distribution of coverage types in 2010 which we retrieved from the recent health insurance historical tables available on the US Census Bureau's website. Total population is reported 306.11 million, but some people have more than one type of insurance that is why the sum in this chart is more than total. As we can see in this chart and also discussed before, most people in the US get coverage through an employer. The percentage of each coverage type or source of payment in 2010 is as following:

- Employer: 55.3%

- Direct Purchase 9.8%
- Medicaid / SCHIP 15.9%
- Medicare 14.5%
- Military 4.2%
- Uninsured 16/3%

2.3 Determinants of Healthcare Coverage

In this section, we introduce the determinants of healthcare coverage and explain some of them briefly. Knowing about these determinants is essential for the attribute selection step which is before modeling. In general, the factors that usually have been analyzed in past healthcare coverage studies are pretty the same. We can partition them to 4 categories: socio-demographic, financial, lifestyle, medical. Examples of each category have been retrieved from [2]. The categories including their examples are in the following:

- *Socio-Demographic:* Gender, race, age, marital, education, employment, etc.
- *Financial:* Income, poverty level, expenditures, etc.
- *Lifestyle:* Smoking, risk-taking, alcohol consumption, etc.
- *Medical:* Disabilities, priority conditions (chronic illness), etc.

A lot of studies have been accomplished about the impacts of various factors on healthcare coverage which the results of these studies strengthened our motivation to do this research. Monheit & Vistnes [28] have explored that lack of healthcare coverage will reduce the access to medical care. More, this issue will also cause more preventable hospitalizations [29]. And finally, poor health for the people without healthcare coverage and early death are the worst impacts that could happen due to this matter [1].

Moreover, Hoffman and Paradise [27] believe health insurance, poverty, are all interconnected in the US. In addition, they showed that health insurance has a strong association with access to primary and preventive care, and treatment and medical management of priority conditions. Usual source of care is one of the popular indicators of access to care. Having usual source of care, such as a health clinic, is a major factor in the quality of primary care which improves the health status [30]. On the other hand, having access to care by protecting people against high medical costs which usually emerge suddenly and also connecting them to healthcare systems is the outcome of having health insurance [27]. Institute of Medicine [31] has indicated in one of their reports that although health insurance alone cannot remove disparities in access to care and health inequalities among different groups of people, but it definitely can improve health quality and life duration.

Race and ethnicity is another factor that can have impact on health insurance from different points of view. Usually, people in different races and ethnics have different cultures, habits, and behaviors. In some specific people of color, due to lower attention to their financial and education issues, some high-risk behaviors happen which might negatively affect their health insurance, such as heavy drinking, driving without a seatbelt, unsafe sex, and lack of preventive care [32].

Sethi and Jain [2] did a feature selection for healthcare coverage modeling using three feature selection methods including chi-square, gain ratio, and info gain. Heavy drinking, state, income, medical cost, education, employment status, marital status, smoking status, and general health were the in common top 10 attributes.

2.4 Data Mining in Healthcare Coverage

As an introduction to this section, we present the major applications of data mining in the health area. Usually there are four types of applications which are listed and defined in the following:

- *Data mining in clinical medicine:* Hospitals and clinical centers are now source of huge amount of data including clinical, laboratory, equipment use, and drug management data. Finding knowledge and manipulating them without data mining techniques seems to be impossible due to the complexity in the data [33].
- *Data mining in public health:* Using data mining techniques for the aim of biomedical surveillance such as medical errors, death rates, and etc. gives new information which can be used in planning, implementation, and evaluation purposes of public health [34].
- *Data mining in healthcare text mining:* This application will be done usually for two purposes. First, text mining the medical and healthcare literature to find hidden relationships between biomedical entities. Second, text mining clinical documents such as patient records to track the presence of a specific illness in a specific population [35].
- *Data mining in healthcare policy and planning:* Policymakers and planners can employ data mining techniques for decision making processes, predicting the outcome of different health situations, and also evaluating the current health conditions.

Our study is in the field of data mining in public health, due to our objective to explore the impacts of different factors in healthcare coverage and recommend more efficient factors for

health planners. In continue, we review the related works and explain some of their results with more details.

In past studies, healthcare coverage disparity of specific populations has been studied using both statistical techniques such as logistic regression [36], and data mining techniques such as decision trees and neural networks [5]. The big advantage of data mining compared to statistics is facing with large amount of data and including more samples to find a pattern [6]. The procedures that the related works have done to contribute this study are very similar. In figure 2.7, we can see the generic model of these types of researches as a framework. The relations between main components have been displayed in this figure.

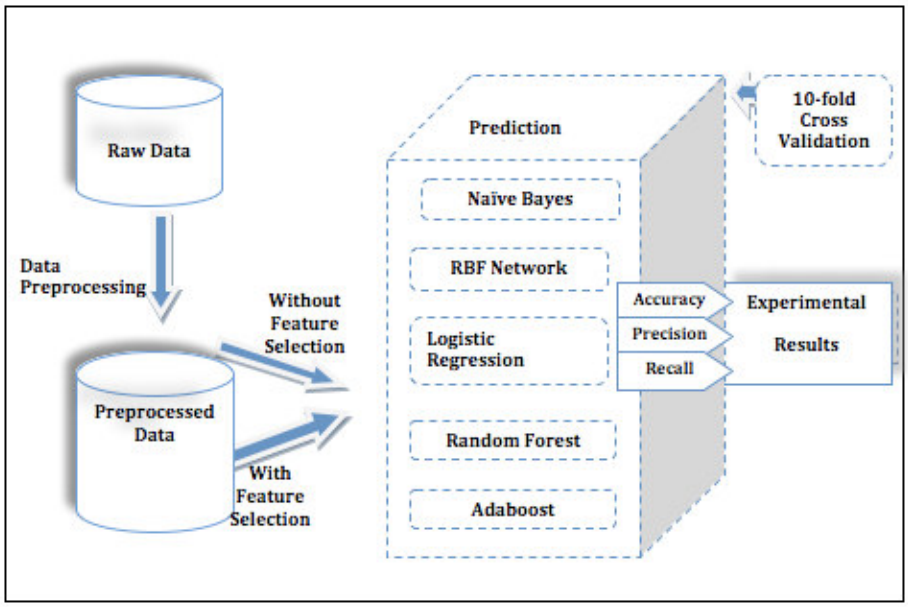


Figure 2.7: Framework of healthcare coverage prediction model [2]

As we mentioned before, Sethi and Jain [2] did a feature selection for healthcare coverage modeling using three feature selection methods including chi-square, gain ratio, and info gain. Heavy drinking, state, income, medical cost, education, employment status, marital status, smoking status, and general health were the in common top 10 attributes. They also built classification models using statistical algorithms using the first 10, 13, and 18 attributes

of each feature selection method. But we believe that the combination of attributes is also important. Therefore, we should reduce the least important attributes one by one and see what the attribute ranking of the new combination is. However, Sethi and Jain [2] have discussed about optimal set of features, but still their conclusion is based on removing a number of attributes from the initial attribute ranking, regardless of the impact of attributes different combinations.

Building a classifier model using either statistical techniques or data mining techniques is a kind of routine and determined method in all related works. For instance, we will go through the results that Delen *et al.* have published in 2009 [5] specifically. They have used Decision Tree (Classification and Regression Tree algorithm) and Neural Networks (Multi-Layer Perceptron algorithm) modeling to predict healthcare coverage using Statistica tool. Note that in data mining context, Neural Networks and Artificial Neural Networks literally have the same meaning. In their results, they have calculated two accuracy rates for each model – one for the records with healthcare coverage, and the other for the ones without healthcare coverage. The mean of these two accuracy rates will be the overall accuracy rate of each model. Accuracy rate will be calculated based on the confusion matrix that will be explained in the research methodology chapter. Their goal was to build models that predict the outcomes 25% better than random chance. As we can see in table 2.1, which is the accuracy rates of all outcome models in this article, the Neural Network modeling has a better accuracy rate compared to the Decision Tree modeling. In other words, NN has predicted the healthcare coverage status better than DT.

Finally, they have ranked the attribute importance for each outcome model. This ranking for the DT model is based on the variable importance measures, and for the NN model is based on the sensitivity analysis. Table 2.2 shows the top section of attribute ranking in this article.

Table 2.1: Accuracy rates of outcome models [5]

Accuracy of classification models

	Overall accuracy (%)	With health coverage accuracy (%)	Without health coverage accuracy (%)
Artificial neural network	78.45	80.05	76.86
Decision tree	74.11	72.71	75.51

Table 2.2: Ranked variable importance [5]

Ranked variable importance for ANN and decision tree models

Neural network		Decision tree	
Variable name	Sensitivity	Variable name	Importance
Income	1.000000	Income	1.000000
Employment status	0.879992	Education	0.619284
Education	0.821284	Marital status	0.501539
Marital status	0.800326	Employment status	0.460933
State	0.650258	Smoking status	0.395934
Age	0.647864	State	0.253011
Race	0.575093	Race	0.208743

Income, employment status, education, marital status, state, and race are the in common important attributes. The importance of this ranking is for the providers or the government so that they can target services to those without coverage and ones more in need, efficiently.

We have to add that in past works, they have just analyzed the coverage status of each sample in the reference time (yes or no). But we believe that this doesn't help the model to give complete information, because the attributes that we are working on are more stable and represent almost the last year situation of each person. Therefore, classifying samples based on their last year coverage status will have more benefits and makes more sense.

Some researchers think that the interviewees might have problems in recalling their previous health insurance status and this might cause errors in the surveys, so it's better to use the

current status as reference period [37]. But in MEPS which uses NHIS's data collecting, we have the report of health insurance status month by month. Therefore, we can build a new insurance coverage variable from the monthly reported variables.

Researchers [38] [39] [40] believe that those who are uninsured for a long period of time have different socio-demographic factors with the ones who are uninsured for just a short period of time. People who cannot provide health insurance for a long period of time have more difficulties obtaining insurances. Therefore, using a cross-sectional survey which reports the current health insurance status of samples may not be a very integrated data. For instance, a person has a very good health with a very high income experiencing health insurances for all times. But, in a specific time that he has been interviewed, his health insurance has been expired and he is attempting to find a better plan. Actually, this person shouldn't be classified as an uninsured sample because of this specific period of time. Most of the past studies in predicting healthcare coverage such as Delan *et al.*'s study [5] have used data with point-in-time (current) reference for the samples' health insurance status. Maybe a repeated cross-sectional or annual look-back time (past 12 months) survey can be more effective in building a health insurance model.

In this chapter we covered a variety of materials from healthcare coverage statistics to data mining in healthcare coverage. Our aim was to present the recent and related contents to our topic in a logical stream. The generic framework of our research study was expressed in this chapter. Regarding this literature review, different stages of this study has been proposed, tested, and analyzed in the next chapters.

Chapter Three

3 RESEARCH METHODOLOGY

In this chapter, we discuss all aspects of our research methodology including the research design, methodology, validation process, and measurement methods. All of these aspects introduce our research study from different point of view. First, the main characteristics of this research have been explained in the research design section. Second, the specific model and framework of this study has come in the methodology section with complete and in detail description. Third, the criteria to validate this study have been expressed in the validation process section. Finally, different methods and tools for measuring the outcome models have been declared in the measurement methods section.

3.1 Research Design

Before explaining the methodology of this study, we introduce the characteristics of this research in a breakdown table to understand the type of this research. Table 3.1 lists the features and characteristics of this research which we have designed to fulfill our objective.

3.2 Methodology

Specifically, the big difference of this study with the related works is that we have added two important components to our approach– Attribute Reduction and Clustering. The reason of adding each of them has been explained in their own step in the following. Figure 3.1 displays our specific model for this research.

Table 3.1: The research characteristics

Category	**Option**	**Reason**
Degree of Problem Crystallization	*Formal*	We are going to test our hypotheses
Method of Data Collection	*Survey*	We have questioned the subjects
Researcher Control of Variables	*Ex Post Facto*	Our data has been collected using sampling, so we have no control over the variables
The Purpose of the Study	*Descriptive*	We are finding out *what* are the factors
The Time Dimension	*Repeated Cross-Sectional*	Our sampling is several snapshots of data in the time dimension
The Topical Scope	*Statistical Study*	We are using quantitative measurements
The Research Environment	*Laboratory*	We are going to use software application to do the analysis in a normal condition

The literature review part has been covered in the Background chapter. Introducing the raw data, selecting the attributes, preprocessing the data, and doing an initial attribute reduction, are included in Data Preparation chapter. In Healthcare Coverage Modeling chapter, we first propose the outcome models, and then, build the models by employing supervised and unsupervised learning (classification and clustering algorithms), and will test them to get their results. Using these results we reduced attributes and again did the modeling. Finally, the results were analyzed in this chapter and discussed with more detail. Each part has been explained briefly in the following:

- Literature Review: Mainly using Medline (All PubMed index databases), SCOPUS (web of science), COMPENDEX, Springer e-Book collection, IEEE Xplore, and other major relevant databases; we will focus on these issues and titles:
    a. Healthcare coverage concepts such as predictive factors and analysis models.

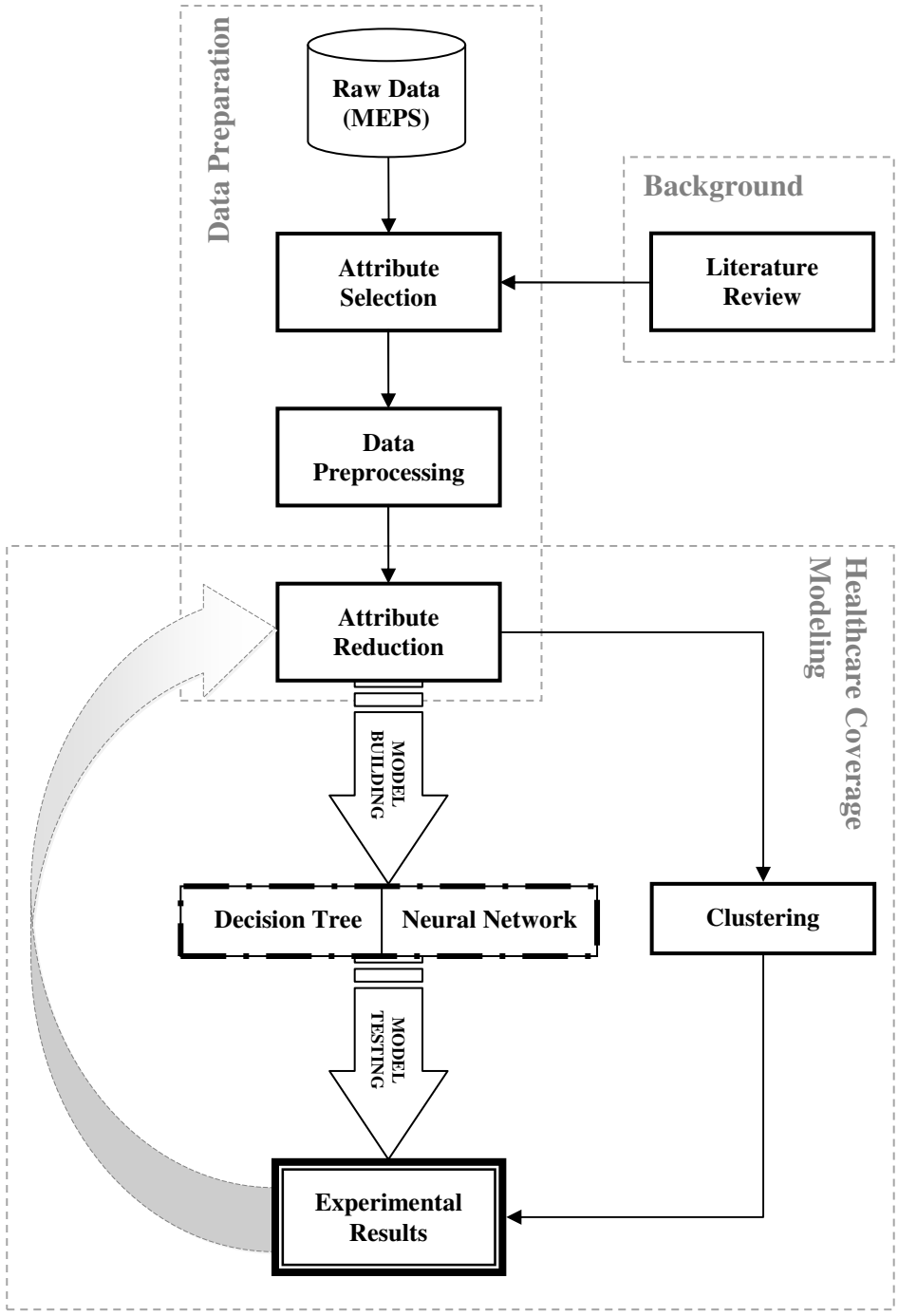


Figure 3.1: Our specific research method

b. Machine learning algorithms, mostly used in biomedical field, with a focus on two main predictive models: Decision Tree and Neural Network.

c. AHRQ's Medical Expenditure Panel Survey (MEPS) website and relevant researches.

- Steps in Research Methodology:

    a. Raw Data (Database): The ASCII format of MEPS-HC 2006-2008 Full Year Consolidated file, for the US non-institutionalized population representing general population of the US in 2006, 2007, and 2008 is available. We first store the codebook information of our data file. Then, we start separating each data using commas. For this part we use a programming language. Lastly, we import the new generated data file in Microsoft Excel program to display our data file in a tabular view.

    b. Attributes Selection: Regarding the information achieved from the literature review, we reduce the original data table to a smaller one to analyze. The outcome of this part is a table listing the selected attributes with a brief description and also the range of values that each of them have. Eventually, we choose about 20 attributes (with relevant support from literature) from these classes of variables present in MEPS-HC 2006-2008 dataset, including:

        - Demographic variables
        - Health status variables
        - Income & tax filling variables
        - Disability-days indicator variables
        - Access to care variables
        - Employment variables
        - Health insurance variables

    c. Preprocessing the data set by:

        - Removing zero-weighted samples

- Removing samples with missing data
- Relabeling attribute values
- Combining more detailed attributes
- Normalization (Data distribution)
- Balancing samples (Not biased)

d. Attribute Reduction: Regarding attribute reduction techniques, we reduce the number of variables to build our model using the smallest amount of number that is possible [41]. This helps us to rank the efficient set of factors more easily at the end. This method is not guaranteed to be optimal due to not considering all combinations of attributes in each stage, but it is an intuitive heuristic that gives a reasonable result and close to optimal.

e. Model Building: Feeding prepared dataset into different data mining's classification algorithms. Note that the dataset will be either cross-validated or separated to a train-test set. For cross-validation, a 10-fold cross-validation is used in order to select the best model. A *v*-fold cross validation method estimates the true error rates of the classification model better than a single train-test set experiment [42]. If there was no chance of cross-validation, a 70-30 percents train-test set is determined for building the model. The train partition of data will be used to build and train the models, then for the classification models, the outcome models will be tested using the test partition of data to see with how much accuracy the model can predict the test data.

f. Clustering: The original dataset and the uninsured samples dataset are clustered separately, respectively before and after the classification modeling. The clustering procedure is done using clustering (segmentation) techniques, such as K-Means clustering algorithm. This gives us useful information about the patients' similarities in the dataset [41]. Based on this information, we complement our analysis and make a more comprehensive conclusion. Note that, all input data has been used to train the unsupervised learning models and there is no data partitioning for this section.

g. Classification Models: Building a model that predicts the healthcare coverage status of samples using IBM SPSS Modeler. To do this classification, we use Decision Tree and Neural Network algorithms. Classification models are applied to find *the healthcare coverage threshold*, *the better healthcare coverage classifier*, and also *the attribute ranking*.

   Logistic regression is also a good option for classification, but because past researchers have employed this method for the same purpose and also logistic is not a pure data mining technique with all its features (it is more statistical technique using regression) we didn't select it for our main tests. It has been used just for the threshold analysis in this study.

   Threshold analysis is done to see whether we can define a threshold for healthcare coverage or not. If so, for example, we can say when a person has been covered in at least 8 months of the last 12 months, this person has the same behavior as a full-covered person has and vice versa.

   Different classification models are applied to see which classifier can predict healthcare coverage the best. After building each model, the input attributes

are ranked in terms of their relevance to output labels and their importance as a predictor. The result of this part helps researchers and organizations to provide efficient solutions for healthcare coverage plans.

h. Model Testing: We use a confusion matrix to test sensitivity and specificity of our models in order to find their final rank in correctness of prediction.

i. Experimental Results: The classification algorithms are ranked according to their potency in predicting suitable class (whether a person has a coverage or not) for test set. Attribute reduction and ranking procedures are reported in comprehensive tables. Results of the threshold analysis are also compared in graphical charts. Also, different clusters of datasets are presented in graphical charts and descriptive tables.

3.3 Validation Process

In the validation process, we define the criteria and scale for our study's concepts based on their measurable definitions. Then, we explain about how we calculate our scales in detail. Table 3.2 displays the validation process of this study.

From another point of view, we have 4 validation aspects:

- *Internal validity:* Medical expenditure dataset factors will be ranked based on their importance rate in each model. Also, our healthcare coverage prediction models will be ranked based on their overall accuracy rates. Finally, we can compare them based on these rankings.
- *External validity:* There is a possibility that the results of this study can be used for other similar societies such as Canada. But validating our models for this purpose is not in our research scope.

Table 3.2: Validation process

Concept / Construct	Operational definition	Properties	Criteria	Indicator	Scale
Healthcare Coverage	Level of healthcare coverage	Healthcare coverage duration	Annual Percentage of coverage	Number of months having healthcare coverage during a complete year	0-12
		Healthcare coverage threshold	Comparing accuracy rates of classification models with different threshold	Minimum acceptable number of months having coverage during a complete year	1-12
Medical Expenditure Dataset Factors	Effectiveness of subset of independent medical expenditure variables	Effectiveness of each factor on the model	Importance of each predictor in a model	Predictor importance	0-1
		Smaller amount of factors in subset	Quantity of factors in subset	Number of input attributes	1-22
Healthcare Coverage Model	Accuracy rate of assigning samples to healthcare coverage classes and groups	Classification Model	Decision Tree	DT's Gain ratio	0-100%
			Neural Network	NN's Gain ratio	0-100%
		Clustering Model	K-Means	Silhouette average	-1 to 1

- *Reliability:* The samples that we are using have weight which shows that how much it is representing the whole population of the US. Based on these weights we can prove that our results are reliable for at least the United States.

In continue, we explain the measurement methods in more detail. Note that we use IBM SPSS Modeler software to build our models and also through this software we do the analysis section and other related calculations.

3.4 Measurement Methods

To analyze the outcome results, we need some valid methods to calculate the accuracies and validities of the outcome models. In this part we introduce the main measurement methods that we used in this project.

3.4.1.1 Confusion Matrix

To accomplish the validation process, we create a confusion matrix (figure 3.2) after building and testing an outcome model in classification techniques, which shows the experimental results of that model. For instance, TP indicates the number of samples with positive target value which have been predicted as positive correctly. On the other hand, FN shows the number of samples with positive target value which have been predicted as negative by mistake.

		Predicted Outcome	
		Positive	Negative
Actual Value	Positive	TP True Positive	FN False Negative
	Negative	FP False Positive	TN True Negative

Figure 3.2: Confusion matrix

3.4.1.2 Accuracy Rate (Correctness)

One of the measures to calculate the accuracy and precision of an outcome model is the accuracy rate or correctness. This parameter shows how much the records have been classified correctly. The accuracy rate or correctness will be calculated using this formula:

$$\text{Accuracy rate (Correctness)} = \frac{TP + TN}{TP + TN + FP + FN}$$

We will get two accuracy rates for each model – one for the training set, and the other for the testing set. The mean of these two accuracy rates make the overall accuracy rate of each model. In IBM SPSS Modeler, these two accuracy rates are usually similar to each other. We compare the accuracy rates of all outcome models to analyze them.

3.4.1.3 G-Mean

For the threshold analysis section, we also calculate another type of accuracy rate called g-mean. The g-mean can show whether the model has a good accuracy for all values, or just for some of them. If the g-mean result is the same as the accuracy rate (correctness), it means that the model is not biased and this accuracy rate is valid for all values of the target variable. G-mean is calculated using this formula:

$$G - mean = \sqrt{Sensitivity \times Specificity} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}$$

3.4.1.4 Predictor Importance

We rank the attribute importance for each outcome model. Usually in other related works, researchers rank each input field based on the strength of its relationship to the specified target, independent of other inputs, which they call it feature selection. But, the predictor importance chart indicates the relative importance of each input for this particular model. For example, if wage and person-level total income are both strongly related to healthcare coverage, then feature selection would indicate that both are important, but the predictor importance will select one of these two inputs if both represent similar information. In modeling and accordingly the predictor importance, interactions and correlations are also

taken into consideration. Particularly, when dealing with large datasets with large numbers of attributes, and predictor importance is more useful in fine-tuning the model [43]. Note that the predictor importance which IBM SPSS Modeler uses is not related to model accuracy, and since the values are relative, the sum of the values for all predictors of a model is 1.

3.4.1.5 Silhouette Average

This analysis method is used to evaluate clustering models. The concept behind this method is combining both cluster cohesion (favoring models which contain tightly cohesive clusters) and cluster separation (favoring models which contain highly separated clusters) in one evaluation method. The average over all cases builds the Silhouette average. The Silhouette measure for each case will be calculated using this formula:

$$(B - A) / max(A, B) ;$$

where A is the distance from the case to the centroid of the cluster which the case belongs to; and B is the minimal distance from the case to the centroid of every other cluster [43]. Regarding IBM SPSS Modeler's application, the range of Silhouette average is between -1 (very bad model) to 1 (very good model). Note that the Silhouette average above 0.2 shows a fair clustering quality, and above 0.5 is for good quality clustering models.

This chapter described our research study from different point of views. Moreover, the step by step framework or methodology was explained and defined. Regarding the information in this chapter, other steps of the project are presented and expressed in the next chapters based on a logical flow.

Chapter Four

4 DATA PREPARATION

Data mining's big ability and advantage is to find meaningful patterns through large amounts of data [6]. Therefore, one of the important phases in data mining is to prepare a large comprehensive dataset. Applying data mining techniques on a deficient dataset might end up in poor and unreliable results. Unreliable analyses might cause high sensitive costs and problems in serious fields of study, such as healthcare coverage. We explored various healthcare data files through internet to find the appropriate data file which has collected all our required information in itself. To understand each data file's structures and aims, we should go through its documentations and explanations.

Going straight into modeling before understanding the meaning of our data might be exciting, but won't help us in achieving acceptable results. Usually, real-world data should get ready before any analysis because they are typically noisy, large in volume, and collected from different sources [41]. Data preprocessing will act a significant role in any data mining process to develop a proper dataset based on the project's objectives.

In this chapter, first we introduce the data file that we selected to work with. Second, the attributes that were selected from the data files are listed with their definitions. Then, we start doing the preprocessing phase to prepare our desired dataset for modeling and data analysis. Finally, we talk about the initial attribute reduction we did before modeling.

4.1 AHRQ's Medical Expenditure Panel Survey

To accomplish our study, we selected a database which had all information we needed for this study. "Medical Expenditure Panel Survey" database is prepared by "Agency for Healthcare Research and Quality", which we call it MEPS for short. MEPS contain different components, but all of them are not available to public. The one that we worked on was the Household component. The MEPS Household Component fields questionnaires to individual household members to collect nationally representative data on demographic characteristics, health conditions, health status, use of medical care services, charges and payments, access to care, satisfaction with care, health insurance coverage, income, and employment [8].

MEPS-HC was initiated in 1996. The samples are from individuals and members of different families. Each year a panel will be designed to collect information from these samples. Each panel contains about 40 questionnaires and will be held in 5 rounds (interviews). It will take two calendar years to collect data for each panel. Sometimes, new sections will be added to new panels. Most of the sections will be presented using a software application – computer-assisted personal interviewing (CAPI) and their data will be collected electronically, but some sections will be presented paper-based which are called supplemental paper questionnaires [44].

The set of households selected for each panel of the MEPS-HC is a subsample of households participating in the previous year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics. The NHIS sampling frame provides a nationally representative sample of the U.S. civilian non-institutionalized population [45].

Figure 4.1 shows how MEPS-HC data collection works visually. Each year, the collected data consist of rounds 1-3 of the new panel and rounds 3-5 of the previous panel.

Questionnaires of each section in different rounds of a specific panel are pretty the same with just a little difference. Questionnaires of each panel consist of Access to Care, Dental Care, Emergency Room, Priority Conditions, and etc. For instance, in the priority conditions, some specific diseases such as asthma, diabetes, and etc. will be covered.

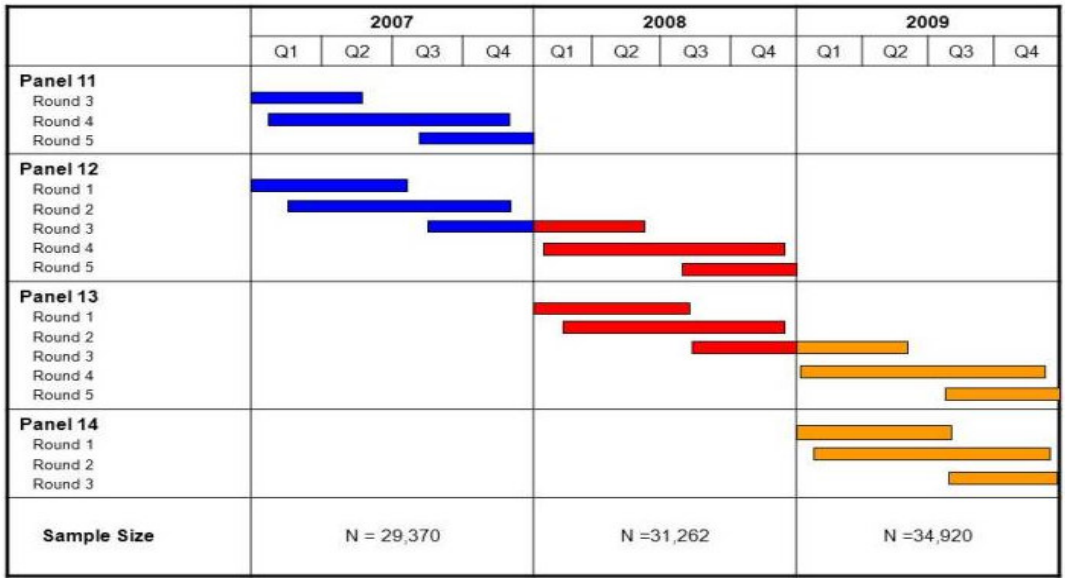


Figure 4.1: MEPS-HC Panel Design and Data Collection Process [44]

Each year, MEPS releases various data files regarding different subjects and events. But, the one that has all the information about the samples of each year is called the consolidated file of full year. The recent consolidated file of a full year is for year 2008. It is a huge data collected which contains mostly all medical and expenditure information of all the samples. Each record contains 5255 columns which represent the values of 1823 variables. In other words, each variable is shown using one or more columns. Each column contains one character and because these characters are in a raw ASCII file, we call each of them a column. More, there are 33066 samples (records) in this data file [46].

Each data file has a documentation file and a codebook. In the documentation file, all materials related to this collected data have been explained completely. However, in the codebook, we can find the title of each variable and its position in each record (Figure 4.2).

MEPS H121 CODEBOOK
2008 FULL YEAR CONSOLIDATED DATA FILE
DATE: November 4, 2010

Variable Name: DSCH0853
Description: DCS: BLOOD CHOLESTEROL CHECKED IN 2008
Format: 2.0
Type: NUM
Start: 1076
End: 1077

VALUE	UNWEIGHTED	WEIGHTED BY PERWT08F
-9 NOT ASCERTAINED	15	130,614
-8 DK	3	32,505
-1 INAPPLICABLE	31,225	286,306,624
1 YES	1,472	14,940,782
2 NO	351	2,965,417
TOTAL	33,066	304,375,942

Figure 4.2: Details of a variable in the codebook [46]

Note that each sample will be weighted to generate MEPS estimates of totals, means, percents, and rates for persons and families. Just a small proportion of samples in MEPS data files are un-weighted. These persons include those who had periods during which they lived in an institution (e.g., nursing home or prison), were in the military, or lived out of the country, as well as those who were born (or adopted) into MEPS sample households or died during the year [45].This weight depends on the sample's effectiveness on the national representative model of MEPS which will be explained in the variable definition part.

Each year's full year consolidated data file of MEPS-HC contains about 30,000 samples. Therefore, we increased the volume of our dataset by aggregating data files of 3 years: 2006 to 2008.The advantage of increasing dataset's volume is that the outcome model will be trained with more samples. The advantage of data mining's techniques is to find meaningful

rules through bulky databases [6], thus, applying data mining modeling on small datasets might cause unreliable results.

In the documentation file of years 2006 and 2007 full year consolidated data file, we noticed that some of the priority condition variables that we wanted to include them in our dataset were missing. Thus, we had to also refer to the medical conditions data file for years 2006 and 2007 to complete our missing values. We explain these data files regarding their collected data in continue.

4.1.1 Full Year Consolidated Data File

These files are released as ASCII files which provide information collected on a nationally representative sample of the civilian non-institutionalized population of the United States for calendar years 2006 to 2008 [45]. Number of variables, logical record length, number of samples, and MEPS survey data obtained for years 2006 to 2008 are in the following table:

Table 4.1: Specifications of full year consolidated data files

Year	Number of Variables	Logical Record Length (characters)	Number of Samples	MEPS Survey
2006	1,672	4,612	34,145	Rounds 3-5 of Panel 10 Rounds 1-3 of Panel 11
2007	1,787	5,173	30,964	Rounds 3-5 of Panel 11 Rounds 1-3 of Panel 12
2008	1,823	5,256	33,066	Rounds 3-5 of Panel 12 Rounds 1-3 of Panel 13

The variables that have been driven from the full year consolidated data files are listed and explained briefly in the following sections. Most of the variables have been reported for all rounds of the interview and also at the end of that calendar year. For instance, age of each person have been recorded in all these phases:

- Round 3 of Panel X or Round 1 of Panel X+1

- Round 4 of Panel X or Round 2 of Panel X+1
- Round 5 of Panel X or Round 3 of Panel X+1
- December 31, 2006 or 2007 or 2008

We have selected the end of year time spot as the reference time, therefore, variables recorded for December 31, 2006 or 2007 or 2008 have been chosen. If this time phase of a variable wasn't available, we have selected the data reported for "Round 5 of Panel X or Round 3 of Panel X+1" phase. In addition, the variables selected for our dataset are based on the literature.

4.1.2 Medical Conditions Data File

Some of the priority conditions (chronic conditions) were not reported directly in the Full Year Consolidated data files of years 2006 and 2007. Therefore, we had to explore the Medical Conditions data files of these two years to complete the missing variables. These data files were also in ASCII format. In these data files, each record represents one medical condition of each person. These data files can be merged with the Full Year Consolidated data files using each person's unique identifier number (DUPERSID).

MEPS have aggregated ICD-9-CM condition codes which are similar to each other and can be grouped as a meaningful category in one clinical classification code (CCCODEX) [47]. For example, a person who has cancer of breast will have a record in the medical condition file with a CCCODEX value of 024. The CCCODEX value of 024 includes the following ICD-9-CM codes: 1740-1750, 1759, 2330, and V103. Using the clinical classification code table in each medical conditions data file's documentation, we found the persons who have the missing priority conditions. Then, we added the results to the Full Year Consolidated data files to fill out the missing values.

Missing data of cancer variable in Full Year Consolidate data files were filled out by joining it with Medical Condition data files based on sample IDs. List of cancer diagnoses which we retrieved using CCCODEX variable of the Medical Condition data files are as following:

Table 4.2: Cancer indicators in Medical Condition data files [47]

CCCODEX	Description
011	CANCER OF HEAD AND NECK
012	CANCER OF ESOPHAGUS
013	CANCER OF STOMACH
014	CANCER OF COLON
015	CANCER OF RECTUM AND ANUS
016	CANCER OF LIVER AND INTRAHEPATIC BIL
017	CANCER OF PANCREAS
018	CANCER OF OTHER GI ORGANS, PERITONEU
019	CANCER OF BRONCHUS, LUNG
021	CANCER OF BONE AND CONNECTIVE TISSUE
024	CANCER OF BREAST
025	CANCER OF UTERUS
026	CANCER OF CERVIX
027	CANCER OF OVARY
028	CANCER OF OTHER FEMALE GENITAL ORGAN
029	CANCER OF PROSTATE
030	CANCER OF TESTIS
031	CANCER OF OTHER MALE GENITAL ORGANS
032	CANCER OF BLADDER
033	CANCER OF KIDNEY AND RENAL PELVIS
035	CANCER OF BRAIN AND NERVOUS SYSTEM
036	CANCER OF THYROID
041	CANCER, OTHER AND UNSPECIFIED PRIMAR

4.2 Attribute Selection

Based on the literature, we selected attributes that were related to healthcare coverage in the MEPS data files. In this part, we introduce the variables that were selected briefly.

4.2.1.1 Variable Categories

MEPS-HC variables are divided into distinct categories. The selected variables grouped into their categories have been listed in the following:

- Survey administration:
    - *DUPERSID, REGION0#, MSA0#*
- Demographic:
    - *AGE0#X, SEX, RACETHNX, MARRY0#X, FTSTU0#X, EDUCYR, HIDEG, ACTDTY53, HONRDC53*
- Income and tax filing:
    - *WAGEP0#X, TTLP0#X, POVCAT0#*
- Person-level condition:
    - *RTHLTH53, MNHLTH53, HIBPDX, CHDDX, ANGIDX, MIDX, OHRTDX, STRKDX, EMPHDX, CHOLDX, DIABDX, JTPAIN53, ARTHDX, ASTHDX*
- Health status:
    - *IADLHP53, ADLHLP53, WLKLIM53*
- Access to care:
    - *HAVEUS42*
- Employment:
    - *EMPST53*
- Health insurance:
    - *INSJA0#X, INSFE0#X, INSMA0#X, INSAP0#X, INSMY0#X, INSJU0#X, INSJL0#X, INSAU0#X, INSSE0#X, INSOC0#X, INSNO0#X, INSDE0#X*
- Person-level weight:

- *PERWT0#F*

The definition of each variable is expressed in the next part.

4.2.1.2 Variable Definitions

All variable definitions are evoked from the data file's documentation [45].

DUPERSID:

A combination of two identity numbers (DU and PERSID). DU stands for Dwelling Unit and PERSID is the person's unique number within the DU. Using this ID number (DUPERSID), each person can be identified uniquely.

REGION0#:

The census region for the living group (e.g. family), which includes the interviewee. Northeast, Midwest, South, and West are the assigned values. For instance, New York is in Northeast, Indiana is in Midwest, Florida is in South, and California is in West.

MSA0#:

This variable indicates whether or not the living group which includes the interviewee is found in a Metropolitan Statistical Area. The definition of metropolitan statistical area has been achieved from the Office of Management and Budget (OMB)'s recent standard definitions.

AGE0#X:

Age of the interviewee is recorded in this variable. The age variable was top-coded at 85 years. Thus, ages of samples are from 0 to 85.

SEX:

This variable shows whether the person is male or female.

RACETHNX:

RACETHNX summarizes both race and ethnicity in one variable. The values that this variable can take are as following:

- Hispanic
- Black (Not Hispanic or other races)
- Asian (Not Hispanic or other races)
- Other race (including White)

MARRY0#X:

Person's marital status is reported through this variable. Possible answers are:

- Married
- Widowed
- Divorced
- Separated
- Never married
- Under age 16

FTSTU0#X:

This variable shows the person's student status (full-time / part-time / not a student) and is only valid for persons between the ages of 17 – 23.

EDUCYR:

Number of years of education is indicated in this variable. Children under the age of 5 were inapplicable for this variable. Also, it has been top-coded at 17 years.

HIDEG:

Highest degree of education is recorded in this variable. Possible answers for this variable are:

- No degree
- GED (General Educational Development)
- High school diploma
- Bachelor's degree
- Master's degree
- Doctorate degree
- Other degree
- Under age 16

ACTDTY53:

Information on active duty military status of the person has been indicated in this variable. Recorded values are:

- Yes – Active duty
- No – Not full-time active duty
- Under age 16
- Over age 59

HONRDC53:

Persons who have been honorably discharged from active duty in the Armed Forces are identified by this variable. Recorded values are:

- Yes – Honorably discharged
- No – Not honorably discharged
- Under age 16
- Now active duty

WAGEP0#X:

It shows person's annual wage and salary income. The minimum and maximum amounts that have been reported are respectively 0 and 671,978 US dollars.

TTLP0#X:

Total person-level income is the sum of all income components with the exception of person's refund and sales incomes. The minimum and maximum amounts reported are respectively -186,193 and 699,225 US dollars.

POVCAT0#:

The percentage result of dividing interviewee's family income by the applicable poverty line – which is based on family size and composition – has been classified into five poverty categories:

- Negative or poor: less than 100%
- Near poor: 100% to less than 125%
- Low income: 125% to less than 200%
- Middle income: 200% to less than 400%
- High income: 400% or greater

Note that the definitions of income, family, and poverty categories used to construct the related variables have been taken from poverty statistics developed by the Current Population Survey (CPS). CPS is a monthly survey of households conducted by the Bureau of Census for the Bureau of Labor Statistics in the United States [48]. Poverty thresholds in 2008 based on age and number of family members is available in Appendix I.

RTHLTH53:

Perceived health status which has been asked by the respondent to rate according to the following categories: excellent, very good, good, fair, and poor.

MNHLTH53:

Perceived mental health status which has been asked by the respondent to rate according to the following categories: excellent, very good, good, fair, and poor.

HIBPDX:

HIBPDX ascertained whether the person (aged 18 or older) had ever been diagnosed as having high blood pressure (other than during pregnancy).

CHDDX:

CHDDX shows whether the person (aged 18 or older) had ever been diagnosed as having coronary heart disease.

ANGIDX:

ANGIDX shows whether the person (aged 18 or older) had ever been diagnosed as having angina, or angina pectoris.

MIDX:

MIDX shows whether the person (aged 18 or older) had ever been diagnosed as having a heart attack, or myocardial infarction.

OHRTDX:

OHRTDX shows whether the person (aged 18 or older) had ever been diagnosed with any other kind of heart disease or condition.

STRKDX:

STRKDX asked if the person (aged 18 or older) had ever been diagnosed as having had a stroke or transient ischemic attack (TIA or mini-stroke).

EMPHDX:

EMPHDX asked if the person (aged 18 or older) had ever been diagnosed with emphysema.

CHOLDX:

CHOLDX ascertained whether the person (aged 18 or older) had ever been diagnosed as having high cholesterol.

CANCERDX:

CANCERDX ascertained whether the person (aged 18 or older) had ever been diagnosed as having cancer or a malignancy of any kind. The type of cancer has been reported in other variables which indicate selection of cancer of the bladder, blood, bone, brain, breast, cervix, colon, esophagus, gallbladder, kidney, larynx, leukemia, liver, lung, lymphoma, melanoma, mouth/tongue/lip, ovary, pancreas, prostate, rectum, skin, soft tissue, muscle, stomach, testis, throat, thyroid, or uterus. This variable hasn't been reported in 2006 and 2007 Full Year Consolidated data files. We have filled out the missing values of this variable using the Medical Condition data files of that year.

DIABDX:

DIABDX indicates whether each person (aged 18 or older) had ever been diagnosed with diabetes (excluding gestational diabetes).

JTPAIN53:

JTPAIN53 asked if the person (aged 18 or older) had experienced pain, swelling, or stiffness around a joint in the last 12 months.

ARTHDX:

ARTHDX asked if the person (age 18 or older) had ever been diagnosed with arthritis. This includes all types of arthritis, such as Rheumatoid Arthritis, Osteoarthritis, and non-specific arthritis.

ASTHDX:

ASTHDX indicates whether a person had ever been diagnosed with asthma.

IADLHP53:

IADLHP stands for Instrumental Activities of Daily Living (IADL) Help which shows this person needs help or supervision in daily activities such as using the telephone, paying bills, taking medications, preparing light meals, doing laundry, or going shopping. For persons under age 13, another question will be asked to check if this help is a result of an impairment or physical or mental health problem.

ADLHLP53:

This disability variable indicates whether the person receives ADL (Activities of Daily Living) help or supervision due to an impairment or physical or mental health problem. ADLs include self-care tasks such as dressing, feeding ourselves, bathing, and working. Logically, the persons who have ADL limitations mostly will also have IADL limitations.

WLKLIM53:

The functional (or walking) limitation variable shows whether the person has difficulties walking, climbing stairs, grasping objects, reaching overhead, lifting, bending or stooping, or standing for long periods of time.

HAVEUS42:

HAVEUS42 (Have Usual Source of Healthcare) ascertains whether there is a particular doctor‘s office, clinic, health center, or other place that the person usually goes to if he/she is sick or needs advice about his/her health.

EMPST53:

This variable asks about employment status of the person (aged 16 or older). Possible employment statuses were defined as following:

- Currently employed: This person has a job right now.

- Has a job to return to: This person is not working now, but he/she has a job to return to.
- Employed during the round, not now: This person doesn't have any job right now, but he/she had a job in the round of survey.
- Not employed: This person doesn't have any job right now, doesn't have a job to return to, and did not work during the round of survey.

INSJA0#X - INSFE0#X - INSMA0#X - INSAP0#X - INSMY0#X - INSJU0#X - INSJL0#X - INSAU0#X - INSSE0#X - INSOC0#X - INSNO0#X - INSDE0#X:

Each of these variables indicate whether the person had any type of insurance coverage including TRICARE, Medicaid, Medicare, SCHIP, or other public hospital/physician or private hospital/physician insurance (including Medigap plans) in each month of the last year or not. The reason that we selected all 12 month variables of insurance coverage status is because of the longer term effect of other variables on the insurance coverage status. This matter has been discussed in the last section of the background chapter

PERWT0#F

A person-level weight has been provided for each person by MEPS to estimate the totals, means, percents, and rates for persons. Briefly, person-level weights were constructed based on different variables, such as age, sex, poverty, race and etc. which try to represent the estimated population of the US. In other words, the person-level weight of each individual roughly shows that this person is representing how many people in the US with accuracy of more than 90% and for sure less than 100% (samples are all civilian non-institutionalized). Zero-weighted persons cannot be considered as valid samples because their data is somehow incomplete and unreliable.

4.3 Data Preprocessing

Data can be in different forms and shapes. It may be even incomplete and inconsistence. To use data mining techniques, the dataset should be prepared and preprocessed. All data mining algorithms expect to have datasets in tabular view as input which each row represents a specific record or sample and the columns show different fields [6].

MS-Excel is popular and user friendly software for analytical purposes. We employed this software for the data preprocessing part so that in addition to preprocessing the data, we can also feel and get more familiar with the data available in MEPS. But, before using our dataset in MS-Excel application, we needed to convert the file format of the original data files to a file format which is understandable for MS-Excel application. To do this task, first we built our final dataset's codebook in MS-Excel application so that it shows the list of our selected variables, plus with their starting and ending column numbers in data files of years 2006 to 2008. Note that the order, type, and number of variables in all three codebooks were the same. Table 4.3 shows a preview of one of these codebooks. Then, we did programming to prepare our dataset in an appropriate file format – CSV file.

Table 4.3: Preview of selected variables codebook in 2008 Full Year Consolidated data file

Name	Start	End	Description
DUPERSID	9	16	PERSID (DUID + PID)
REGION08	73	74	CENSUS REGION AS OF 12/31/08
MSA08	81	82	MSA AS OF 12/31/08
AGE08X	193	194	AGE AS OF 12/31/08 (EDITED/IMPUTED)
SEX	216	216	SEX
RACETHNX	221	221	RACE/ETHNICITY (EDITED/IMPUTED)
MARRY08X	231	232	MARITAL STATUS-12/31/08 (EDITED/IMPUTED)
FTSTU08X	263	264	STUDENT STATUS IF AGES 17-23 - 12/31/08

Medical Condition data files of years 2006 and 2007 were also explored and their CCCODEX variable was retrieved. Based on the sample ID, we detected which samples of years 2006 and 2007 were diagnosed with cancer. Note that 2006 and 2007 Medical Condition data files were also converted to CSV files, and then imported to MS-Excel application for the data preprocessing operations.

4.3.1 Programming

MEPS data files are released in two formats: ASCII format and SAS transport format. For this study, we used the ASCII format data files and retrieved the data and tables using C++ programming language. To build our desired dataset we needed two different C++ programs to prepare the dataset we were looking for. First, we had to read these ASCII data files, split the data columns, select the columns we wanted, and write the data files in a format that was understandable for MS-Excel application. Second, we aggregated the data files of several years in one file. Note that we employed Microsoft Visual Studio 2005 Standard Edition as IDE to code these programs and run them as console applications.

4.3.1.1 ASCII to CSV

This program convert's a MEPS's ASCII data file to a CSV (Comma-Separated Values) file based on a specific codebook. Each data file in MEPS has its own codebook which shows the location of each variable in the samples (records). But we don't need all variables from a data file. Therefore, we determine the starting and ending column of the variables we want, and write them in a separate text file which will be used as an input file for this program. In addition, we will set the "NumberOfColumns" and "Year" variables which show respectively that data file's total column numbers and the year of that data file. Note that some of the variables didn't exist in all data files, thus, we inserted "-" (dash character) as

their data. Finally, we will get a CSV data file regarding our selected variables (based on their starting and ending columns) in an output file (figure 4.3). For more details about this conversion program, go to Appendix II which displays this program's coding plus explanatory comments.

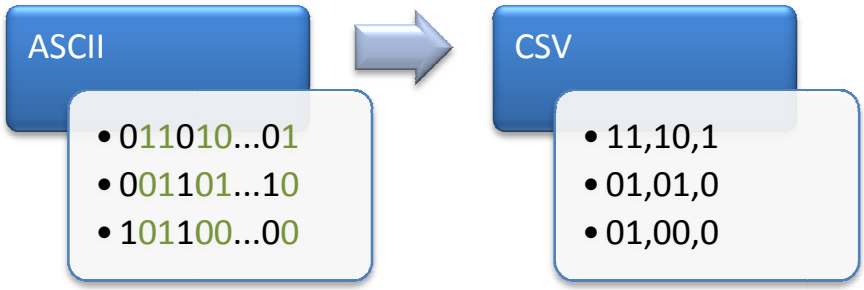


Figure 4.3: ASCII to CSV program diagram

4.3.1.2 Aggregate

In this study, we used Full Year Consolidated files of years 2006, 2007, and 2008. After the data files have been converted from ASCII file formats to CSV files based on their specific codebooks, we created a final codebook which shows our selected variables in one dataset next to each other. To combine these datasets, we need a program to aggregate the input datasets (CSV files) and build a bulky output dataset (a CSV file) based on the final codebook (figure 4.4). Finally, the aggregated data file is ready to be imported to MS-Excel software. For more details about this conversion program, go to Appendix III which displays this program's coding plus explanatory comments.

4.3.2 Data Cleaning

In this part, we apply different operations and processes on the real data to remove incomplete samples and prepare it for modeling. Each section will explain the operations step by step.

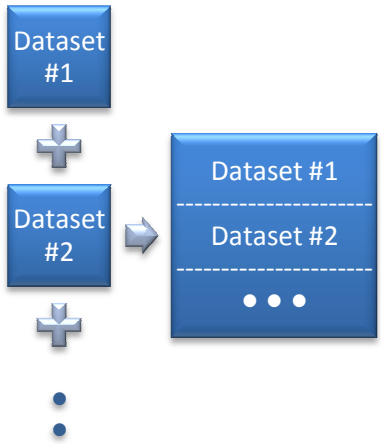


Figure 4.4: Aggregate program diagram

4.3.2.1 Raw Data

The dataset was prepared in a CSV file. Thus, it was imported in MS-Excel software regarding this condition that the commas separate the columns. After importing our final data file, now we can see the data in a tabular format. Samples of years 2006 and 2007 were checked whether they have been diagnosed with any type of cancer or not and the missing data were completed. This check was done by looking up at the dataset retrieved from the Medical Condition data files which showed the people who had any type of cancer – sample's ID and clinical classification code (CCCODEX). This look up and join was done using the following formula in MS-Excel:

=IF(LOOKUP(A1,Z1:Z100)=A1,1,0)

We call this version of the dataset as "Raw data". The raw dataset contains 98,175 samples from years 2006 to 2008.

Note that the codebook of the raw dataset has been provided which explains the meaning of each value of all attributes. Appendix IV shows the codebook of insurance coverage dataset which will be introduced later.

Table 4.4: Number of records in raw datasets

Data File	Number of Records
2006 Full Year Consolidated Data File	34,145
2007 Full Year Consolidated Data File	30,964
2008 Full Year Consolidated Data File	33,066
(Total)	**98,175**

4.3.2.2 Weighted and Cleaned Data

In the next step, we removed the zero-weighted samples which are not valid for analysis because they were not in-scope of the survey (civilian non-institutionalized) or didn't respond in the periods they were in-scope (missing in data), regarding MEPS's advice mentioned in pages C-2 and C-11 of [45]. The number of zero-weighted samples which were removed was 4,966 samples. In the raw data, due to the data file aggregation we did and the structure of panel surveys of MEPS, there are more than one sample available for some interviewees. It is better not to have duplicated records (equal samples with same IDs) in one dataset. But, in our dataset we don't have equal samples with same IDs. About the samples that have same IDs but different data, we can assume them as two different samples. Table 4.5 shows the total number of zero-weighted samples, plus number of equal samples in the dataset.

The next task is to remove incomplete samples after identifying the samples which contain missing data. In some cases, the interviewee wasn't sure about the answer (coded -7) or didn't answer the question (coded -8). Sometimes, the interviewer had made a mistake and didn't record the answer (coded -9). Even some samples weren't eligible to answer a

Table 4.5: Number of zero-weighted and equal samples

Title	Number of Records
Zero-weighted samples	4,966
Samples with same IDs	29,300
Equal samples with different IDs	3,250
Equal samples with same IDs	0

question for any reason (coded -1). Note that some attributes have "inapplicable" value (coded -1 or another number) which shows a specific group of people. For example, being a full-time student is only valid for samples aged 17 to 23, others are inapplicable and this doesn't mean this sample has a missing data.

Table 4.6: Missing data codes in all variables

Variable	Missing Data Code	Variable	Missing Data Code	Variable	Missing Data Code
DUPERSID	-	TTLP0#X	-	DIABDX	-7, -8, -9
REGION0#	-1	POVCAT0#	-	JTPAIN53	-7, -8, -9
MSA0#	-1	RTHLTH53	-1, -7, -8, -9	ARTHDX	-7, -8, -9
AGE0#X	-1	MNHLTH53	-1, -7, -8, -9	ASTHDX	-7, -8, -9
SEX	-	HIBPDX	-7, -8, -9	IADLHP53	-7, -8, -9
RACETHNX	-	CHDDX	-7, -8, -9	ADLHLP53	-7, -8, -9
MARRYO#X	-7, -8, -9	ANGIDX	-7, -8, -9	WLKLIM53	-7, -8, -9
FTSTU0#X	-8, -9	MIDX	-7, -8, -9	HAVEUS42	-1, -7, -8, -9
EDUCYR	-7, -8, -9	OHRTDX	-7, -8, -9	EMPST53	-7, -8, -9
HIDEG	-7, -8, -9	STRKDX	-7, -8, -9	INSJA0#X – INSDE0#X	-1
ACTDTY53	-1, -7, -8, -9	EMPHDX	-7, -8, -9		
HONRDC53	-1, -7, -8, -9	CHOLDX	-7, -8, -9	PERWT0#F	-
WAGEP0#X	-	CANCERDX	-7, -8, -9	YEAR	-

MEPS have certain coding for these cases and we can find the missing data using these coding. If a sample has even one of these missing data, it has been determined as an incomplete sample and was removed from the dataset. Overall, there were 5,959 incomplete samples which had any kind of missing data and were removed. In table 4.6, we listed the variables with their possible missing data codes. If a variable doesn't have a missing data code in its values, it means that whether that codes has another meaning for the variable or the variable doesn't have that kind of missing data.

After these operations on our dataset, we called this version as "Weighted and Cleaned data".

4.3.2.3 Original Data

By this stage, we prepared a clean dataset which doesn't have any missing data and all samples are valid. We snapped a version of this data as an "Original data" which can be used for different purposes. This dataset contains 87,250 valid records with MEPS's original values without any relabeling and attribute combination. For this study, we continue the preprocessing phase due to some changes in the attributes and their values.

Note that the codebook of the original dataset has been provided which explains the meaning of each value of all attributes. Appendix IV shows the codebook of insurance coverage dataset which will be introduced later.

4.3.2.4 Relabeled and Combined Data

In this part, we standardize the values of all attributes so that all have similar meanings. For instance, in some attributes, negative or false answer has been coded with 0; while in other attributes, it has been coded with 2. Thus, we relabeled some of the values so that a specific pattern exists. For flag (binary) variables the values were all relabeled to 0 and 1.

All of our attributes were in a categorical type of measurement except three of them. Hence, we converted the other continuous (numeric) variables – the age, wage, and total income variables – to categorical. Our criteria in binning the continuous variables were the usual statistical reports in the US. In table 4.7, we listed the relabeling items that applied on the variables. The values of other variables that don't exist in this table didn't change and were fine.

Table 4.7: Relabeling values of variables

Variable	**Relabeling rule**	**Variable**	**Relabeling rule**
AGE0#X	0-17→1, 18-40→2, 41-65→3, 66-85→4	STRKDX	2→0
		EMPHDX	2→0
WAGEP0#X	[0,-25K)→1, [25K,50K)→2, [50K,75K)→3, [75K,100K)→4, [100K,+∞)→5	CHOLDX	2→0
		CANCERDX	2→0
		DIABDX	2→0
TTLP0#X	(-∞,-25K)→1, [25K,50K)→2, [50K,75K)→3, [75K,100K)→4, [100K,+∞)→5	JTPAIN53	2→0
		ARTHDX	2→0
		ASTHDX	2→0
HIBPDX	2→0	IADLHP53	2→0
CHDDX	2→0	ADLHLP53	2→0
ANGIDX	2→0	WLKLIM53	2→0
MIDX	2→0	HAVEUS42	2→0
OHRTDX	2→0	INSJA0#X-INSDE0#X	2→0

In addition, we combined the heart disease indicators (CHDDX, ANGIDX, MIDX, and OHRTDX) to one variable (HEARTDS) with this criteria that if even one of the heart disease

indicators was diagnosed for a person, we considered him as a heart disease patient. Then, number of priority conditions for each sample was counted and summarized in one variable (PCCOUNT). In other words, all ten priority conditions were combined and summed into one variable which got values from zero to ten. We provided this attribute as a new constructed attribute in this study to explore the effect of number of priority conditions on healthcare coverage.

In all these processes, a person who didn't know or didn't answer about one or more of his priority condition indicators, or even wasn't eligible to answer due to his gender or age, we considered that he didn't suffer from that priority condition and didn't count that priority condition for him. Finally, we combined the monthly insurance coverage variables to one variable (INSCOV) with different criteria. After these operations on our dataset, we called this version as "Relabeled and Combined data".

4.3.3 Target Variable Selection

Insurance coverage is the target variable in this study. Each combined insurance coverage variable which was built based on different criteria and for different aims in the previous stage was separated to introduce a new dataset for itself. These different conditions for the target variable are explained in continue.

4.3.3.1 Nominal Target Variable

First, samples were divided into three classes (no-coverage, partial-coverage, and full-coverage) based on the numbers of months they've been covered during the last 12 months with this condition:

No-coverage: 0 month / Partial-coverage: 1-11 months / Full-coverage: 12 months

This dataset was called “Nominal target variable” and was used for the Nominal vs. Flag target variable test.

4.3.3.2 Threshold-*x* Flag Target Variable

Due to the weak results of nominal target variable, we designed a threshold analysis that will be discussed later in this document. For the threshold analysis part we needed 12 datasets with a flag target variable that their insurance coverage variable differed from each other. Depending on the number of threshold, the persons whose number of insured months in the last 12 months was greater or equal to the number of threshold were considered as insured and others as uninsured. Each of these datasets had their own names which represents their aim, such as “Threshold-1 flag target variable”, “Threshold-2 flag target variable”, and etc.

4.3.3.3 Insurance Coverage Dataset

In addition to the results of threshold analysis, we also explored for a definition on insurance coverage in MEPS. In the insurance coverage summary variable, it has been indicated that the persons who didn’t have coverage during the whole calendar year were known as uninsured and others were considered as insured. This definition of insurance coverage summary exists in page C-83 of [45]. Moreover, CPS which is a broadly cited statistical survey in the US, labels those who had no coverage for the whole calendar year as uninsured [9]. Actually, this definition is compatible with the results of threshold analysis and equals to the threshold-1 flag target variable dataset. We call this dataset the “Insurance Coverage dataset”.

Note that for each dataset – nominal target variable, threshold-1 target variable, and etc. – a codebook has been provided which explains the meaning of each value in all attributes.

Appendix IV shows the codebook of insurance coverage dataset which was used for the analysis of this research's main objective – efficient factors in healthcare coverage.

4.3.4 Data Balancing

To build a classification model using classical data mining techniques, all datasets should be balanced. Imbalance dataset would usually bias the majority class and the outcome model can predict samples from the minority class rarely [49]. Balancing a dataset based on a specific target variable means to have equal portions of samples for the values of the target variable. For instance, if the target variable has 3 values, one third of the samples should have target value of 1, one third should have target value of 2, and one third should have target value of 3.

There are several techniques to balance a dataset. We selected the random under-sampling method to apply due to its simplicity and reliability. The way that it works is very simple. Assume that we have 10,000 samples which 8,000 of them are insured and the others are not. To balance this dataset using random under-sampling technique, we should remove 6,000 samples from the insured samples randomly. Consequently, we will have two equal portions of insured and uninsured samples each containing 2,000 samples. There is a node in IBM SPSS Modeler called "Sample" that we can do data balancing using its stratification options. But, we did all the data balancing using a manual procedure in MS-Excel software. The data distribution results show that the random number generator of MS-Excel has worked properly in generating random numbers.

Finally, after balancing all of our datasets, the total number of records each has is reported in the table 4.8.

Table 4.8: Total number of records of all datasets

Dataset	Number of Records
Nominal target variable	31,770
Threshold-1 flag target variable	26,932
Threshold-2 flag target variable	27,678
Threshold-3 flag target variable	28,814
Threshold-4 flag target variable	30,298
Threshold-5 flag target variable	32,110
Threshold-6 flag target variable	33,962
Threshold-7 flag target variable	35,696
Threshold-8 flag target variable	37,848
Threshold-9 flag target variable	40,160
Threshold-10 flag target variable	42,600
Threshold-11 flag target variable	45,446
Threshold-12 flag target variable	48,112
Insurance Coverage Dataset	26,932

After data balancing, the Data Distribution and the Correlation of variables were checked with before data balancing so that there won't be any remarkable changed in the data distribution of the datasets. Note that the data distribution of all datasets didn't change a lot after being balanced. Table 4.9 shows the difference of data distribution of imbalanced and balanced insurance coverage datasets.

4.4 Attribute Reduction

In the data preparation, columns will also be explored and examined to be consistent. The columns or the attributes cannot look like whatever they are. Some of them should be

Table 4.9: Difference of data distribution of imbalanced and balanced insurance coverage datasets in percentage

Difference of Imbalanced and Balanced																								
values	REGI	MSA	AGEC	SEX	RACE	MAR	FTST	EDUC	HIDE	ACTI	HON	WAG	TTLP	POV	RTHL	MNH	PCCC	IADL	ADLH	WLK	HAVE	EMP	INSC	PERV	YEAR
-1	0	0	0	0	0	0	-2.7	-2.7	0	0	0	0	0	0	0	0	0	0	0	0	0	-5.6	0	0	0
0	0	0.2	0	0	0	0	0	-0.7	0	0	0	0	0	0	0	0	4.1	0.8	0.5	1.5	14	0	35	0	0
1	-2.3	-0.2	-5.9	1.9	9	0.1	-0.2	-0.1	6.7	0	-1.2	4.1	5.1	3.7	-1.8	-1.4	-0.1	-0.8	-0.5	-1.5	-14	4.3	-35	0	0
2	-2.7	0	8.6	-1.9	-1.1	-0.9	0.2	-0.1	0.8	10	6.9	-1.2	-1.6	1.4	-0.6	-0.2	-0.7	0	0	0	0	0	0	0	0
3	3.9	0	1.4	0	-0.6	0.8	2.6	-0.1	2.1	-5.6	-5.7	-1.7	-2.1	3.1	1.7	1.4	-1	0	0	0	0	1.1	0	0	0
4	1.1	0	-4.1	0	-7.2	0.6	0	-0.1	-1.8	-4.4	0	-0.7	-0.9	-0.9	0.8	0.1	-1	0	0	0	0	0.1	0	0	0
5	0	0	0	0	0	5	0	0.1	-1.2	0	0	-0.4	-0.6	-7.3	-0.1	0.1	-0.7	0	0	0	0	0	0	0	0
6	0	0	0	0	0	-5.6	0	1.2	-0.4	0	0	0	0	0	0	0	-0.5	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	-0	-0.4	0	0	0	0	0	0	0	-0.1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0.3	-5.8	0	0	0	0	0	0	0	-0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1.5	0	0	0	0	0	0	0	0	-0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	-0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	1.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	2.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	-0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	-0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	-1.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	-1.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.2
2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2
	0	-0	-0	0	0	0	0	-0	0	0	-0	0	-0	0	0	-0	0	-0	-0	0	0	0	0	0	-0

ignored for a data mining modeling. We discuss this matter as the initial attribute reduction we do for this study and apply it for all of our tests.

4.4.1 Column Ignorance Rules

There are some rules in data mining that should be applied on the columns. The outcome of these rules might be ignoring some columns (reducing attributes). Linoff and Berry [6] have listed these rules as following which believe these attributes are often not very useful:

1. Unary columns which have only one value.
2. Almost unary columns that around 95 to 99% of the column's value have only one value – almost unary.
3. Columns with unique values, such as person's identity number.
4. Synonym columns with the target attribute which are highly correlated with each other.

Table 4.10: Data distribution of imbalanced insurance coverage dataset in percentage

Value Appearance Rates of Imbalanced Dataset																								
values	REGI	MSA	AGEC	SEX	RACE	MAR	FTST	EDUC	HIDE	ACTI	HON	WAG	TTLP	POV	RTHL	MNH	PCCC	IADL	ADLH	WLK	HAVE	EMP	INSC	PERV	YEAR
-1	0	0	0	0	0	0	89	8.8	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0
0	0	16	0	0	0	0	0	4.3	0	0	0	0	0	0	0	0	54	97	98	91	22	0	15	0	0
1	15	84	29	47	27	39	5.1	2.1	18	0	5.9	74	69	20	30	40	18	2.8	1.5	8.9	78	45	85	0	0
2	20	0	30	53	18	4.5	0.9	2.1	3	59	67	15	18	6.4	32	30	10	0	0	0	0	0.1	0	0	0
3	38	0	30	0	4.8	7.8	5.1	2.3	31	25	27	6.1	7.2	17	27	24	7.3	0	0	0	0	3.2	0	0	0
4	27	0	10	0	51	1.8	0	2.1	10	16	0	2.4	2.9	29	8.8	5.1	5.1	0	0	0	0	27	0	0	0
5	0	0	0	0	0	22	0	2.2	3.9	0	0	1.9	2.5	28	2.7	1.2	3.1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	25	0	3.7	1.1	0	0	0	0	0	0	0	1.5	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	2.5	5.1	0	0	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	3.5	28	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	4.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	4.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	4.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	2.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	9.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	6.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32
2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34
	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	100

Table 4.11: Data distribution of balanced insurance coverage dataset in percentage

Value Appearance Rates of Balanced Dataset																								
values	REGI	MSA	AGEC	SEX	RACE	MAR	FTST	EDUC	HIDE	ACTI	HON	WAG	TTLP	POV	RTHL	MNH	PCCC	IADL	ADLH	WLK	HAVE	EMP	INSC	PERV	YEAR
1	0	0	0	0	0	0	86	6.2	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
0	0	16	0	0	0	0	0	3.6	0	0	0	0	0	0	0	0	58	98	99	93	35	0	50	0	0
1	13	84	23	49	36	39	4.9	2	25	0	4.7	78	74	24	29	38	18	2.1	1.1	7.4	65	49	50	0	0
2	17	0	39	51	16	3.6	1.1	2.1	3.8	69	74	14	17	7.8	31	29	9.4	0	0	0	0	0.1	0	0	0
3	42	0	32	0	4.3	8.7	7.7	2.2	33	20	21	4.4	5.1	20	28	26	6.3	0	0	0	0	4.3	0	0	0
4	28	0	6.4	0	43	2.4	0	2	8.3	12	0	1.7	2	28	9.6	5.2	4.1	0	0	0	0	27	0	0	0
5	0	0	0	0	0	27	0	2.3	2.7	0	0	1.4	1.9	21	2.6	1.3	2.3	0	0	0	0	0	0	0	0
6	0	0	0	0	0	20	0	4.8	0.8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	2.4	4.7	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	3.8	22	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	5.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	4.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	6.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	4.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	7.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	2.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	4.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31
2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34
	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	100

To check these rules for our datasets, we built Data Distribution Table and Correlation Matrix for both imbalanced and balanced versions of each dataset. In the data distribution

table we calculated the percentage of each value in each attribute to check rules number 1 and 2. There wasn't any unary columns in our datasets, but the limitation attributes (IADLHP53, ADLHLP53, and WLKLIM53) were almost unary. Although we could have used rule number 2 to remove the limitation attributes, but we included them in our datasets with this idea that they might have significant impacts on healthcare coverage. In tables 4.10 and 4.11, the data distribution of imbalanced and balanced insurance coverage datasets are displayed.

The correlation matrix also showed that there isn't any attribute correlated with the target attribute (INSCOV0#) – rule number 4. The correlation degree is between -1 to 1. Highly correlated variables will have a correlation degree near 1 or -1 (if their trend is in reverse), and they are uncorrelated if their correlation degree is around 0. As a matter of fact, one of the reasons we didn't select the total medical expenditure variable from MEPS databases was that we believe it will have high correlation with healthcare coverage. The logic is very simple: the more amounts of medical costs, the more possibility of being insured. In other words, whoever has considerable amount of medical costs will surely have healthcare coverage, and people with very low medical costs might have coverage and might not. This means that a high portion of this attribute is correlated with the target variable.

Although we didn't have any attribute which was highly correlated with the target variable, but there were two attributes which had a lot of dependency to each other. In other words, one of them had been derived from the other one. These two attributes are the wage income and the person-level total income. They have a correlation degree of more than 0.9. Therefore, we decided to calculate the sum of absolute correlation values that each of them have with all other attributes. This number shows to some extent each attribute is independent related to all other attributes. Eventually, total income variable was more

dependent to other variables than the wage variable. Logically, we can also say that the wage variable is more independent than the income variable, because the total income variable includes the wage income in its definition. But, we didn't remove the total income variable in the data preprocessing phase, and kept it as the first candidate to be removed in the attribute reduction procedure. Another achievement from this sum is that the age variable is the most dependent variable to all other variables which is reasonable. Correlation matrices of threshold-1 dataset for both imbalanced and balanced version have been displayed in table 4.12. As you can see in this table, the correlation of variables didn't change a lot after data balancing.

For rule number 3, we had these kinds of attributes such as person ID, person-level weight, and year. Therefore, we didn't include them in our modeling.

4.4.2 Principal Component Analysis (PCA)

PCA is a mathematical procedure to convert a set of correlated variables to a set of uncorrelated variables called principal components. It is a kind of data-reduction technique to reduce the complexity of data. For example, when you have 22 input variables and you want to reduce to principal components, this procedure will introduce a number of factors (e.g. 5 factors) built using a linear combination of the original inputs. In other words, these 5 uncorrelated factors have been derived from the original 22 correlated variables.

Although principal component analysis seems to reduce the attributes to a minimum number, actually it isn't applicable for our study. The reason is that the principal components are not real-world variables; they are a linear combination of the real-world variables. In other words, each principal component does not represent a specific variable that can be found in the real world. Therefore, our attempt to reduce the attributes using the PCA method was

useless because we needed a smaller number of real-world variables, not constructed variables.

Table 4.12: Correlation matrices of insurance coverage datasets

Correlation of Imbalanced Dataset

	REGIC	MSAC	AGEO	SEX	RACE	MARR	FTSTU	EDUC	HIDE	ACTD	HONR	WAGI	TTLPC	POVC	RTHL	MNHI	PCCO	IADLH	ADLH	WLKL	HAVE	EMPS	INSCC	Total
REGIC	1	0.03	-0	-0	-0.2	0.02	0.02	-0.1	-0	-0	0.03	-0	-0	-0.1	0.01	0	-0.1	-0	0	-0	-0.1	-0	-0.1	1.86
MSAC	0.03	1	-0	-0	-0.1	0.04	0	0.02	0.03	-0.1	0.02	0.07	0.06	0.08	-0.1	-0.1	-0.1	-0	-0	-0.1	-0	-0	0	1.92
AGEO	-0	-0	1	0.05	0.2	-0.7	-0.1	0.62	-0.6	0.2	-0.7	0.22	0.31	0.19	0.33	0.22	0.63	0.19	0.14	0.35	0.02	0.62	0.01	7.62
SEX	-0	-0	0.05	1	-0	-0	-0	0.04	-0	0	0.06	-0.1	-0.1	-0.1	0.05	0.03	0.05	0.04	0.02	0.07	0.08	0.11	0.04	2.03
RACE	-0.2	-0.1	0.2	-0	1	-0.2	-0.1	0.26	0.01	0.07	-0.2	0.18	0.22	0.33	-0	-0	0.18	0.03	0.01	0.1	0.14	0.08	0.19	3.82
MARR	0.02	0.04	-0.7	-0	-0.2	1	0.18	-0.6	0.56	0.08	0.68	-0.3	-0.4	-0.3	-0.2	-0.1	-0.4	-0.1	-0	-0.2	0.03	-0.5	0.03	6.69
FTSTU	0.02	0	-0.1	-0	-0.1	0.18	1	0.11	-0.2	-0.3	-0.1	-0.1	-0.1	-0.1	-0.1	-0	-0.2	-0	-0	-0.1	-0.2	0.17	-0.1	3.32
EDUC	-0.1	0.02	0.62	0.04	0.26	-0.5	0.11	1	-0.5	-0.3	-0.8	0.42	0.46	0.33	0.13	0.07	0.29	0.03	0.01	0.11	-0.1	0.53	-0	6.79
HIDEC	-0	0.03	-0.6	-0	0.01	0.56	-0.2	-0.5	1	0.22	0.73	-0.1	-0.1	-0	-0.3	-0.2	-0.3	-0.1	-0.1	-0.2	0.18	-0.6	0.2	6.41
ACTD	-0	-0.1	0.2	0	0.07	0.08	-0.3	-0.3	0.22	1	0.16	-0.2	-0.1	-0	0.06	0.04	0.32	0.15	0.12	0.22	0.22	-0	0.2	4.13
HONR	0.03	0.02	-0.7	0.06	-0.2	0.68	-0.1	-0.8	0.73	0.16	1	-0.3	-0.3	-0.2	-0.2	-0.2	-0.4	-0.1	-0.1	-0.2	0.11	-0.6	0.09	7.33
WAGI	-0	0.07	0.22	-0.1	0.18	-0.3	-0.1	0.42	-0.1	-0.2	-0.3	1	0.92	0.44	-0.1	-0.1	0.04	-0.1	-0.1	-0.1	0	-0.1	0.11	5.03
TTLPC	-0	0.06	0.31	-0.1	0.22	-0.4	-0.1	0.46	-0.1	-0.1	-0.3	0.92	1	0.49	-0.1	-0.1	0.11	-0.1	-0	-0	0.02	-0	0.12	5.23
POVC	0.1	0.08	0.19	0.1	0.33	0.3	0.1	0.33	0	0	0.2	0.44	0.49	1	0.1	0.1	0.06	0.1	0	0.1	0.09	0	0.17	4.35
RTHLT	0.01	-0.1	0.33	0.05	-0	-0.2	-0.1	0.13	-0.3	0.06	-0.2	-0.1	-0.1	-0.1	1	0.66	0.45	0.25	0.19	0.38	0.04	0.31	-0	5.12
MNHI	0	-0.1	0.22	0.03	-0	-0.1	-0	0.07	-0.2	0.04	-0.2	-0.1	-0.1	-0.1	0.66	1	0.29	0.22	0.17	0.26	0.02	0.24	-0	4.18
PCCO	-0.1	-0.1	0.63	0.05	0.18	-0.4	-0.2	0.29	-0.3	0.32	-0.4	0.04	0.11	0.06	0.45	0.29	1	0.27	0.19	0.5	0.18	0.44	0.11	6.54
IADLH	-0	-0	0.19	0.04	0.03	-0.1	-0	0.03	-0.1	0.15	-0.1	-0.1	-0.1	-0.1	0.25	0.22	0.27	1	0.6	0.43	0.06	0.21	0.05	4.06
ADLH	0	-0	0.14	0.02	0.01	-0	-0	0.01	-0.1	0.12	-0.1	-0.1	-0	-0	0.19	0.17	0.19	0.6	1	0.34	0.05	0.15	0.04	3.4
WLKL	-0	-0.1	0.35	0.07	0.1	-0.2	-0.1	0.11	-0.2	0.22	-0.2	-0.1	-0	-0.1	0.38	0.26	0.5	0.43	0.34	1	0.09	0.32	0.06	5.11
HAVE	-0.1	-0	0.02	0.08	0.14	0.03	-0.2	-0.1	0.18	0.22	0.11	0	0.02	0.09	0.04	0.02	0.18	0.06	0.05	0.09	1	-0.1	0.34	3.1
EMPS	-0	-0	0.62	0.11	0.08	-0.5	0.17	0.53	-0.6	-0	-0.6	-0.1	-0	-0	0.31	0.24	0.44	0.21	0.15	0.32	-0.1	1	-0.1	6.24
INSCC	-0.1	0	0.01	0.04	0.19	0.03	-0.1	-0	0.2	0.2	0.09	0.11	0.12	0.17	-0	-0	0.11	0.05	0.04	0.06	0.34	-0.1	1	3.14

Correlation of Balanced Dataset

	REGIC	MSAC	AGEO	SEX	RACE	MARR	FTSTU	EDUC	HIDE	ACTD	HONR	WAGI	TTLPC	POVC	RTHLT	MNHI	PCCO	IADLH	ADLH	WLKL	HAVE	EMPS	INSCC	Total
REGIC	1	0.03	-0	-0	-0.2	-0	0.03	-0.1	-0	-0	0.02	-0	-0	-0.1	0.03	0.02	-0.1	-0	-0	-0	-0.1	-0	-0.1	2.03
MSAC	0.03	1	-0	-0	-0.2	0.03	-0	0.01	0.02	-0	0.02	0.05	0.05	0.06	-0.1	-0.1	-0.1	-0	-0	-0	-0	-0	0	1.88
AGEO	-0	-0	1	0.05	0.16	-0.7	-0.1	0.56	-0.6	0.12	-0.7	0.21	0.27	0.15	0.32	0.21	0.57	0.17	0.12	0.31	0.01	0.54	0.01	6.97
SEX	-0	-0	0.05	1	0	-0	-0	0.03	0	0.02	0.06	-0.1	-0.1	-0.1	0.06	0.04	0.06	0.04	0.03	0.07	0.11	0.13	0.05	2.15
RACE	-0.2	-0.2	0.16	0	1	-0.1	-0.1	0.26	0.1	0.11	-0.1	0.19	0.22	0.3	-0	-0	0.2	0.05	0.02	0.12	0.2	0.03	0.26	3.92
MARR	0	0.03	0.7	0	0.1	1	0.22	0.5	0.51	0.11	0.61	0.3	0.3	0.2	0.2	0.1	0.3	0.1	0	0.1	0.03	0.4	0.04	5.93
FTSTU	0.03	-0	-0.1	-0	-0.1	0.22	1	0.12	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0	-0.2	-0	-0	-0.1	-0.2	0.16	-0.1	3.4
EDUC	-0.1	0.01	0.56	0.03	0.26	-0.5	0.12	1	-0.4	-0.3	-0.7	0.36	0.39	0.26	0.1	0.05	0.24	0.02	0	0.1	-0.1	0.46	-0	6.07
HIDEC	0	0.02	0.6	0	0.1	0.51	0.2	0.4	1	0.32	0.71	0	0	0.06	0.3	0.2	0.2	0.1	0	0.1	0.23	0.6	0.29	6.04
ACTD	-0	-0	0.12	0.02	0.11	0.11	-0.2	-0.3	0.32	1	0.24	-0.1	-0.1	0.03	0.03	0	0.29	0.14	0.12	0.19	0.26	-0.1	0.3	4.22
HONR	0.02	0.02	-0.7	0.06	-0.1	0.61	-0.1	-0.7	0.71	0.24	1	-0.2	-0.3	-0.1	-0.2	-0.2	-0.4	-0.1	-0	-0.2	0.13	-0.6	0.14	6.85
WAGI	-0	0.06	0.21	-0.1	0.19	-0.3	-0.1	0.36	-0	-0.1	-0.2	1	0.93	0.43	-0.1	-0.1	0.05	-0	-0	-0.1	0.04	-0.1	0.17	4.78
TTLPC	-0	0.05	0.27	-0.1	0.22	-0.3	-0.1	0.39	-0	-0.1	-0.3	0.93	1	0.48	-0.1	-0.1	0.12	-0	-0	-0	0.07	-0.1	0.19	4.95
POVC	-0.1	0.06	0.15	-0.1	0.3	-0.2	-0.1	0.26	0.06	0.03	-0.1	0.43	0.48	1	-0.1	-0.1	0.06	-0	-0	-0	0.13	-0.1	0.24	4.17
RTHLT	0.03	-0.1	0.32	0.06	-0	-0.2	-0.1	0.1	-0.3	0.03	-0.2	-0.1	-0.1	-0.1	1	0.67	0.44	0.2	0.15	0.35	0.04	0.28	-0.1	4.84
MNHI	0.02	-0.1	0.21	0.04	-0	-0.1	-0	0.05	-0.2	0	-0.2	-0.1	-0.1	-0.1	0.67	1	0.28	0.19	0.15	0.24	0.02	0.21	-0	3.97
PCCO	-0.1	-0.1	0.57	0.06	0.2	-0.3	-0.2	0.24	-0.2	0.29	-0.4	0.05	0.12	0.06	0.44	0.28	1	0.26	0.18	0.5	0.22	0.34	0.16	6.19
IADLH	-0	-0	0.17	0.04	0.05	-0.1	-0	0.02	-0.1	0.14	-0.1	-0	-0	-0	0.2	0.19	0.26	1	0.55	0.39	0.07	0.17	0.09	3.72
ADLH	-0	-0	0.12	0.03	0.02	-0	-0	0	-0	0.12	-0	-0	-0	-0	0.15	0.15	0.18	0.55	1	0.31	0.06	0.12	0.07	3.12
WLKL	-0	-0	0.31	0.07	0.12	-0.1	-0.1	0.1	-0.1	0.19	-0.2	-0.1	-0	-0	0.35	0.24	0.5	0.39	0.31	1	0.12	0.25	0.09	4.73
HAVE	-0.1	-0	0.01	0.11	0.2	0.03	-0.2	-0.1	0.23	0.26	0.13	0.04	0.07	0.13	0.04	0.02	0.22	0.07	0.06	0.12	1	-0.1	0.4	3.61
EMPS	0	0	0.54	0.13	0.03	0.4	0.16	0.46	0.6	0.1	0.6	0.1	0.1	0.1	0.28	0.21	0.34	0.17	0.12	0.25	0.1	1	0.1	5.82
INSCC	-0.1	0	0.01	0.05	0.25	0.04	-0.1	-0	0.29	0.3	0.14	0.17	0.19	0.24	-0.1	-0	0.16	0.09	0.07	0.09	0.4	-0.1	1	4.01

To have some feeling about the data that has been prepared, figure 4.5 shows snapshots of first step of our raw data and the last step of prepared data which is completely understandable and can be used for any modeling purposes.

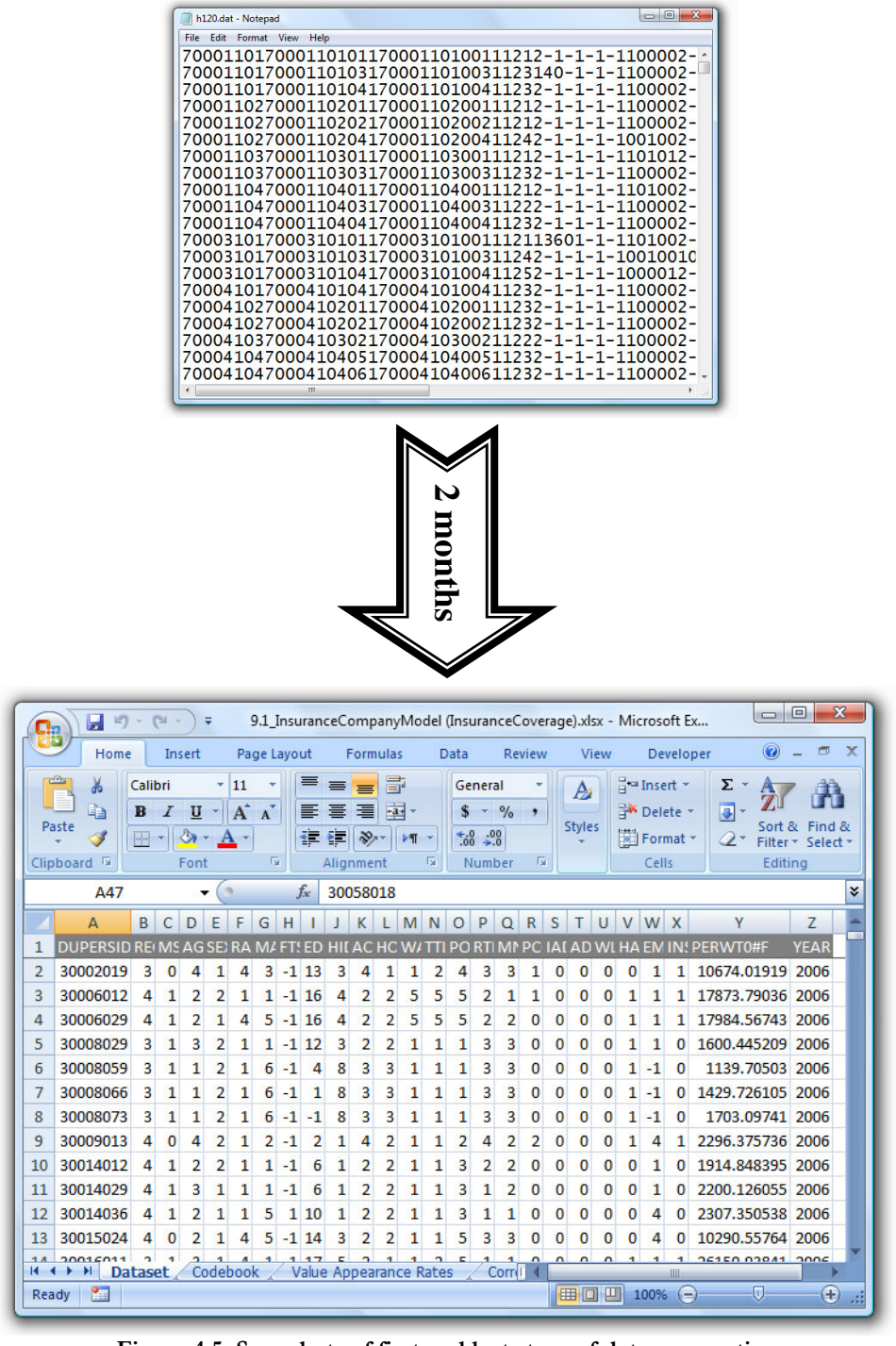


Figure 4.5: Snapshots of first and last steps of data preparation

4.5 Summary

Data preparation is an essential phase for all data mining studies. Different operations might be applied to the dataset based on the needs and objectives. The main items we presented and prepared in this chapter are as following:

- Medical Expenditure Panel Survey (MEPS) data files
- Selecting attributes from data files
- Converting data files from ASCII file format to CSV file format
- Cleaning the data: removing missing data, relabeling values, and combining attributes
- Datasets with different target variables
- Balanced and imbalanced Insurance Coverage dataset
- Initial attribute reduction procedure

5 HEALTHCARE COVERAGE MODELING

Modeling the healthcare coverage with different aims and employing various techniques has been done in the past studies. In this study, we aim to discover the efficient factors in healthcare coverage as our main objective. Other findings and analyses will also outcome during this study which is presented beside the main results. In this chapter, first we briefly introduce the data mining modeler tool that we employed; second, the models that we proposed are explained; and finally, outcome models and their results are analyzed and discussed.

5.1 IBM SPSS Modeler

IBM SPSS Modeler is a data mining, modeling, and reporting tool which allows us to build predictive and descriptive models very quickly and easily using a good graphical interface. We employed the version 14.1 of this software in this project which was the latest one. To some aspects it works pretty similar to other popular data mining applications such as WEKA. The big advantage of this product is having a very user friendly interface which helps the users to find their desired tools to enhance their modeling a lot. However, there are some rewordings that the most important ones we should consider are “segmentation” and “field”. The word “segmentation” in IBM SPSS Modeler means “clustering” which is the well-known unsupervised learning in data mining. “Field” is also a rewording for “attribute” or “variable” which describes each column of a dataset that has a specific meaning, such as ID, age, sex, and etc.

5.1.1 Stream of Nodes

Operating data mining tools in IBM SPSS Modeler is via flow of data through streams. We have several types of nodes that each of them does a specific operation on our data. The nodes will be connected to each other to build a stream. Each stream builds a specific model and reports the related results. Nodes in IBM SPSS Modeler are divided into four types: Source, Process, Modeling, and Output. Note that the order of different types of nodes should be considered as a rule in streams.

5.1.1.1 Source Nodes

Source nodes represent the actual data that we are going to work on. Each source node allocates a specific data file to itself. It reads all the data of that data file and prepares for operation. Source nodes can be from text files (VAR node), excel files, and different database file formats. From the user's perspective, we set the configuration of source node just once, and then whenever we need our data source, we only work with this one node. In IBM SPSS Modeler, all properties of a node can be set by double-clicking on that node and changing its specifications.

5.1.1.2 Process Nodes

The second type of nodes is the process nodes. They are separated to Records operations and Field operations. Using these nodes we can apply some processes on our data to prepare them for modeling. It includes important operations for data preprocessing, such as:

- Type node: Determining the type and role of each field.
- Filter node: Allows fields to be renamed or removed.
- Binning node: Converts numeric fields to sets (categories).
- Partition node: Split the data into separate subsets (training and testing sets).

- Select node: Selects a subset of records based on a specified condition.
- Sample node: Provides different methods for sampling data.
- Balance node: Corrects imbalances using specified conditions.

Different data preprocessing streams can be built for one specific data source which will end up in different data mining models. Note that the measurement type of dataset's fields should be determined initially in the data preprocessing section using the Type node. The measurement types are as following:

- Continuous: Integer, real number, date/time, etc.
- Categorical: String values (number of them is unknown) – IBM SPSS Modeler will automatically change this type to one of flag, nominal, or ordinal types.
- Flag: Two distinct values (true/false)
- Nominal: Multiple distinct values (red/blue/green) – 250 set maximum
- Ordinal: Multiple distinct values have an inherent order (children/youth/adults/elders)
- Typeless: No role (person's ID)

In addition, we should determine each field's role in this node. For classification modeling we need at least one field with the role of target. Other fields should be set as inputs. The field roles are as following:

- Input: Predictor.
- Target: Output.
- Both: Predictor and output together (only for Apriori modeling node).
- None: Ignored.
- Partition: Manual train-test set.
- Split: Building separate models for each value of this field.

- Frequency: Frequency weighted factor (for C&RT, CHAID, … modeling nodes)
- Record ID: Only for linear modeling, others ignore it.

Table 5.1 lists the attributes of our final datasets with their measurements and roles.

5.1.1.3 Modeling Nodes

Modeling nodes are divided into three categories: Association, Classification, and Segmentation. Each modeling node presents a data mining technique or algorithm which builds a predictive or descriptive model. We can set the different configurations for each modeling node. For instance, a K-Means clustering node can have the value of 2 for the number of K, and the other can have 5, and so on.

There are also some nodes in the modeling section called "AutoClassifier" and "AutoCluster" which classifies and clusters the dataset using different techniques and ranks them based on their accuracy. These nodes can be used initially to explore different algorithms and see which algorithm works better on a specific dataset. But for more detailed results and analysis, it's better to use the specific algorithm's modeling node.

5.1.1.4 Output Nodes

After we run a stream, a new model node will be built. This model is the outcome of our modeling. It shows the specific trained model. We can see the results and analyses in different ways using Output nodes such as charts, tables, graphical diagrams, and etc. These nodes can be found in the Graph, Export, and Output tabs of IBM SPSS Modeler. For instance, to get the accuracy of outcome models we should use the Analysis node. Note that some results of the outcome model can be seen inside the built model node and doesn't need any kind of output node, such as the predictor importance.

Table 5.1: Measurements and roles of the variables

Variable	Measurement	Role in Classification	Role in Clustering
DUPERSID	Typeless	None	None
REGION0#	Nominal	Input	Input
MSA0#	Flag	Input	Input
AGE0#X	Ordinal	Input	Input
SEX	Flag	Input	Input
RACETHNX	Nominal	Input	Input
MARRYO#X	Nominal	Input	Input
FTSTU0#X	Nominal	Input	Input
EDUCYR	Ordinal	Input	Input
HIDEG	Ordinal	Input	Input
ACTDTY53	Nominal	Input	Input
HONRDC53	Nominal	Input	Input
WAGEP0#X	Ordinal	Input	Input
TTLP0#X	Ordinal	Input	Input
POVCAT0#	Ordinal	Input	Input
RTHLTH53	Ordinal	Input	Input
MNHLTH53	Ordinal	Input	Input
PCCOUNT	Ordinal	Input	Input
IADLHP53	Flag	Input	Input
ADLHLP53	Flag	Input	Input
WLKLIM53	Flag	Input	Input
HAVEUS42	Flag	Input	Input
EMPST53	Nominal	Input	Input
INSCOV0#X	Flag/Nominal	Target	Input
PERWT0#F	Typeless	None	None
YEAR	Typeless	None	None

To sum up, we can see a typical stream of nodes in IBM SPSS Modeler regarding their order in figure 5.1.

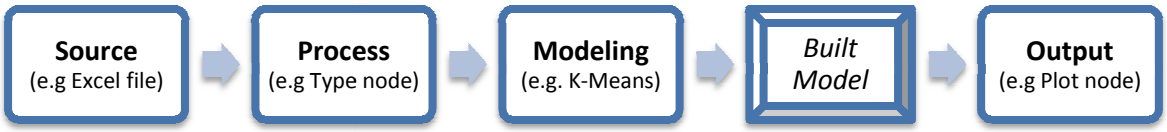


Figure 5.1: Stream of nodes in IBM SPSS Modeler

5.2 Proposed Models

The models that we proposed in this study are in two parts: unsupervised learning and supervised learning. In the unsupervised learning, clustering techniques were applied to achieve descriptive information about the insurance coverage dataset. On the other hand, the aim in the supervised learning is to build a predictive model to predict the healthcare coverage status and also discover the most important attributes in this modeling. In this section, we introduce our proposed models. The analyses of the outcomes are presented in the next section.

5.2.1 Unsupervised Learning

In the first step of the modeling, an unsupervised learning process is applied on the dataset to discover various hidden groups. This task helps the researchers to understand their dataset and different populations that exist in their dataset better. Before building the healthcare coverage prediction models as the supervised learning, we employ K-means clustering algorithm on the imbalanced insurance coverage dataset that we prepared in the data preparation section. K-Means clustering is a method which tries to partition the samples into K clusters in which each sample fit in the cluster with the nearest mean.

Two unsupervised learning models are proposed in this study: whole population clustering, uninsured population clustering. In the whole population clustering, the whole dataset is

clustered into several groups and the important characteristics of each cluster are presented. The aim of this modeling is to find the main groups of people in our sampling population which is a national representative data. The uninsured population clustering is done after the supervised learning to provide additional information about the people without health insurance. Therefore, only the uninsured population of the dataset is clustered. Introducing the main groups of uninsured people and their significant features is the goal of this modeling.

5.2.2 Supervised Learning

The aims of models built in the supervised learning section are to classify and predict the healthcare coverage status of samples, and rank the attributes based on their importance. The insurance coverage variable is the target variable in the classification models. The balanced insurance coverage dataset that we prepared in the data preparation section is used for classification models. Having a balanced dataset helps the models to provide unbiased outcomes.

As a supervised learning, both decision tree and neural network modeling are applied on the prepared dataset. We employ C5.0 algorithm as the decision tree modeling and Multi-Layer Perceptron (MLP) algorithm as the neural network modeling. Different procedures were applied in building these models which are explained in the next section. Before, we explain the operating features of C5.0 and MLP algorithms briefly in continue.

C5.0 employs a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. In the tree structure, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. C5.0 is significantly faster and more memory efficient than C4.5. Other advantages of C5.0

compared to C4.5 are building smaller decision trees, support for boosting, weighting different cases, and winnowing the attributes.

A multilayer perceptron (MLP) is a feed-forward artificial neural network model that will map input nodes onto the output node(s) using multiple layers of nodes in a directed graph which each layer is fully connected to the next one. Each node – except the input nodes – is a neuron (or processing element) with a nonlinear activation function. MLP uses back-propagation for training the network which is a supervised learning technique and can distinguish data that is not linearly separable. It is a modification of the standard linear perceptron.

Figure 5.2 shows the stream of nodes that we used in IBM SPSS Modeler to build and analyze our desired classification models. Note that in the C5.0 modeling node, it has a setting for cross-validation which we selected a 10-fold cross validation for all tests; but in the MLP (Neural Network) modeling node, it doesn't have any setting for cross-validation, therefore we applied a 70-30 train-test set partitioning using a partition node before modeling.

5.3 Analysis of Results

In this section we report the results of all tests and discuss their results. Each outcome is a chain of findings stream in our study. They are all related to each other, and each of them separately gives new information and output. In other words, each of these outcomes has their own analysis and also the whole procedure and stream has an overall analysis that we discuss them all in this section. Results and analysis of the two unsupervised learning models, and also the supervised models are reported and discussed in continue.

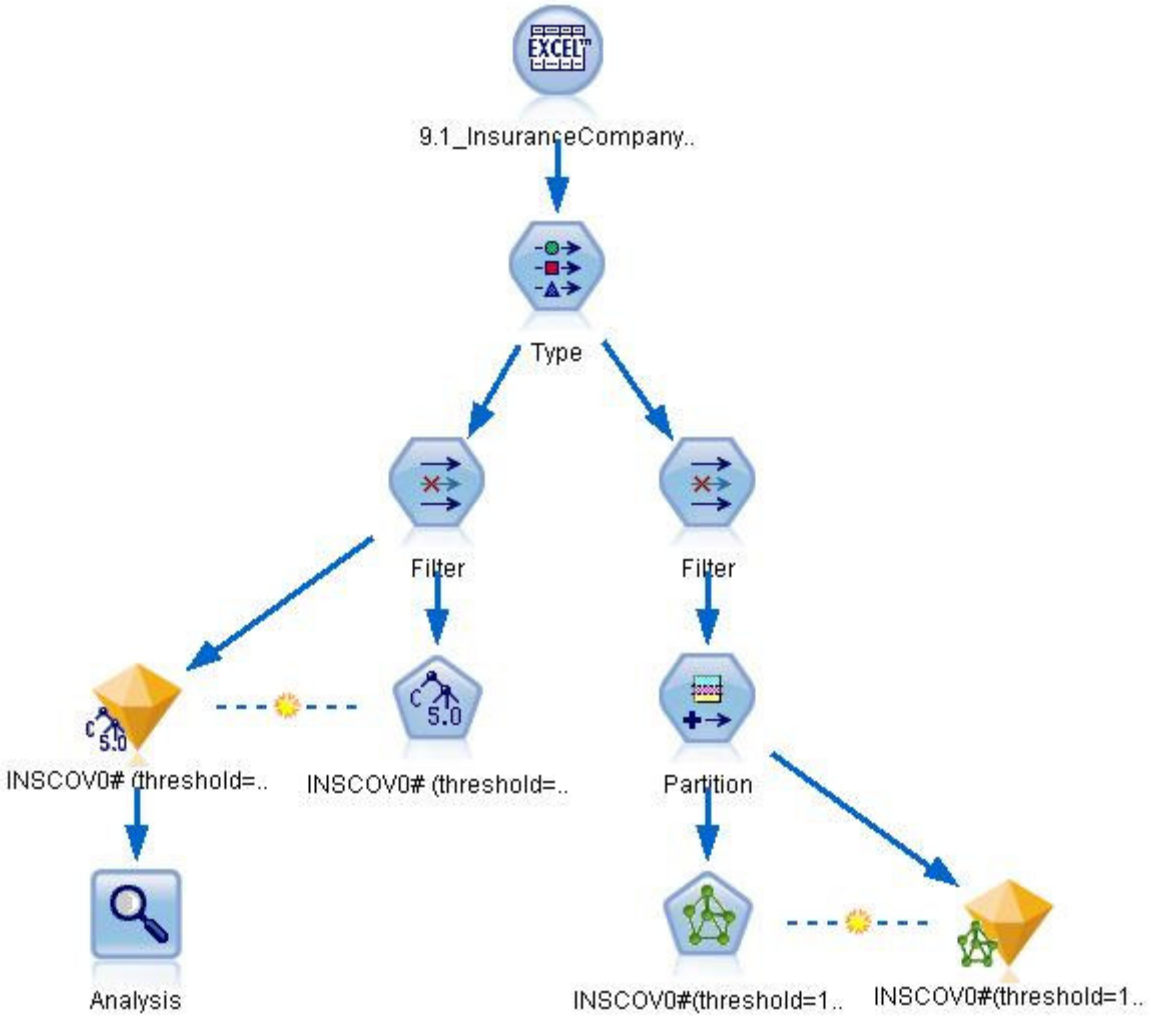


Figure 5.2: Classification streams of nodes

5.3.1 Nominal vs. Flag: Insurance Coverage Measurement

In most of the related works, healthcare coverage was reported as a binary variable: true or false. This test was done initially and before the unsupervised learning modeling to experiment whether we can build healthcare coverage prediction model by having a 3-value target variable or not. As we mentioned in the previous chapter, first we had a nominal target variable with three class labels called “no-coverage”, “partial-coverage”, and “full-coverage”. No-coverage means that a person didn’t have healthcare coverage for even one

month in the last 12 months. Partial-coverage represents the samples that had between 1 to 11 months coverage in the last 12 months. And finally, full-coverage is for the people who had complete coverage during the last 12 months.

C5 (decision tree) and MLP (neural network) algorithms were applied to build the classification models. Although we balanced the dataset (nominal target variable dataset), the results were not acceptable. Thus, the predictor importance results for ranking attributes were not reliable. The problem was with the partial-coverage group in which the samples didn't have much similarity with each other. Therefore, we had to assign them to one of the two other groups: no-coverage or full-coverage.

The sensitivity of partial-coverage samples was around 40%. This means that the outcome model can predict a sample which has partial-coverage as a partial-coverage with only 40% probability. Thus, with 60% probability it will predict the sample as either a no-coverage or full-coverage sample. This result shows that the samples which are grouped and labeled as partial-coverage do not have a similar pattern with each other. In other words, they shouldn't be grouped in one class. As a result, we found from this test that we cannot use three pre-defined classes for healthcare coverage with the labeling that we had due to the disparity in the partial-coverage population and it's better to use a flag labeling for healthcare coverage.

Populations of samples which have from 1 month to 11 months coverage are from 0.5% to 1.5% of the total population. Figure 5.3 shows the population of each group. If we want to have a 3-value target variable, the middle label includes a set of very small portion of the population with no relevance in their behaviors. Note that the two side of this middle section (samples having from 1 to 11 months coverage) are near to no-coverage and full-coverage sections and they do not have similar patterns and behaviors. Therefore, the partial-coverage cannot be well predicted.

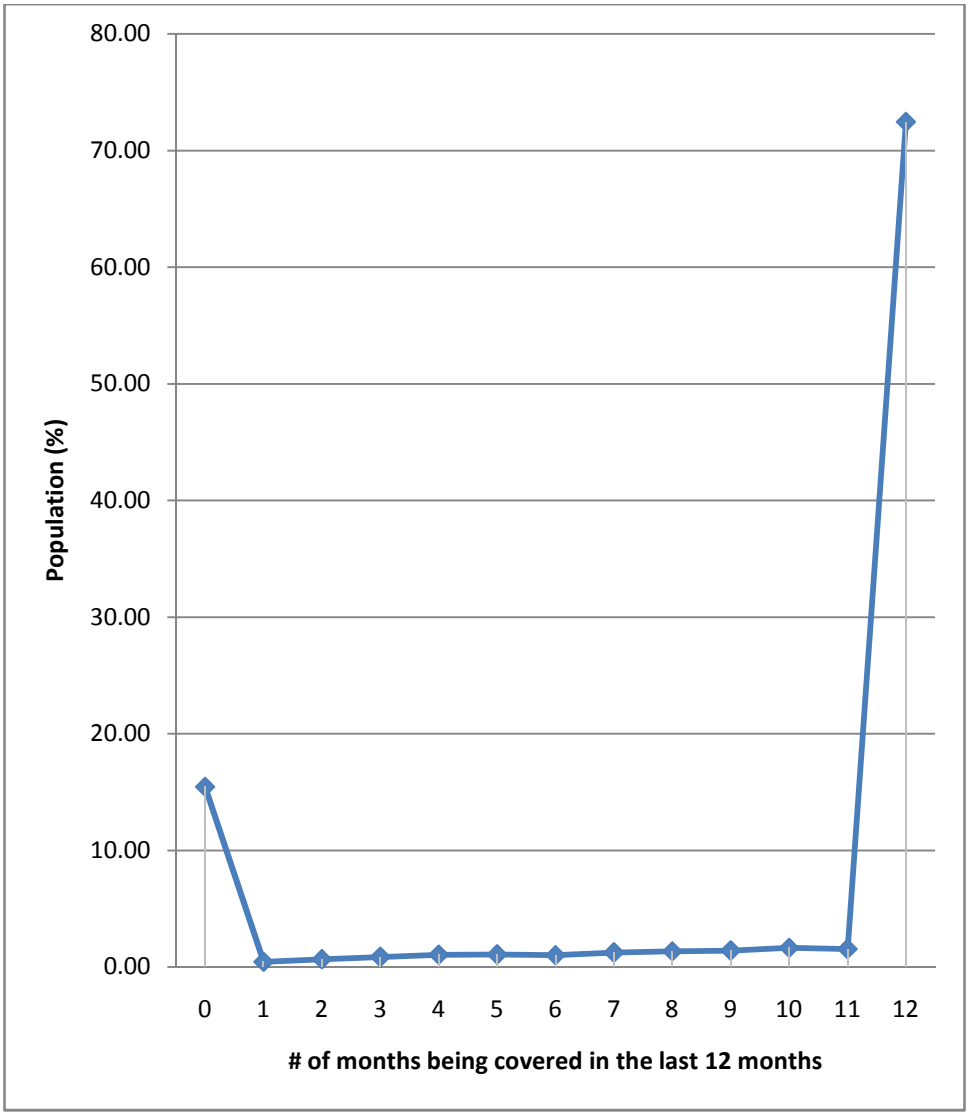


Figure 5.3: Healthcare coverage distribution based on number of months

Due to the small size of partial-coverage's population, the solution is to divide them and merge them with the other two existing classes. To find the border of this division, it's better to experiment different number of months as threshold.

5.3.2 Insurance Coverage Threshold Analysis

Because of the weak results of predicting healthcare coverage using the nominal target variable dataset, a threshold analysis test was designed to examine the healthcare coverage

threshold. In other words, we want to see having at least how many months coverage in the last 12 months can be indicated as being insured in a binary healthcare coverage status.

We argue that by finding a number as threshold, the samples which have coverage months more than this number in the last 12 months have similar behavior and pattern, and also for the samples which have coverage months less than the threshold. We evaluate this statement by classifying different threshold datasets with this assumption that the higher classification accuracy rate, the better similar pattern among samples. In other words, whenever the accuracy of classifier model peaks during the test, that point is the best threshold to divide the partial-coverage section.

We built 12 datasets with a binary healthcare coverage status which represent thresholds from 1 to 12 months. For instance, the threshold-1 dataset means that samples that had at least 1 month coverage during the last 12 months has been considered as insured. In addition to C5 (decision tree) and MLP (neural network) algorithms, C&RT (classification and regression trees) and Logistic Regression techniques were also applied to build models which predict the two-label healthcare coverage. Having more analytical results can help us in finding the general trend of healthcare coverage threshold.

Four different algorithms were employed to model 12 healthcare coverage prediction datasets with different healthcare coverage thresholds. In addition to the correctness accuracy rate, we also calculated the g-mean accuracy rate which is more sensitive about biased models. But most of the outcome models were unbiased and therefore their correctness accuracy rate was equal to their g-mean accuracy rate. For more information about different types of accuracy rates in classification models you can refer to the measurement methods section in the Introduction chapter.

Figures 5.4 to 5.7 show the results of these four modeling techniques in finding a trend for healthcare coverage threshold. All the methods show a descending trend with some exception in their results. C5 was the best classifier with higher accuracy rates. Others seem to have pretty the same results. MLP has more oscillations than others, and had more difficulties in finding a trend for different threshold patterns.

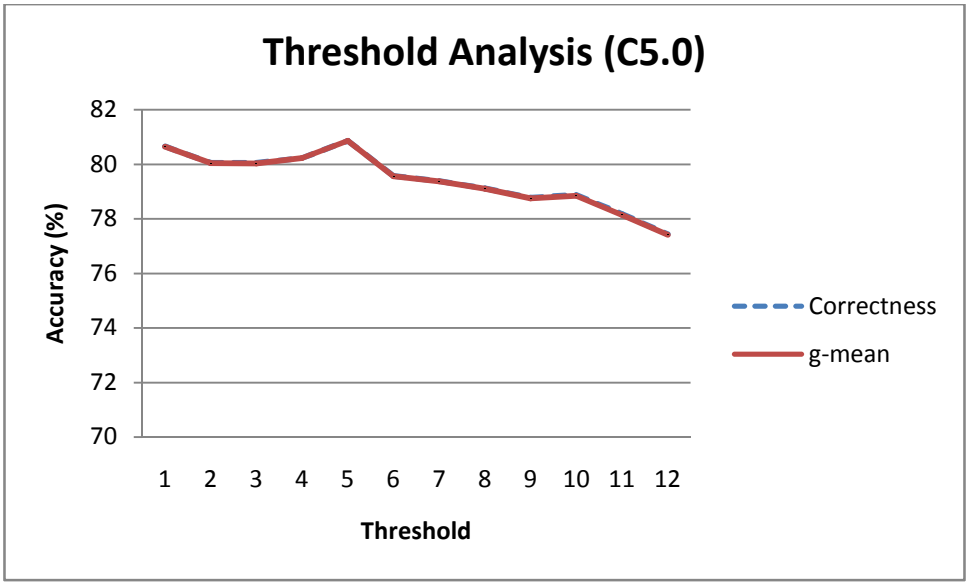


Figure 5.4: C5's threshold analysis results

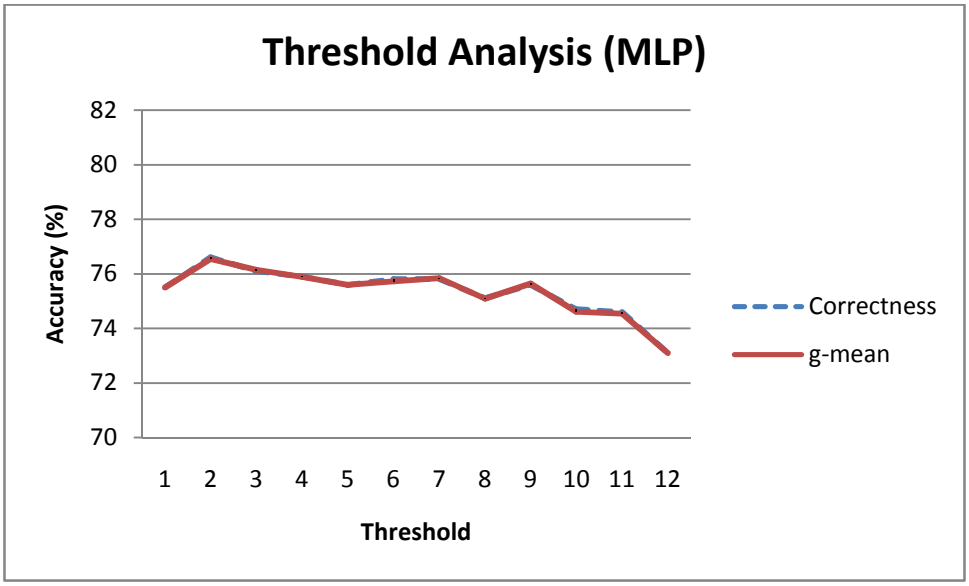


Figure 5.5: MLP's threshold analysis results

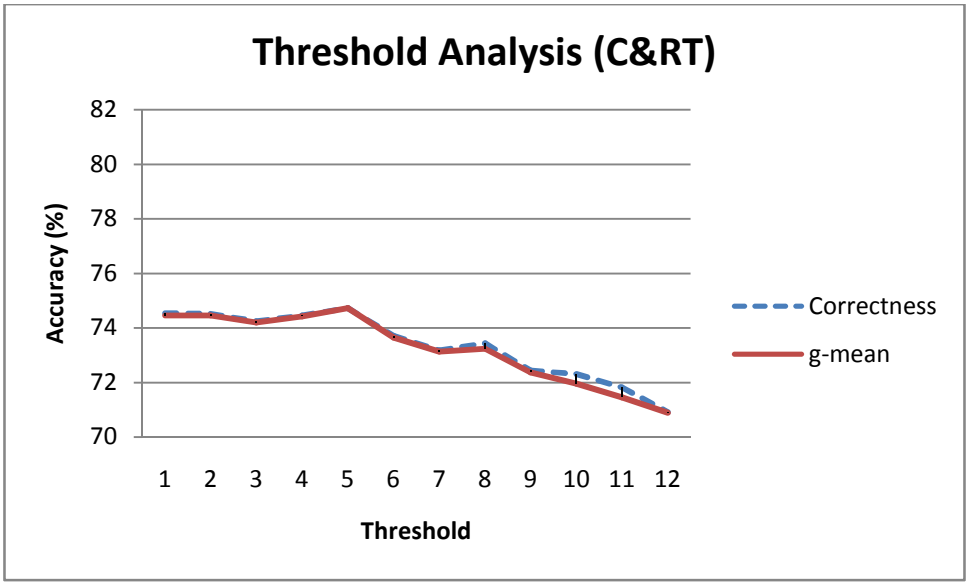


Figure 5.6: C&RT's threshold analysis results

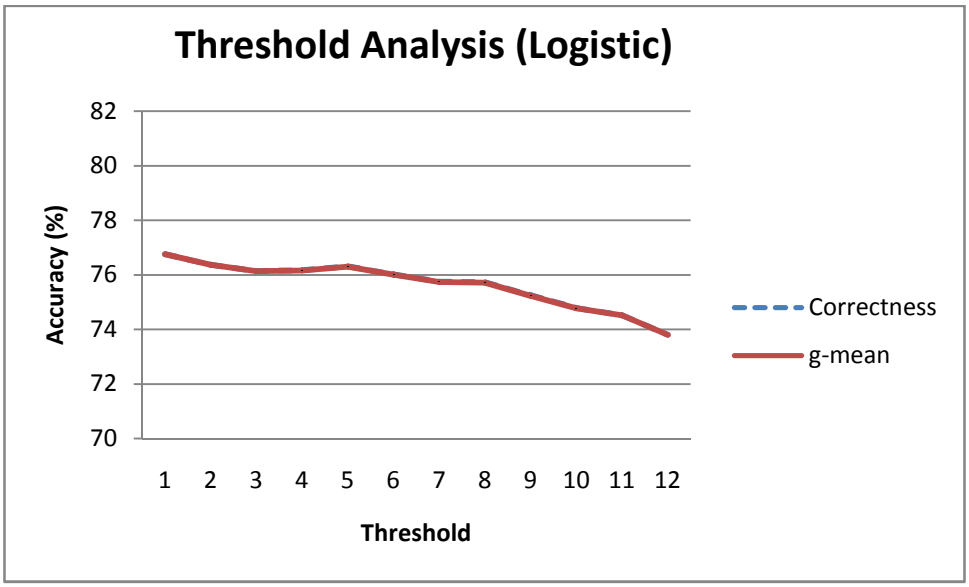


Figure 5.7: Logistic's threshold analysis results

The result that we expected was a kind of arch curve that by increasing the threshold the accuracy rate goes high and then down. For instance, figure 5.8 shows the result of a similar work done in another field of study. But, the generic trend that we can see in the threshold analysis results is a descending line which the accuracy rate goes down when the threshold increases.

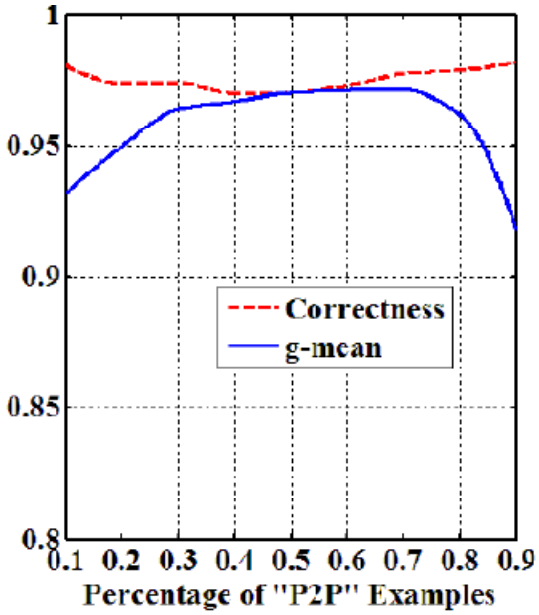


Figure 5.8: Performance of C4.5 on the P2P traffic data with different imbalances [49]

There are also some exceptions in the results, e.g. threshold-5 which is in common. Logically, this does not really make sense that threshold-5 gives better accuracy rate than threshold-1, and simultaneously on the other hand, threshold-4 gives lower accuracy rate than threshold-1. Because if the people who have more than or equal to 5 months coverage in the last 12 months have very similar behavior to each other about healthcare coverage, how can people who have more than or equal to 4 months coverage have less similar behavior than people who have more than or equal to 1 month coverage?!

In addition, the 4 months coverage samples build only 1% of the imbalanced dataset and when threshold-5 dataset was balanced, these samples formed only 3% of the population. How come the transfer of only 3% of the population from one target group to another is that much important that the model's accuracy rates grow significantly?!

The only possible answer for these questions is that maybe people with 4 months coverage have similar behaviors to people without any coverage, and on the other hand, people with 2

and or 3 months coverage have similar behaviors to people with full coverage. But, we keep the safe-side and based on the generic trends of all four models, we ignore these exceptions.

More, we believe that the effect of including 4 months coverage samples in uninsured people instead of insured group on the accuracy rate of model is true. But this effect is not very significant so that we select 5 months coverage as the healthcare coverage threshold. Maybe one of the reasons that amplify this effect is error in experiment. But if one wants to study the behavior of samples with different numbers of coverage months deeply, we recommend considering 5 months coverage in the last 12 months as a noteworthy borderline. It seems that people who have from 5 to 12 months coverage in one group and ones who have 0 to 4 months coverage in another group have similar behaviors and patterns in their groups.

To sum up, we didn't select any number as the healthcare coverage threshold directly from the results of this test. But, regarding the generic trends we found in the threshold analysis and the definition of insurance coverage summary indicated in MEPS full year consolidated data file documentation [45] we decided to work on threshold-1 dataset. We also call this dataset as Insurance Coverage dataset which has classified samples with 0 month coverage as uninsured and samples with 1 to 12 months coverage in the last 12 months as insured.

5.3.3 Whole Population Clustering

After selecting the insurance coverage dataset as the final dataset for this study, the unsupervised modeling was accomplished. The aim of this section was to find the main groups in our dataset which is a sample of the US population. We employed K-Means clustering algorithm on the imbalanced insurance coverage dataset. All 87,250 records and all 23 attributes were included in training the models. Different number of Ks from 2 to 5 were experimented and based on the outcome clusters that made more sense in each model

and the Silhouette average, the best model was selected. Note that K-Means clustering is a method which tries to partition the samples into K clusters in which each sample fit in the cluster with the nearest mean.

Among the outcome models, the model with 4 clusters was selected as the best outcome. The Silhouette average of this clustering model was 0.2 which is fair regarding IBM SPSS Modeler's labeling. One of the reasons that the Silhouette measure is not very high is due to the nature of healthcare data which is highly skewed [50]. This data is a sample of all people of the US and it's completely logical that the data will be scattered (distributed) a lot. Finding groups among various types of people can be very hard; therefore, we can only highlight some majorities in the society. Other samples will be spread in different parts of this grouping. Figure 5.9 shows the size of outcome clusters.

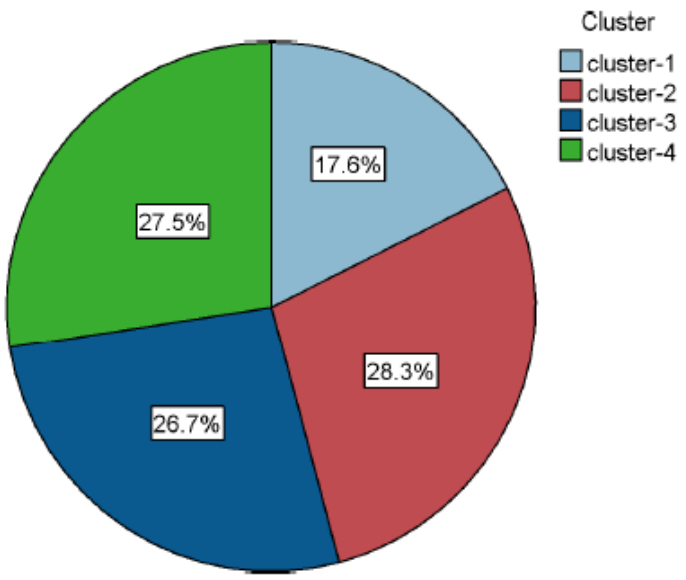


Figure 5.9: Cluster sizes of whole population clustering

The size of smallest cluster containing 15,342 samples is 17.6%, and on the other hand, the size of largest cluster containing 24,661 samples is 28.3%. Hence, the ratio of size of largest cluster to size of smallest cluster is 1.61 which shows the samples are well distributed

between clusters. Results of the whole population clustering has come in Appendix V showing the four outcome clusters which list and compare the percentage of major values of all attributes in each cluster and also the importance of each attribute.

Table 5.2 displays the clusters of the whole population. In this table, the title and characteristics of each cluster comes from the attributes which have one or two values in majority for that cluster (in bold format) and has that specific meaning in overall. For instance, samples that have been gathered in a cluster are all in the age of 0 to 18; therefore, this cluster shows the children which might also have other properties based on the values of other attributes. In a cluster, some attributes might cover two or more values which we can explore them using the cell distribution feature in IBM SPSS Modeler. For example, figure 5.10 shows the cell distribution of employment status in cluster-4.

Table 5.2: Clusters of the whole population

Cluster	Cluster-2	Cluster-4	Cluster-3	Cluster-1
Size	28.3%	27.5%	26.7%	17.6%
Title	Youth & adults with low income	Insured employees with high income	Insured children	Insured elders with access to care
Characteristics	**Youth & adults.** Good health. **Low income.** **No active duty.** **No limitation.**	**Employed.** Educated. Married. Good health. High income. **With coverage.** Other-race (white). **No active duty.** **No limitation.**	**Children.** Good health. **With coverage.** Have access to care. No priority condition. **No limitation.**	**Adults & elderly people.** Not employed. **With coverage.** **Have access to care.** Other-race (white).

From this part we found that the main groups among our dataset which is a sample of the US population are as following:

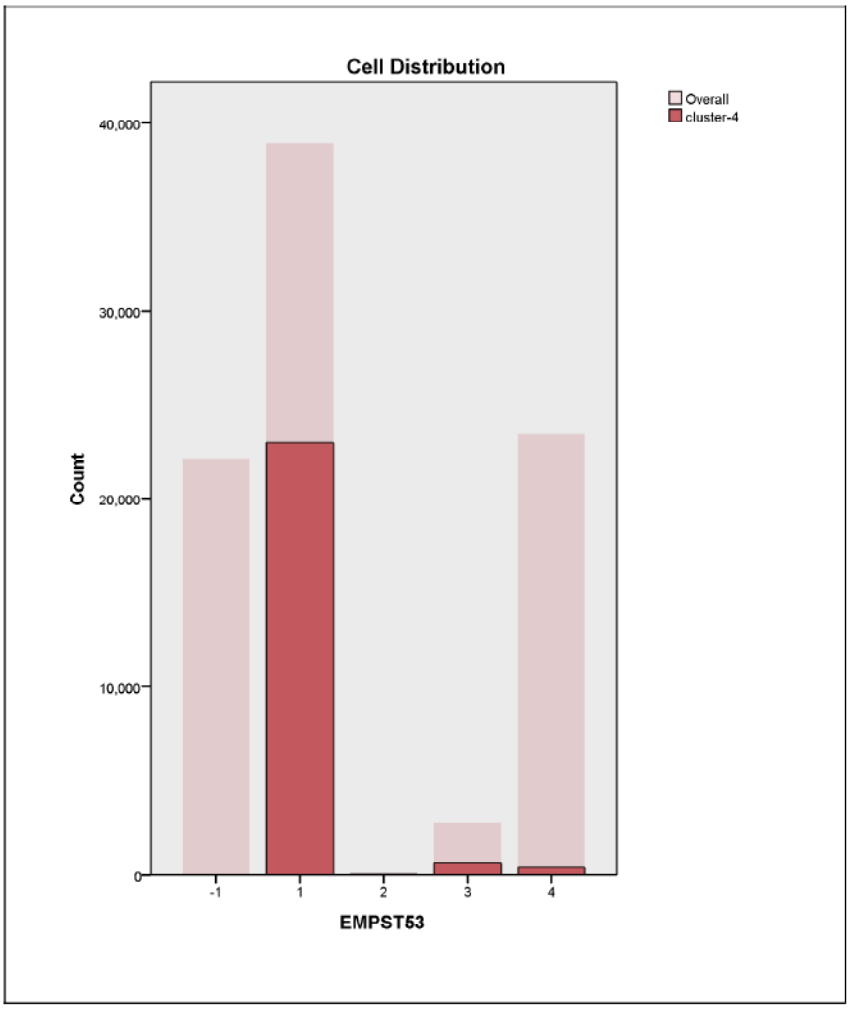


Figure 5.10: Example of cell distribution in whole population clustering

- Youth and adults which usually have good health but they don't earn much income. They also don't have any disabilities or limitations and not active duty as military status.
- The employees and workers who usually earn high incomes and have health insurance. Most of them are educated and in a marital status and experiencing a good health. In addition, most of them are White people or from other races except Hispanic, Black, and Asian. They also don't have any disabilities or limitations and not active duty as military status.

- Children who don't usually have much problem with their health. They also have insurance coverage, access to usual source of healthcare, and no disabilities or limitations.
- Elders who are usually not employed but have health insurance and also access to usual source of healthcare. In addition, most of them are White people or from other races except Hispanic, Black, and Asian.

5.3.4 Correlation Ranking

According to the result of threshold analysis section, we considered the insurance coverage dataset as the final dataset. Before attribute reduction procedure and exploring the outcome results, especially the predictor importance rankings of each model, we can see the ranking of attributes based on their correlation with the insurance coverage variable as descriptive information in the following table:

Table 5.3: Ranking of attributes based on their correlation with the target variable

No.	Variable	No.	Variable
1	HAVEUS42	12	REGION0#
2	ACTDTY53	13	WLKLIM53
3	HIDEG	14	IADLHP53
4	RACETHNX	15	ADLHLP53
5	POVCAT0#	16	RTHLTH53
6	TTLP0#X	17	SEX
7	WAGEP0#X	18	MNHLTH53
8	PCCOUNT	19	EDUCYR
9	HONRDC53	20	MARRY0#X
10	FTSTU0#X	21	AGE0#X
11	EMPST53	22	MSA0#

Although this table shows that HAVEUS42 (have access to care), ACTDTY53 (active duty status), and HIDEG (highest education degree) are the three top attributes that have high correlation with the target variable (insurance coverage), this ranking doesn't give any specific and useful result related to predicting healthcare coverage status or efficient factors in healthcare coverage. However, the only issue that we can find from this ranking is that having access to usual sources of care is more correlated to having healthcare coverage which this score is equal to 0.4. Maybe the people who have more access to care are more liked to go for a healthcare coverage. Note that on the other hand, having a very high correlation with the target variable is not good. The main reason that we didn't include the total expenditure variable is this study was this matter that the total medical expenditure variable would have had a high correlation with the insurance coverage variable; therefore, it is going to be a kind of derived attribute and in the data preprocessing should have been removed.

5.3.5 Attribute Reduction

Our main objective in this study is to find the efficient factors in healthcare coverage. As we can see in the research methodology of this project, we have planned to reduce the less important attributes and keep the efficient ones. The way that we are going to implement this process is building classification models to achieve the predictor importance ranking (attribute ranking) and then remove the least important attribute and again rebuild the models. Each of these iterations is called a "stage" which starts from zero. In other words, the stage shows how many attribute has been reduced for building this model. Only in the first stage, we will remove the attributes that are highly correlated to each other. In our test there is only one attribute with this specification and it is the total income variable

(TTLP0#). Note that the attribute reduction procedure will be continued until only one attribute remains. This method is not guaranteed to be optimal due to not considering all combinations of attributes in each stage, but it is an intuitive heuristic that gives a reasonable result and close to optimal.

C5 (decision tree) and MLP (neural network) algorithms were applied to build models which predict the two-label healthcare coverage. Because our dataset is balanced the correctness accuracy rate and the g-mean accuracy rate of the outcome models will be pretty the same. Therefore, we only calculate the correctness accuracy rate and report it as the model's accuracy rate. We stored accuracy rate and predictor importance ranking of each model. After doing the attribute reduction test, the results were gathered and displayed in a graphical chart to discover the optimum point of each method.

There is also an exception in stage18 of MLP modeling which we removed the last two least importance attributes. The reason was that by removing only the least importance attribute, MLP couldn't build any kind of model using the remaining attributes. Therefore, we removed the last two attributes, skip the outcome of stage 18, and build the model of stage 19. The accuracy rate of stage 18 of MLP was calculated using the mean of the accuracy rates of the previous and next stages (stages 17 and 19).

Forty four classification models were built (22 C5 and 22 MLP models). Figure 5.11 shows the accuracy rate of these models in a comparison chart. As we can see in this chart, in overall, C5 has better accuracy in predicting healthcare coverage than MLP. In stage 0 (when no attribute has been reduced), the accuracy rate of C5 is 80.65% while MLP's accuracy rate is just 76.2%. After stage 7, C5's accuracy starts dropping, and this drop continues until the end. On the other hand, MLP has a more stable accuracy trend whereas until stage 16 the accuracy rate doesn't change a lot and just has little oscillations, and after that it starts

dropping. In stage 16, both C5 and MLP have the same accuracy rate, and after that they both drop down equally.

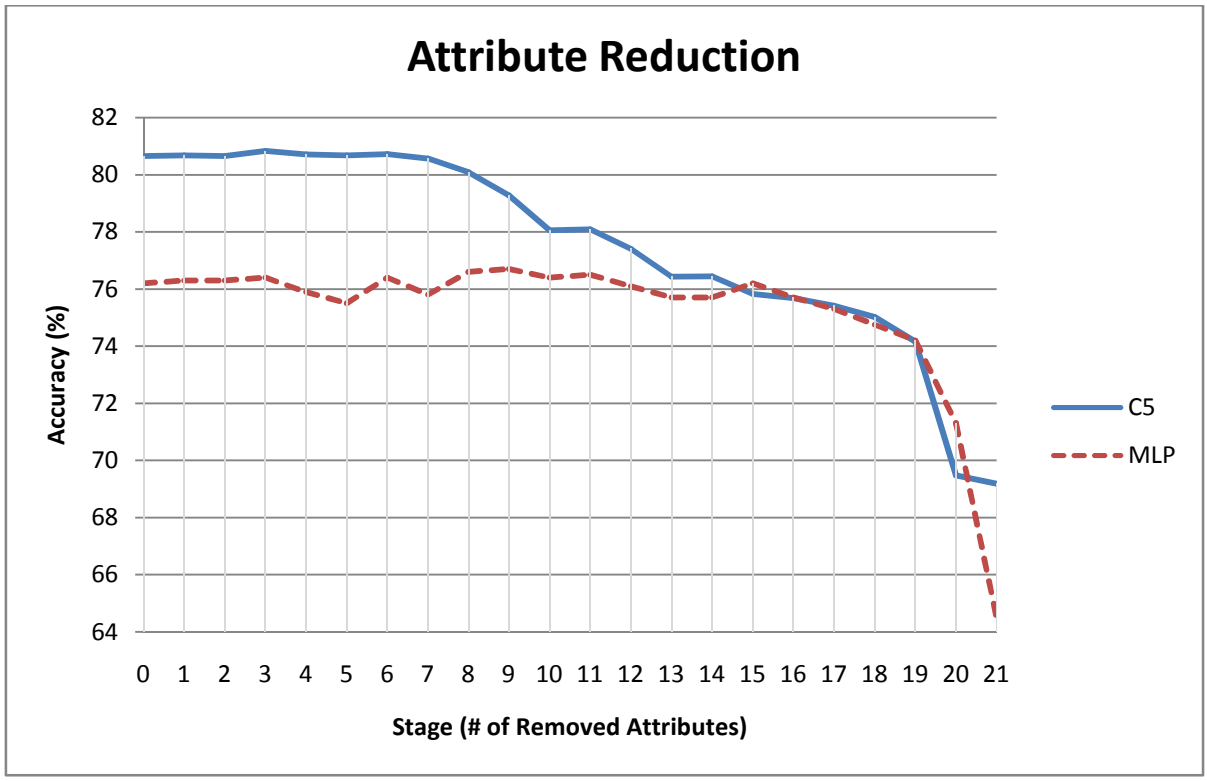


Figure 5.11: Attribute reduction comparison chart of C5 and MLP

5.3.5.1 Model Accuracy

Before discussing about the attribute rankings, we explain how the accuracy rates (correctness) were calculated briefly. Table 5.4 displays stage 0 of C5 model's confusion matrix. As we can see from this table, 80.42% of uninsured samples have been predicted as uninsured, and also 80.88% of insured samples have been predicted as insured. Therefore, the accuracy rate of this model is 80.65%. In addition, due to the balanced data, the model is not biased and predicts both insured and uninsured samples equally. You can also see the

distribution chart of this result in figure 5.12. Note that other models also have pretty similar distribution to this model in case of unbiased modeling results.

Table 5.4: Stage 0 of C5 model's confusion matrix

		Predicted	
		Uninsured	Insured
Observed	Uninsured	10,829	2,637
		80.42%	19.58%
	Insured	2,574	10,892
		19.12%	80.88%

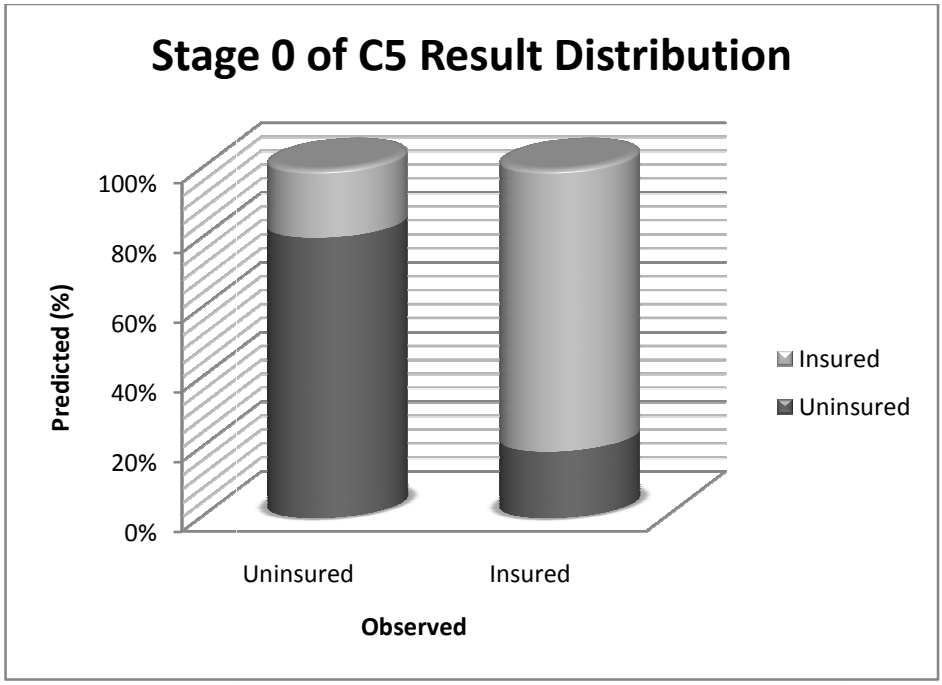


Figure 5.12: Stage 0 of C5 model's result distribution

5.3.5.2 Attribute Ranking

Appendices VI and VII show the complete attribute ranking of C5 and MLP modeling in each stage based on the predictor importance of each model which IBM SPSS Modeler calculates. Although all attributes have ranking, the attributes which had same predictor importance were binned into a same ranking group. Depending on how much accuracy rate is acceptable to a person, he can pick the attribute ranking related to that accuracy rate. For instance, assume that the minimum acceptable accuracy rate for person A is 80% and for person B is 75%. Person A has to forget about using MLP (neural networks) to predict healthcare coverage and his solution is to use C5 as the classifier. He can also refer to attribute ranking of C5 modeling and pick the standings of stage 8 which says the important predictors are respectively age, access to care, family's poverty level, and etc. But, person B has two options. He can go either for C5 or MLP to predict healthcare coverage. Also for important predictors, he can select the attribute ranking of either stage 18 of C5 or stage 17 of MLP.

Our goal is finding the model with more removed attributes (less input variables) and higher accuracy rate. The attribute ranking of that model is going to be the efficient factors in healthcare coverage. As we can see in figure 5.11, the maximum accuracy rate of C5 happens in stage 3 with the rate of 80.83% and MLP's is in stage 9 with the rate of 76.7%. Stages before these two are ineffective due to their low reduced attributes and lower accuracy rate. If our criterion is the accuracy rate of stage 0 which is the model that uses all the attributes to predict healthcare coverage, we can select stage 7 of C5 and stage 15 of MLP which have the same accuracy rate with their stage 0's one.

As a recommendation, and based on the results of the past studies, we believe that models with an accuracy rate of more than 80% are very good and also models with an accuracy rate

of more than 75% are acceptable due to the available accuracy rates in healthcare coverage prediction literature. Based on this recommendation, stage 8 of C5 is a very good model, and also stage 18 of C5 and stage 17 of MLP are acceptable. Table 5.5 and 5.6 show the attribute rankings of these stages in C5 and MLP.

Table 5.5: Attribute ranking of C5's notable stages

Title	Beginning Accuracy	Maximum Accuracy	Same Accuracy as Beginning	More than 75% Accuracy
Stage	**0**	**3**	**7**	**18**
Accuracy (%)	**80.65**	**80.83**	**80.56**	**75.02**
Rank 1	AGE0#X	AGE0#X	AGE0#X	HAVEUS42
Rank 2	HAVEUS42	HAVEUS42	HAVEUS42	AGE0#X
Rank 3	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#
Rank 4	RACETHNX	RACETHNX	EDUCYR	RACETHNX
Rank 5	EDUCYR	WAGEP0#X	REGION0#	-
Rank 6	WAGEP0#X	REGION0#	WAGEP0#X	-
Rank 7	REGION0#	EDUCYR	RACETHNX	-
Rank 8	TTLP0#X	PCCOUNT	PCCOUNT	-
Rank 9	MARRY0#X	MARRY0#X	MARRY0#X	-
Rank 10	EMPST53	FTSTU0#X	SEX	-
Rank 11	SEX	SEX	ACTDTY53	-
Rank 12	HIDEG	ACTDTY53	FTSTU0#X	-
Rank 13	PCCOUNT	EMPST53	ADLHLP53	-
Rank 14	RTHLTH53	ADLHLP53	MNHLTH53	-
Rank 15	MNHLTH53	WLKLIM53	HIDEG	-
Rank 16	IADLHP53	MNHLTH53	-	-
Rank 17	MSA0#	HONRDC53	-	-
Rank 18	ACTDTY53	HIDEG	-	-
Rank 19	ADLHLP53	RTHLTH53	-	-
Rank 20	HONRDC53	-	-	-
Rank 21	WLKLIM53	-	-	-
Rank 22	FTSTU0#X	-	-	-

As we can see in rankings of both C5 and MLP, the attribute rankings in different stages are not the same. The ranking of some attributes subsequently change from a stage to another. This shows that the combination of attributes also is important in having a good prediction for healthcare coverage. Also, in rankings of both C5 and MLP, especially the more than 75% accuracy column, the attributes that have the most importance are mostly in common. Having access to care, age, family's poverty level, and race/ethnicity are the important

factors in healthcare coverage. In other words, using just these 4 factors we can predict the healthcare coverage status of people with accuracy of 75%. In the next part, we discuss more about the values of these four factors in both insured and uninsured people using national statistics.

Table 5.6: Attribute ranking of MLP's notable stages

Title	Beginning Accuracy	Maximum Accuracy	Same Accuracy as Beginning	More than 75% Accuracy
Stage	**0**	**9**	**15**	**17**
Accuracy (%)	**76.2**	**76.7**	**76.2**	**75.3**
Rank 1	HAVEUS42	AGE0#X	AGE0#X	AGE0#X
Rank 2	AGE0#X	HAVEUS42	HAVEUS42	HAVEUS42
Rank 3	RACETHNX	EDUCYR	EDUCYR	POVCAT0#
Rank 4	IADLHP53	WAGEP0#X	PCCOUNT	EDUCYR
Rank 5	WAGEP0#X	PCCOUNT	POVCAT0#	RACETHNX
Rank 6	EDUCYR	RACETHNX	RACETHNX	-
Rank 7	HIDEG	HIDEG	WAGEP0#X	-
Rank 8	POVCAT0#	POVCAT0#	-	-
Rank 9	MARRY0#X	MARRY0#X	-	-
Rank 10	PCCOUNT	REGION0#	-	-
Rank 11	ACTDTY53	FTSTU0#X	-	-
Rank 12	EMPST53	ACTDTY53	-	-
Rank 13	TTLP0#X	HONRDC53	-	-
Rank 14	REGION0#	-	-	-
Rank 15	FTSTU0#X	-	-	-
Rank 16	MNHLTH53	-	-	-
Rank 17	HONRDC53	-	-	-
Rank 18	RTHLTH53	-	-	-
Rank 19	WLKLIM53	-	-	-
Rank 20	ADLHLP53	-	-	-
Rank 21	MSA0#	-	-	-
Rank 22	SEX	-	-	-

In addition, if we want to introduce another factor as an important factor in predicting healthcare coverage, we will introduce the Education Year attribute based on the attribute ranking of neural network which has more than 75% accuracy (in stage 17). We can also see the effects of this attribute in table 5.8 which shows the uninsured population characteristics in identifying groups of people regarding their healthcare coverage status.

Figure 5.13 shows a snapshot of one of the 22 neural network models that have been built in this part. In this figure we can see the input attributes in left, the hidden nodes in the middle, the target node in right, and also the connection between the input nodes and hidden nodes based on their relevance. Also, a snapshot of an outcome decision tree has also been shown in figure 5.14.

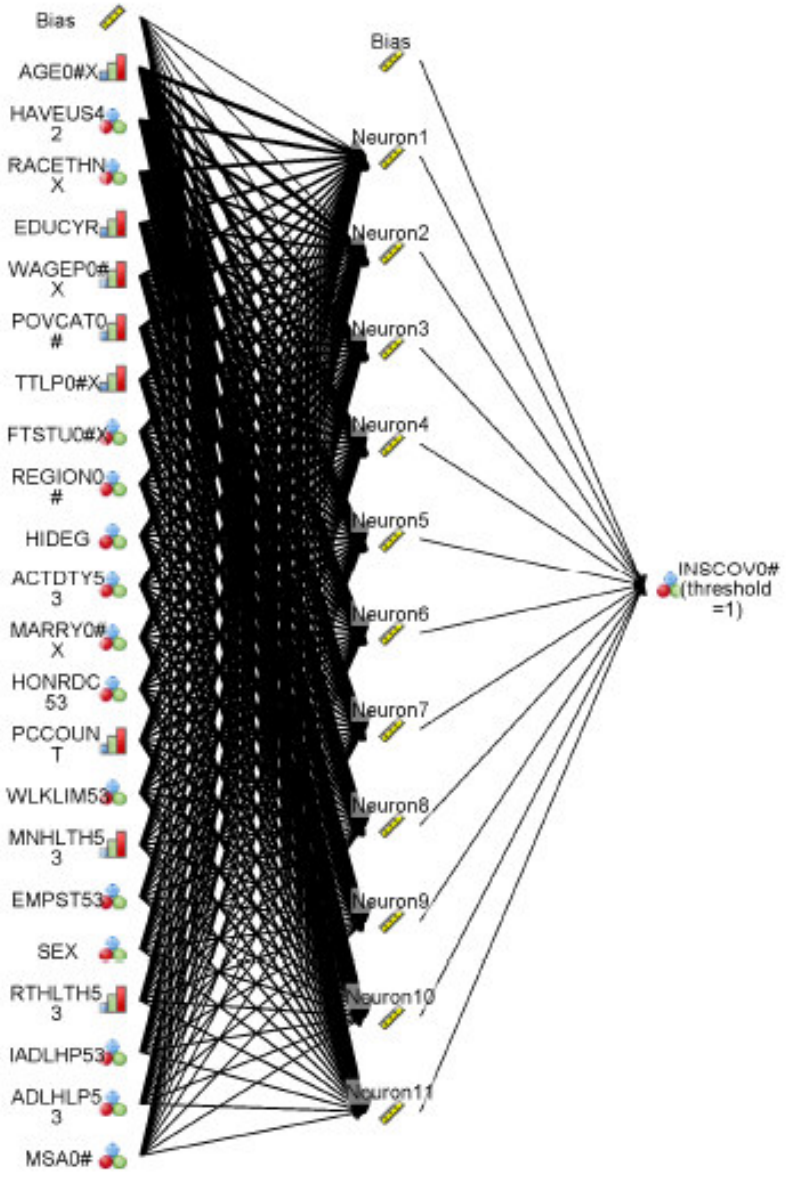


Figure 5.13: Snapshot of an outcome neural network

5.3.6 Efficient Factors

In this section, we explore the values of the four efficient attributes to see which value is the majority, which values have the most problem with insurance coverage, and what is the distribution of each value regarding the insurance coverage, using very simple statistical methods. To explore the data distribution, we included the person-level weight parameter which MEPS calculates for each sample. By including this parameter in the imbalanced insurance coverage dataset which includes data of 2006 to 2008, the data distribution and statistics generated are almost nationally representative. Regarding the documentation of full year consolidated data file in MEPS [45], the person-level weight estimates the population with more than 90% accuracy.

Before, we explore the C5 decision tree in stage 18 of attribute reduction procedure. This decision tree was built by the four efficient factors. Therefore, the classes appeared as leaves are the end groups in this tree which the model predicts the samples based on these paths and leaves.

5.3.6.1 Decision Tree of Efficient Factors

The C5 decision tree in stage 18 of attribute reduction is built by the four efficient factors. Figure 5.14 shows this decision tree. As we can see in this figure, all the records are available in the top basket. Based on the predictor importance, the samples have been divided into 2 or more baskets. Finally, all of them are included in one of the end nodes (leaves). In each leaf, the samples with this specific path (characteristics) will be predicted as insured or uninsured, depending on which type of coverage status are in majority between the samples of that leaf.

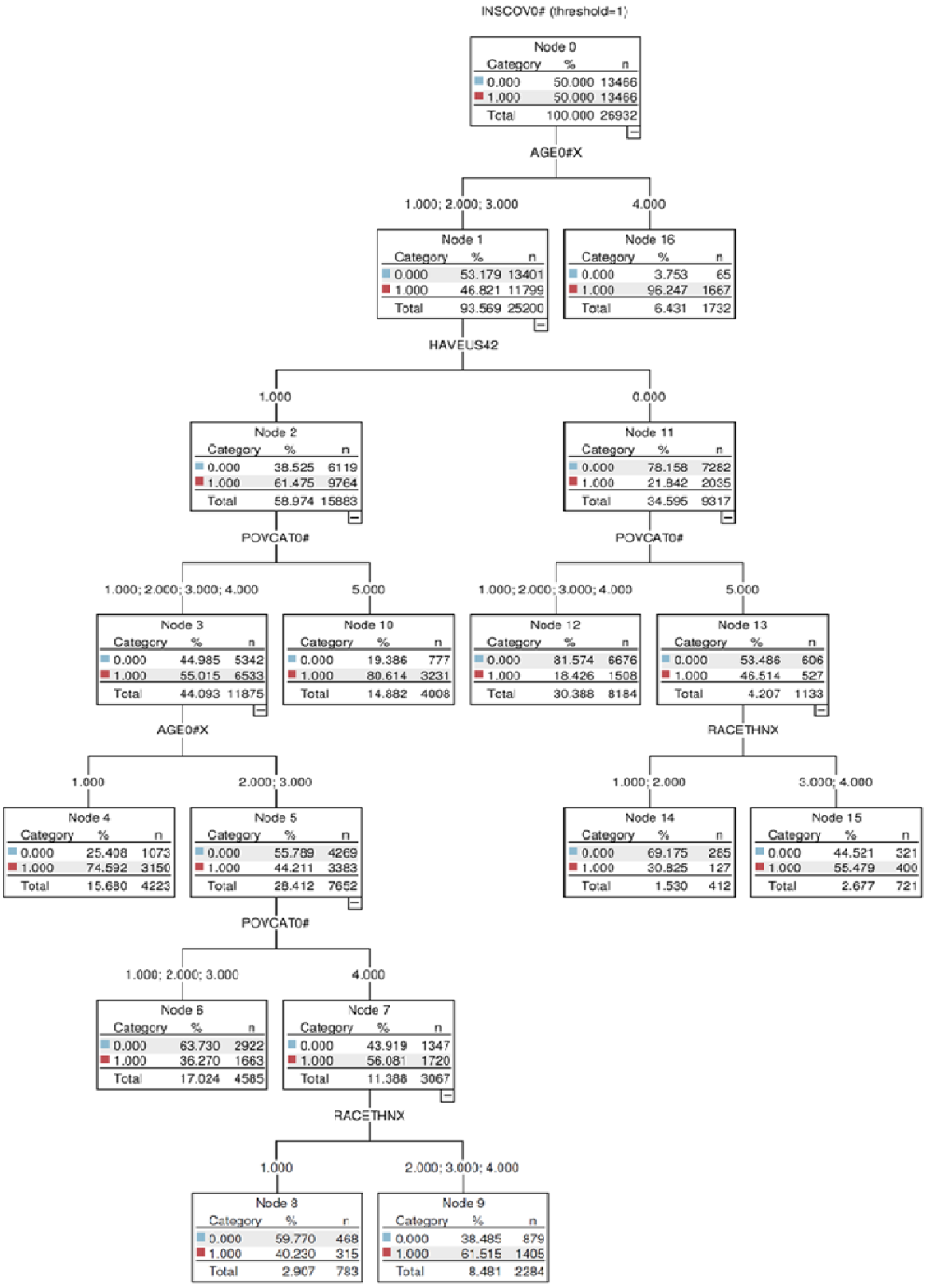


Figure 5.14: C5 decision tree in stage 18

In this section, we interpret the meaning of each path that has ended up as a leaf in this tree. This path and leaf introduce a group of people which somehow have a specific healthcare coverage status, based on this decision tree. Note that the accuracy rate of this decision tree is more than 75%. Table 5.7 displays the characteristics of group, group size, and uninsured population inside the group for each class in the balanced insurance coverage dataset. Each row is a path in this decision tree which has ended up with a leaf as a class. The table is sorted based on the group size.

Table 5.7: Classes of C5 decision tree in stage 18 of attribute reduction

No.	HAVEUS42	AGE0#X	POVCAT0#	RACETHNX	Uninsured Population	Group Size
1	0	1,2,3	1,2,3,4	-	**81.57%**	30.39%
2	1	2,3	1,2,3	-	**63.73%**	17.02%
3	1	1	1,2,3,4	-	25.40%	15.68%
4	1	1,2,3	5	-	19.38%	14.88%
5	1	2,3	4	2,3,4	38.48%	8.48%
6	-	4	-	-	3.75%	6.43%
7	1	2,3	4	1	**59.77%**	2.91%
8	0	1,2,3	5	3,4	44.52%	2.68%
9	0	1,2,3	5	1,2	**69.17%**	1.53%
Total						**100%**

When this model receives a new sample, it will predict the healthcare coverage status of that sample using this table. The input sample will be fitted in one of these groups, depending on the values of these attributes. If the uninsured population of that group is more than 50% (in bold format) it means the model will predict this sample as uninsured, and vice versa.

Therefore, based on this decision tree which is only 75% accurate, these groups will be predicted as uninsured:

- Group 1: Children, youth, and adults who don't have access to care and aren't in high income families.
- Group 2: Youth and adults that have access to care, but are in poor, near poor, or low income families.
- Group 7: Hispanic youth and adults who have access to care and are in middle income families.
- Group 9: Hispanic and Black nonelderly people who don't have access to care and are in high income families.

Note that these classes are only for this decision tree to predict healthcare coverage status with 75% accuracy rate. In the next sections we provided nationally representative statistics using the weight variable of MEPS.

5.3.6.2 Have Access to Care

This attribute has two values: Yes and No, which very simply reports that whether each sample has access to usual sources of care or not. As we can see in the table of figure 5.15 which shows the distribution of the US population based on access to care and insurance coverage, 80% have access to care, and on the other hand, 20% don't. These percentages are built using person-level weight variable of MEPS for national estimates in the imbalanced insurance coverage dataset which covers years 2006 to 2008.

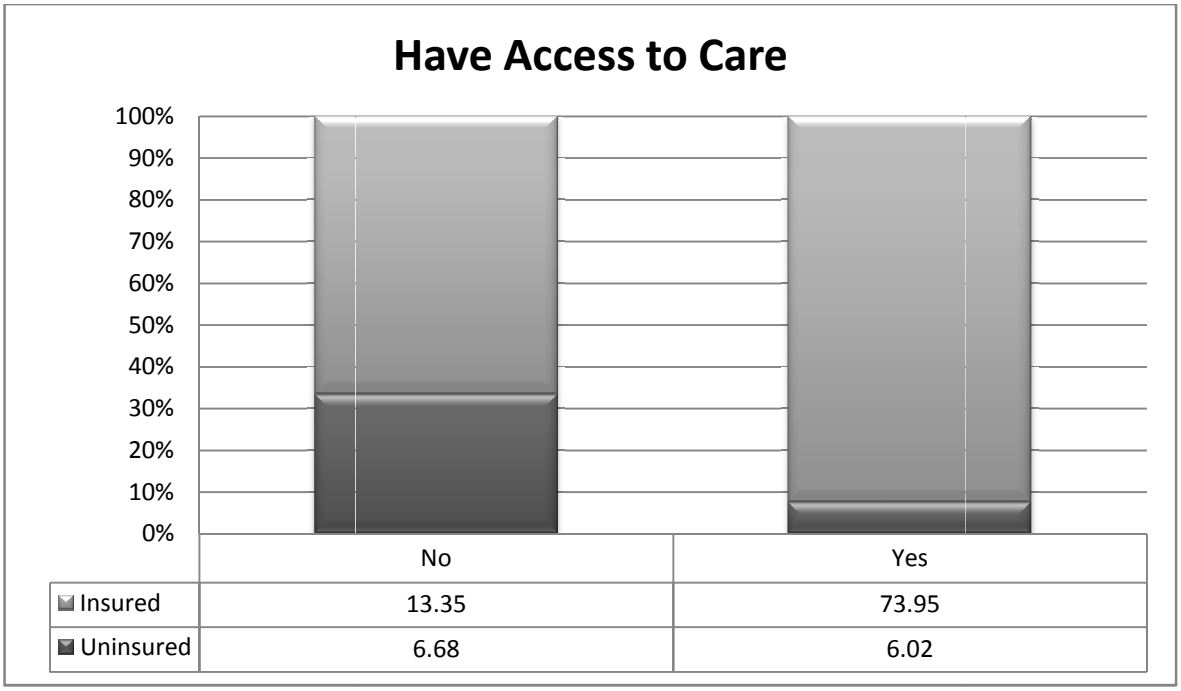


Figure 5.15: Access to care and insurance coverage distribution in the US

We understand from this chart that most (92%) of the people who have access to care and most (67%) of the people who don't have, are insured (respectively 74% and 13% of total). But, a significant number (33%) of the people who don't have access to care are not insured (7% of total). In addition, although a small portion (8%) of the people who have access to care don't have insurance coverage (6% of total), this amount in the total population is equal to the number of people who don't have access to care and don't have coverage, which is weird. In overall, we can see the impact of healthcare coverage on access to care, and vice versa in this chart. Having healthcare coverage will increase the possibility of having access to care.

5.3.6.3 Age

The age attribute was binned into four groups: 0-18, 19-40, 41-65, and 66-85. As the table of figure 5.16 which shows the distribution of the US population based on age and insurance coverage display, the population of age values are respectively 24%, 32%, 33%, and 11%.

These percentages are built using person-level weight variable of MEPS for national estimates in the imbalanced insurance coverage dataset which covers years 2006 to 2008.

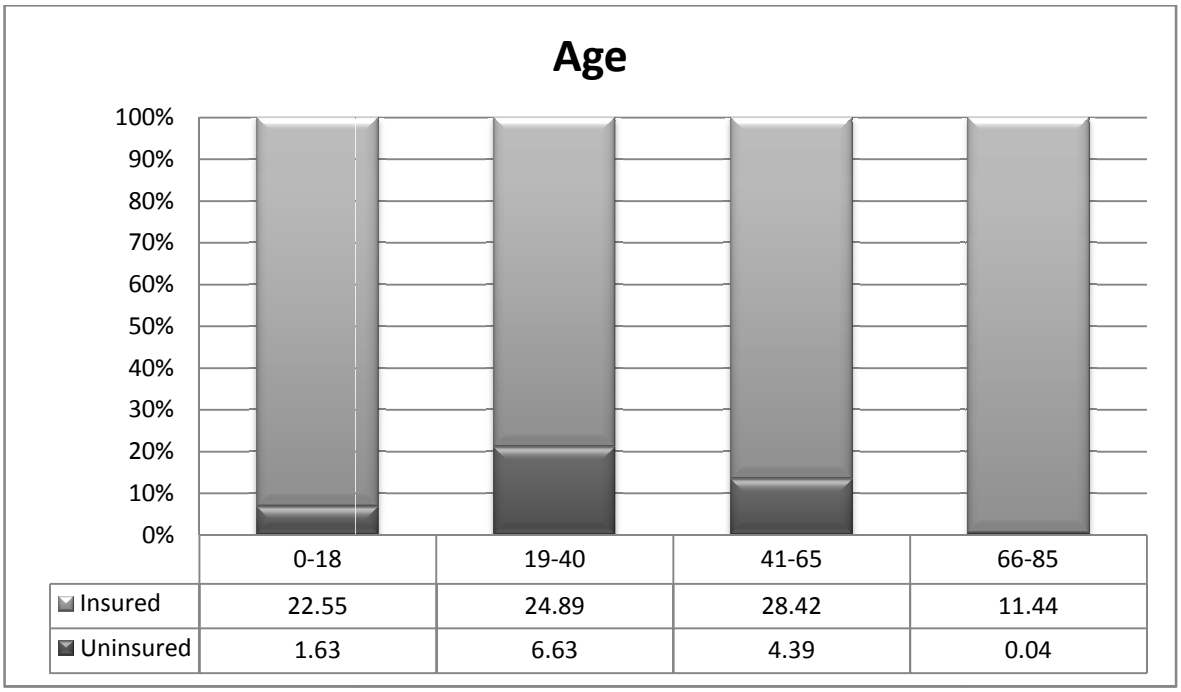
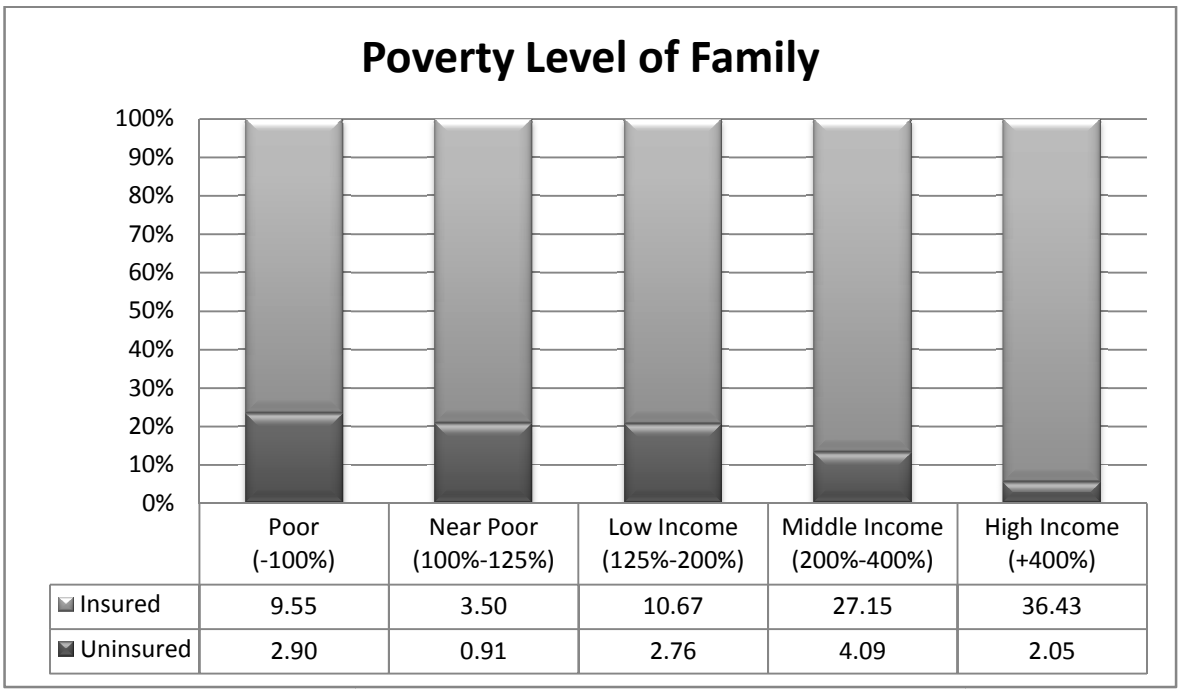


Figure 5.16: Age and insurance coverage distribution in the US

This chart shows a significant number (21%) of the youth aged 19 to 40 are uninsured (6% of total). Some reasons for this issue may be that the youth are healthier, more risk taking, and gain less amount of income. Adults aged 41 to 65 are in the second step of uninsured (4% of total). Most (93%) of children and youngsters aged 0 to 18 and most (99%) of the old people aged 65 to 85 have insurance coverage (respectively 22% and 11% of total) which seems logical due to their necessary needs to medical care and the available public health plans. Elderly people are mostly covered by Medicare, thus, health coverage gaps usually exist in nonelderly group of people [32]. Although public health plans are available for children in poor family, 7% of them are uninsured.

5.3.6.4 Poverty Level of Family

The poverty attribute shows the family income of the sample divided by the poverty line in five categories: Poor (-100%), Near Poor (100%-125%), Low Income (125%-200%), Middle Income (200%-400%), and High Income (+400%). As we can see in the table of figure 5.17 which shows the distribution of the US population based on poverty level of family and insurance coverage, this attribute's instance rates are respectively 12.5%, 4.5%, 13.5%, 31%, and 38.5%.



	Poor (-100%)	Near Poor (100%-125%)	Low Income (125%-200%)	Middle Income (200%-400%)	High Income (+400%)
Insured	9.55	3.50	10.67	27.15	36.43
Uninsured	2.90	0.91	2.76	4.09	2.05

Figure 5.17: Poverty level of family and insurance coverage distribution in the US

In the related studies, one of the main items that effect healthcare coverage is the socioeconomic situation of people. People are usually divided to two socioeconomic levels: lower socioeconomic groups and higher socioeconomic groups. Smoking, poor diet, and lack of physical activity are signs of poor lifestyle which usually exist in lower socioeconomic groups [51] [20]. On the other hand, frequent medical checkups and other types of preventive cares mostly happen in higher socioeconomic groups [20]. One of the main reasons of

socioeconomic inconsistencies in health is the financial difficulty to achieve proper health services [52]. Health insurance plans in the US are usually expensive and lower socioeconomic groups cannot afford buying costly insurances and this might cause having less access to care, and suffering from more illnesses [53].

Figure 5.17 shows chart that most (87% and 95%, respectively) of the people in middle and high income families have insurance coverage (27% and 36.5% of total, respectively). On the other hand, a significant number of the people in poor, near poor, and low income families (20% of them) are not insured (respectively 9.5%, 3.5%, and 10.5% of total).Although health insurances in the US are costly and hardly affordable for families with low income, 6% of people in middle and high income families still don't have insurance coverage.

5.3.6.5 Race/Ethnicity

Finally, the forth efficient factor in healthcare coverage is race/ethnicity which has four values: Hispanic, Black, Asian, and Other (including White). As we can see in the table of figure 5.18 which shows the distribution of the US population based on race/ethnicity and insurance coverage, the population of this attribute's values are respectively 15%, 12%, 4%, and 69%.

As this chart shows us, most (90%) of the race/ethnicity majority which is the "Other" (includes White and etc.) are insured (62% of total). On the other hand, a significant number (27%) of the Hispanic people don't have insurance coverage (4% of total). Also, about 11% of the Black and Asian people are uninsured (respectively 2% and 0.5%). Any heterogeneous in different races/ethnicities having insurance coverage is not logical, and for sure there are some specific reasons for this matter.

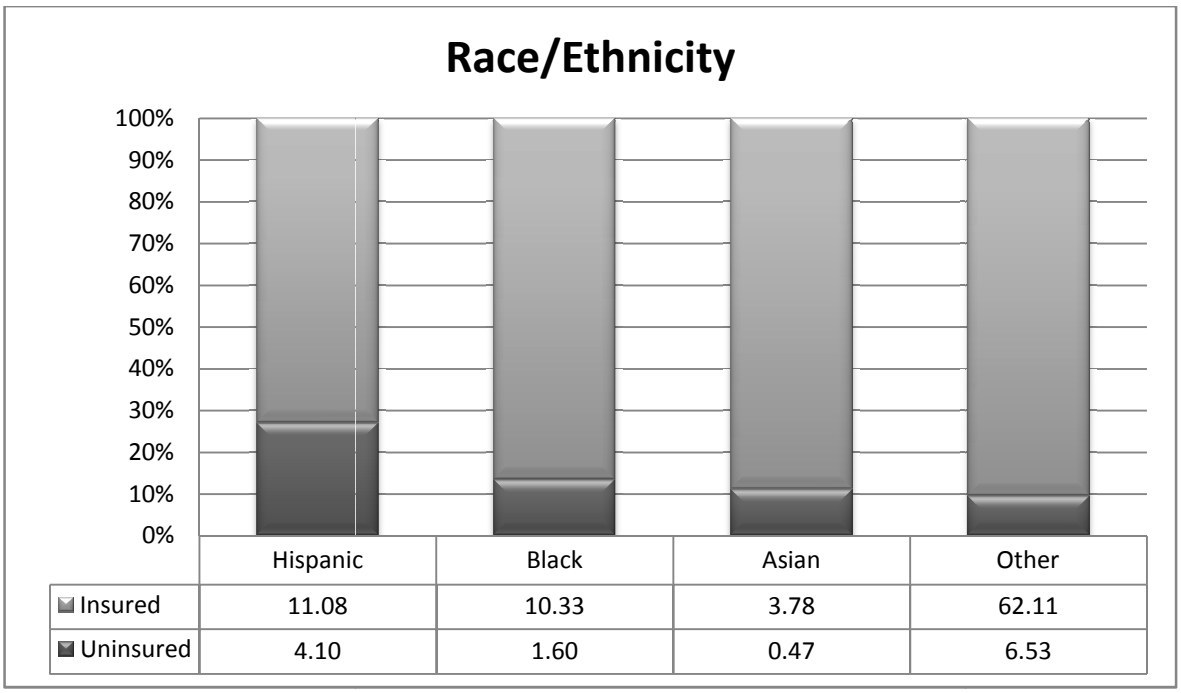


Figure 5.18: Race/ethnicity and insurance coverage distribution in the US

Buying health insurance for people with low incomes is usually tough. On the other hand, many of low income people are men of color, including Hispanic, Asian American, African American, and etc. [32]. Special conditions of people of color have been expressed in continue. Figure 5.19 shows a snapshot of lower income uninsured people by race and employment source in the US.

In addition to health insurance costs, another problem that some people in color face is the quality of health services that they receive. Because English is not the primary language of some segments such as most of Hispanics, there are problems in the communication between the provider and the patient; this can result in giving poor health quality services to these people [32]. According to Actuarial Research Corporation (ARC) analysis of March 2004 CPS data, about 44% of the uninsured people eligible for public programs who could have enrolled in Medicaid and/or SCHIP were Hispanic. There are several reasons for this issue,

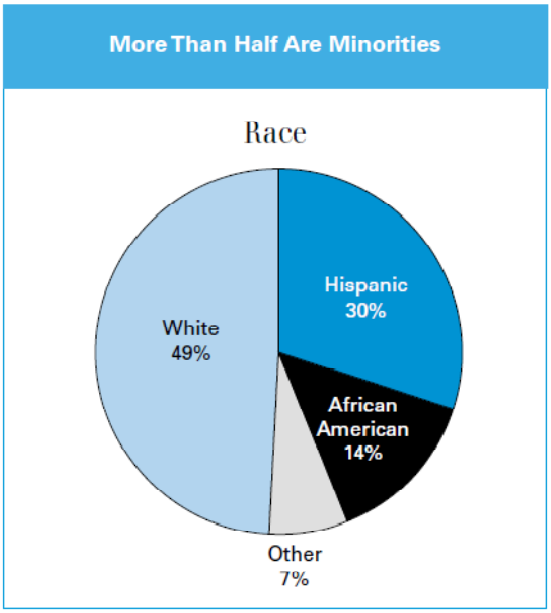
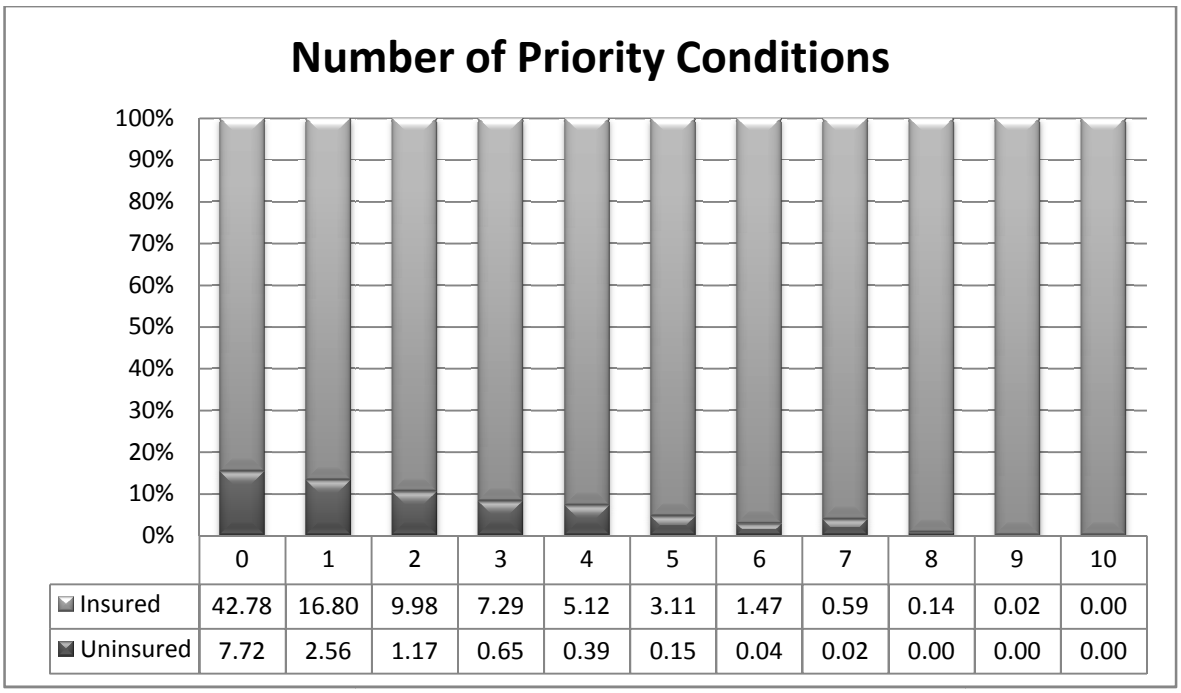


Figure 5.19: Snapshot of lower-income uninsured: United States, 2004 [14]

such as being unaware of this matter, unable to go through complicated application procedures, and etc. Also, one of the possible reasons of this fact is the language barriers for Hispanic people [14].

5.3.6.6 Number of Priority Conditions

The PCCOUNT attribute was in the top six important attributes in both C5 and MLP modeling. We didn't count it as an efficient factor in healthcare coverage, but because we constructed this attribute from the priority condition variables and it hasn't been experienced in the related works, we decided to report its value distribution based on insurance coverage. We checked the nine priority conditions for each sample and counted them for each person. Thus, the possible values are from 0 to 10. As we can see in the table of figure 5.20 which show the distribution of the US population based on number of priority conditions and insurance coverage, the population of this attribute's values are respectively 49.5%, 19.4%, 11.1%, 8%, 5.5%, 3.3%, 1.5%, 0.6%, 0.1%, 0.02%, and 0.005%.



	0	1	2	3	4	5	6	7	8	9	10
Insured	42.78	16.80	9.98	7.29	5.12	3.11	1.47	0.59	0.14	0.02	0.00
Uninsured	7.72	2.56	1.17	0.65	0.39	0.15	0.04	0.02	0.00	0.00	0.00

Figure 5.20: Number of priority conditions and insurance coverage distribution in the US

Fortunately, about half of the population doesn't have any priority condition which 15% of them are uninsured (8% of total). Although Medicaid health plan also covers poor people disabled by chronic conditions, about 5% of the US population, which is a significant number, have priority condition and aren't insured. This will cause high cost pressures on that person and his family. There is also a trend in this chart that is very logical and completely makes sense. This trend says that, as the number of priority conditions increases in the samples, more people will get insurance coverage. The trend continues till the samples which have all the priority conditions and even 9 priority conditions out of 10 are completely insured (0.005% and 0.02% of total). People with more priority conditions have more health needs, though, are more likely to have coverage. This chart is good evidence which shows clearly that most of the people use their insurance coverage as sick-care, not healthcare.

5.3.7 Uninsured Population Clustering

In the last section of this study, we discovered the main groups of people in the uninsured population of the US to figure out more information about this specific population and combine the information and statistics we reviewed in the previous section. One of the ways to get additional information from a specific dataset is finding its hidden groups using a clustering method. K-Means clustering was again employed to cluster the uninsured population with the aim of presenting main uninsured segments of people. Balanced and imbalanced insurance coverage dataset had the same number of uninsured samples (13,466 samples) because the uninsured population was the minority compared to the insured population and therefore none of their samples were under-sampled during the data balancing. All attributes were included in the uninsured population dataset for the purpose of clustering. Again, different number of Ks from 2 to 5 were experimented that the best model was presented. This model introduced the main uninsured groups which resulted in clusters with labels that made more sense and present specific values in more attributes.

K-Means clustering was employed and the model with 5 clusters was selected as the best outcome. The Silhouette average of this clustering model was 0.15 which is near fair. Again we believe that the reason of not getting a good Silhouette score for this clustering model is because of the highly skewed data and also the data distribution among an uneven-behavior population with no much cohesion. Figure 5.21 shows the size of outcome clusters.

The size of smallest cluster containing 1,717 samples is 12.8%, and on the other hand, the size of largest cluster containing 3,398 samples is 25.2%. Hence, the ratio of size of largest cluster to size of smallest cluster is 1.98 which shows the samples are well distributed between clusters. Results of the uninsured population clustering has come in Appendix VIII

showing the five outcome clusters which list and compare the percentage of major values of all attributes in each cluster and also the importance of each attribute.

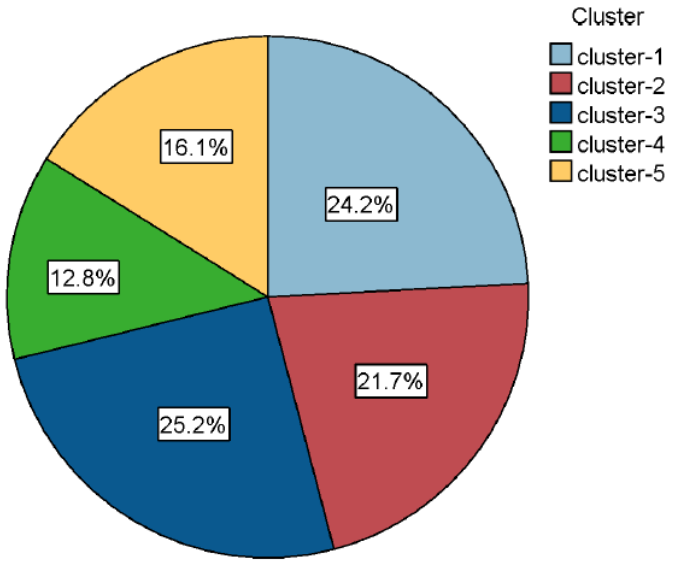


Figure 5.21: Cluster sizes of uninsured population clustering

Table 5.8: Clusters of the uninsured population

Cluster	Cluster-3	Cluster-1	Cluster-2	Cluster-5	Cluster-4
Size	25.2%	24.2%	21.7%	16.1%	12.8%
Title	Healthy employed bachelors with low income & no access to care	Uneducated Hispanic wives with low income & no access to care	Educated ladies with priority conditions, low income, & access to care in other races	Educated husbands with middle income & no access to care	Healthy Hispanic children with access to care
Characteristics	Men. **Youth.** **Never married.** Educated. Employed. **Low income.** No priority condition. Good health. No access to care. **No limitation.** **No active duty.**	Women. **Youth & adults.** Married. Not educated. No degree. Employed. **Low income.** Fair health. No access to care. **Hispanic race.** **No limitation.** **No active duty.**	Women. Adults. Educated. Diploma degree. **Low income.** Have priority condition. Fair health. Have access to care. Other-race (white).	Men. **Youth & adults.** Married. Educated. Middle income. No access to care. **No limitation.** **No active duty.**	**Children.** **No priority condition.** Good health. Have access to care. Hispanic race. **No limitation.**

Table 5.8 displays the clusters of the uninsured population clustering. In this table, the title and characteristics of each cluster comes from the attributes which have one or two values in majority (in bold format) for that cluster and has that specific meaning in overall. In a cluster, some attributes might cover two or more values which we can explore them using the cell distribution feature in IBM SPSS Modeler. For example, figure 5.22 shows the cell distribution of race/ethnicity in cluster-1.

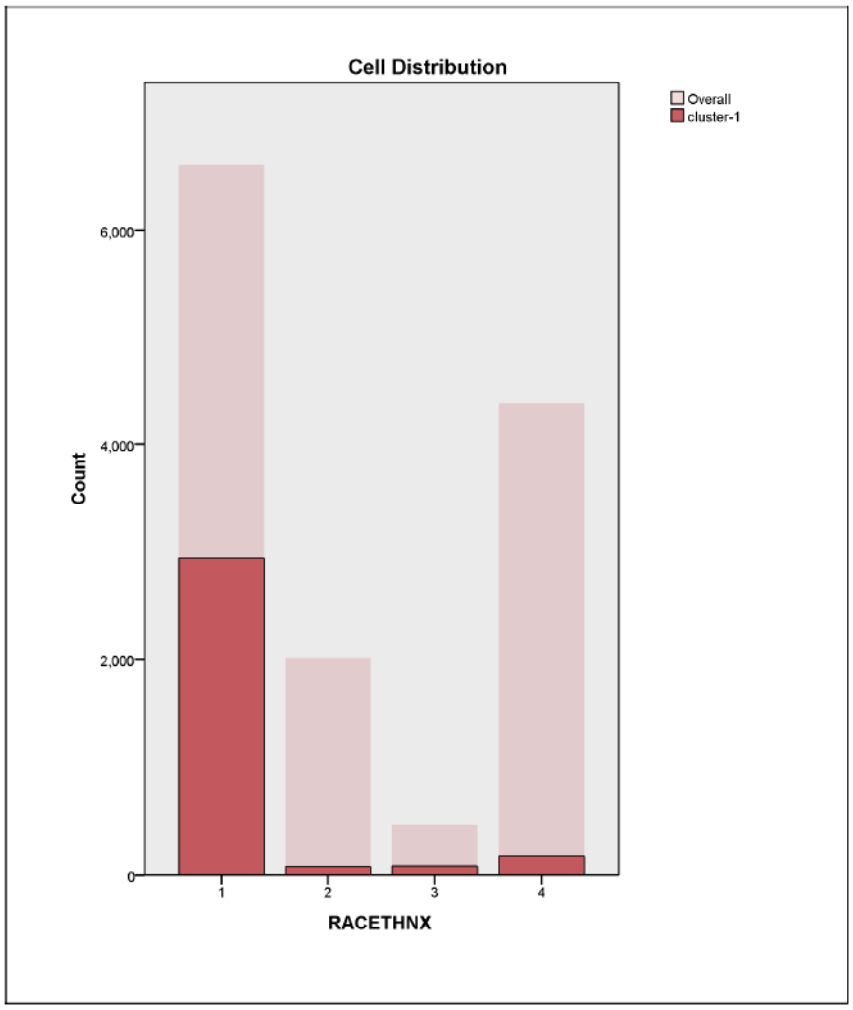


Figure 5.22: Example of cell distribution in uninsured population clustering

From this part we found that the most common characteristic among main groups of uninsured population is having low income. In other words, health insurances are not

affordable for low income population who are not eligible for public health plans. The main groups of uninsured people in the US are as following:

- Young men who have never married. They have a job, but their income is low. They are mostly in a good health status, but they don't have access to usual source of healthcare. They also don't have any disabilities or limitations and not in active duty of military.
- Hispanic young and adult wives who are not very educated. They work but they don't earn much. They don't have a good health status and they don't have access to usual source of healthcare. They also don't have any disabilities or limitations and not in active duty of military.
- Adult women who are educated, earn low income, and mostly from the White or other races except Hispanic, Black, and Asian. They don't have a good health status and usually suffer from a priority condition, but they have access to usual source of healthcare.
- Young and adult husbands who are educated, earn middle income, and don't have access to usual source of healthcare. They also don't have any disabilities or limitations and not in active duty of military.
- Hispanic children who neither have any priority condition, nor any disabilities or limitations. They mostly experience a good health with having access to usual source of healthcare.

Table 5.9 shows the major characteristics of main uninsured groups of US population regarding the efficient factors of healthcare coverage discovered in the supervised learning models of this study. This table compares the results of supervised learning and unsupervised

learning models of this study comprehensively and indicates that although different algorithms were used in each of them, but the results are validating and supporting each other. The characteristics of uninsured people in unsupervised learning models are mostly identified by the efficient factors achieved from the supervised learning results.

Table 5.9: Main uninsured groups of US population regarding efficient factors of healthcare coverage

Cluster	Have Access to Care	Age	Poverty Level of Family	Race/ Ethnicity	Other
1	No	Youth	Low income	-	Healthy Employed Bachelors
2	No	Youth/Adults	Low income	Hispanic	Uneducated Wives
3	Yes	Adults	Low income	Other races	Educated Ladies Chronic illnesses
4	No	Youth/Adults	Middle income	-	Educated Husbands
5	Yes	Children	-	Hispanic	Healthy

5.4 Summary

In this chapter, we discussed all the materials in healthcare coverage modeling. Supporting contents and information in the previous chapters helped us to design a comprehensive modeling for this study. Proposing the models, presenting their results, and analyzing the outcomes were the main parts of this chapter. In the following, we listed the highlights of this chapter's findings:

- Insurance coverage variable should have binary values (insured or uninsured) in data modeling.
- The generic trend of insurance coverage threshold from having 1 month coverage in the 12 months to having 12 months coverage is descending.
- The medical expenditure dataset is highly skewed.
- C5 predicts healthcare coverage status more accurate than MLP.
- The combination of attributes has impact in better predicting healthcare coverage.
- Access to care, age, poverty level of family, and race/ethnicity factors have significant effects on healthcare coverage.
- People without having access to care are more likely to be uninsured.
- Youth and adults form the major segment of uninsured population in the US.
- Poverty level of family affects healthcare coverage status directly.
- Minorities of race/ethnicity, especially Hispanics, mostly have problems in their healthcare coverage.
- People with more priority conditions have more health needs, though, are more likely to have coverage.
- The most common characteristic among main groups of uninsured population is having low income. Health insurances are not affordable for low income population who are not eligible for public health plans.

Chapter Six

6 CONCLUSIONS

This chapter concludes the thesis with a summary of its finding, expressing the study's limitations, and suggesting future works.

6.1 Summary of the Thesis

In this thesis we start with an introduction on healthcare coverage and the effects of its disparity. Lack of healthcare coverage will cause main problems in people's lives, such as poor health and even early death. The research questions of this study lead us to find important and efficient factors in healthcare coverage and also discovering groups of people with health coverage problems and inconsistencies. The method to fulfill the objectives of the study is by developing data mining models including supervised and unsupervised modeling. The aim of employing supervised learning is to predict healthcare coverage status of samples and examine the impacts of attributes on healthcare coverage, and by employing unsupervised learning we want to describe specific populations of our dataset and discover the main groups that exist in that population. Note that we considered the whole year coverage status in building our models. Based on our initial results, healthcare coverage prediction models, efficient factors of healthcare coverage, and main uninsured groups of people, were achieved with this assumption that not having coverage in the whole year is considered as uninsured and others are considered as insured.

In the background section, a large number of articles, reports, and other supporting documents were reviewed. Healthcare coverage disparity in the United States was explained

using recent statistics which the coverage gap is getting wider subsequently. The percentage of uninsured people has reached 16.3% in 2010 which means around 50 million people. Different reasons which lead to the existing disparity, such as costly health insurances, were presented and discussed. We also investigated the universal coverage in Canada comparing it to the United States. Sources of payment for healthcare coverage were presented including different types of private coverage and public coverage. Other past studies which had built healthcare coverage models employing data mining and statistical techniques were also reviewed in this past.

The dataset used in this study was retrieved from Medical Expenditure Panel Survey (MEPS) databases which are provided by the Agency for Healthcare Research and Quality of the US Department of Health and Human Services. MEPS release nationally representative samples from all parts of the country containing enormous amounts of data with different varieties. We aggregated the data files of years 2006, 2007, and 2008 and after several operations of data preprocessing, such as attribute selection, data cleaning, and data balancing, the final datasets were prepared. Different target variables using different conditions were constructed for several tests. The final supervised learning was applied on the dataset which originally contained 98,175 records and after data cleaning and balancing contained 26,932 records including 23 variables.

IBM SPSS Modeler was employed as a data mining, modeling, and reporting tool experiencing a visual user friendly interface. The proposed models were built and tested in this environment. C5 decision tree and MLP neural network algorithms were employed for supervised modeling; on the other hand, K-Means clustering was employed for unsupervised modeling. The results and analysis of over 50 outcome models were discussed using tables, charts, and figures comprehensively. The outlines of main outcomes and findings are:

- Exploring a massive medical expenditure dataset:
    - Insurance coverage variable should have binary values.
    - The medical expenditure dataset is highly skewed.
- Building both supervised and unsupervised learning models.
- C5 predicts healthcare coverage status more accurate than MLP.
- Access to care, age, poverty level of family, and race/ethnicity factors have significant effects on healthcare coverage.
- People with more priority conditions have more health needs, though, are more likely to have coverage.

In addition to public health findings of this study, the outcome prediction models are also good for insurance companies that are seeking potential customers. In other words, the outcome models predict the people who are likely to have insurance coverage due to our definition for insurance coverage variable (0 month coverage are uninsured, 1 to 12 months coverage are insured). When the model predicts someone as insured, it means that person has had 1 to 12 month coverage so has tendency to have insurance coverage.

6.2 Contributions of the Thesis

As a conclusion, the contributions of this study are listed in the following:

- Exploring, preprocessing, and preparing a large dataset retrieved from the Medical Expenditure Panel Survey.
- Employing 2 supervised modeling to predict healthcare coverage based on different combinations of attributes.
- Ranking attributes of each model and discovering efficient factors in healthcare coverage.

- Characterizing main groups of uninsured people in the United States.

6.3 Limitations

In this section we express the limitations that we had during this study. The limitations that we faced can be divided to two parts. First, data restriction in data files; Second, operating limitations in software tools. In continue we state some examples of these two types of limitations.

The nature of health data is highly skewed, especially if it covers the population of a whole country. In other words, various behaviors and patterns exist in a healthcare dataset that in the unsupervised modeling, this fact causes problem in clustering the data. The cluster quality might not be as we expect and in the best case it will be fair. This was one of the data restrictions we faced in the unsupervised learning. Another data restriction was during the aggregation procedure in data preprocessing phase. Some variables don't exist in all data files of MEPS and we had to search for them in other data files. Eventually, by exploring the new data file we tried to merge it with the previous data file to fill out the missing values.

There were also some limitations in operation of some tools in IBM SPSS Modeler. Filling the memory of computer after several data modeling and stop working is one of them. Another one is the limitation in some modeling nodes. For example, some of the modeling nodes couldn't be cross-validated and therefore we had no choice to select the traditional train-test partitioning for them.

6.4 Future Works

Further works to expand the results and methods of this study are suggested as following:

- Health related data modeling can be done by adding the impact of person-level weight variable of MEPS to develop models based on nationally representative estimates, not just having a well-distributed sample.
- Compare the outcomes of intuitive heuristic approach and optimal approach in attribute reduction. The intuitive heuristic approach is the method used in this study, whereas the optimal approach is to experiment all combinations of attributes in each stage of attribute reduction (e.g all combinations of 4 attributes out of 22 for stage 18) and selecting the best combination which has the highest accuracy rate.
- Specify whether a cross-sectional study will have a better result in predicting healthcare coverage (just cross a specific part of the time, e.g. end of calendar years) or a longitudinal study (add the change of situation by passing time to the study). Regarding the results of threshold analysis in this study, longitudinal study can provide dynamics of health insurance and also the results of being uninsured.
- Study the relationship of population in public and/or private health plans with healthcare coverage separately. MEPS have defined the values of insurance coverage summary variable with: public, private, and uninsured. Also, due to the current movement in the US and argues about the private insurance companies, explore different groups and important attributes for current insurance companies and compare them with results of this study.
- Investigate healthcare coverage from the point of view of government or a health system researcher which even lacking one month coverage in a year is not acceptable.

REFERENCES

[1] B. Herring, "The effect of the availability of charity care to the uninsured on the demand for private health insurance," *Journal of Health Economics*, vol. 24, no. 2, pp. 225-252, Mar. 2005.

[2] P. Sethi and M. Jain, "A comparative feature selection approach for the prediction of healthcare coverage," *Communications in Computer and Information Science*, vol. 54, no. 7, pp. 392-403, 2010.

[3] "Health Insurance Historical Tables," *U.S. Census Bureau*. [Online]. Available: http://www.census.gov/hhes/www/hlthins/data/historical/HIB_tables.html. [Accessed: 26-Oct-2011].

[4] "Proposed List Of Demands For Occupy Wall St Movement!," *OccupyWallStreet*, 25-Sep-2011. [Online]. Available: http://occupywallst.org/forum/proposed-list-of-demands-for-occupy-wall-st-moveme/. [Accessed: 07-Oct-2011].

[5] D. Delen, C. Fuller, C. McCann, and D. Ray, "Analysis of healthcare coverage: A data mining approach," *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 995-1003, Mar. 2009.

[6] G. S. Linoff and M. J. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd ed. John Wiley and Sons, 2004.

[7] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. Wiley, 2009.

[8] "Medical Expenditure Panel Survey Household Component Overview," *Agency for Healthcare Research and Quality Website*. [Online]. Available: http://www.meps.ahrq.gov/mepsweb/survey_comp/household.jsp. [Accessed: 06-Sep-2011].

[9] R. A. Cohen, D. M. Makuc, A. B. Bernstein, L. T. Bilheimer, and E. Powell-Griner, "Health insurance coverage trends, 1959-2007: estimates from the National Health Interview Survey," *National Health Statistics Reports*, no. 17, pp. 1-25, Jul. 2009.

[10] C. DeNavas-Walt, B. D. Proctor, and J. C. Smith, "Income, Poverty, and Health Insurance Coverage in the United States: 2010," U.S. Census Bureau, Washington, DC: U.S., Sep. 2011.

[11] P. Fronstin, "Sources of Health Insurance and Characteristics of the Uninsured: Analysis of the March 2000 Current Population Survey," *SSRN eLibrary*.

[12] S. A. Glied and P. C. Borzi, "The Current State of Employment-Based Health Coverage," *Journal of Law, Medicine and Ethics*, vol. 32, p. 404, 2004.

[13] L. Clemans-Cope, B. Garrett, and C. Hoffman, "Changes in Employees'Health Insurance Coverage, 2001–2005," *Henry J. Kaiser Family Foundation*, Oct. 2006.

[14] "The Uninsured in America," BlueCross BlueShield Association, Jan. 2006.

[15] Institute of Medicine, *America's uninsured crisis: consequences for health and health care*. National Academies Press, 2009.

[16] P. J. Veugelers and A. M. Yip, "Socioeconomic disparities in health care use: Does universal coverage reduce inequalities in health?," *Journal of Epidemiology and Community Health*, vol. 57, no. 6, pp. 424 -428, Jun. 2003.

[17] "National health expenditure trends 1975 to 2010." Canadian Institute for Health Information, Oct-2010.

[18] "OECD Health Data 2010." Organisation for Economic Co-operation and Development, Jun-2010.
[19] B. Starfield, "Is US Health Really the Best in the World?," *JAMA: The Journal of the American Medical Association*, vol. 284, no. 4, pp. 483 -485, Jul. 2000.
[20] K. Davis, M. Gold, and D. Makuc, "Access to Health Care for the Poor: Does the Gap Remain?," *Annual Review of Public Health*, vol. 2, pp. 159-182, May 1981.
[21] S. J. Katz, T. P. Hofer, and W. G. Manning, "Physician use in Ontario and the United States: The impact of socioeconomic status and health status.," *Am J Public Health*, vol. 86, no. 4, pp. 520-524, Apr. 1996.
[22] S. J. Katz, T. P. Hofer, and W. G. Manning, "Hospital utilization in Ontario and the United States: the impact of socioeconomic status and health status," *Canadian Journal of Public Health. Revue Canadienne De Santé Publique*, vol. 87, no. 4, pp. 253-256, Aug. 1996.
[23] K. Gorey, E. Holowaty, G. Fehringer, E. Laukkanen, N. Richter, and C. Meyer, "An international comparison of cancer survival: relatively poor areas of Toronto, Ontario and three US metropolitan areas," *Journal of Public Health*, vol. 22, no. 3, pp. 343 -348, 2000.
[24] C. DeNavas-Walt, B. D. Proctor, and J. C. Smith, "Income, poverty, and health insurance coverage in the United States: 2007," U.S. Census Bureau, Washington, DC: U.S., Aug. 2008.
[25] J. Gabel, K. Dhont, H. Whitmore, and J. Pickreign, "Individual Insurance: How Much Financial Protection Does It Provide?," *Health Affairs*, 2002.
[26] R. Helman and R. Christensen, "Findings From the 2003 Health Confidence Survey: Americans Increasingly Worried About Health Care Costs." Employee Benefit Research Institute, Oct-2003.
[27] C. Hoffman and J. Paradise, "Health Insurance and Access to Health Care in the United States," *Annals of the New York Academy of Sciences*, vol. 1136, no. 1, pp. 149-160, Jun. 2008.
[28] A. C. Monheit and J. P. Vistnes, "Race/Ethnicity and Health Insurance Status: 1987 and 1996," *Medical Care Research and Review*, vol. 57, pp. 11-35, Dec. 2000.
[29] Services USDoHaH, "National healthcare disparities report," US Government Printing Office, 2003.
[30] B. Starfield and L. Shi, "The medical home, access to care, and insurance: a review of evidence," *Pediatrics*, vol. 113, no. 5, pp. 1493-1498, May 2004.
[31] Institute of Medicine, "Care Without Coverage: Too Little,Too Late," National Academies Press, Washington, DC, 2002.
[32] W. A. Leigh, "Factors Affecting Health of Men of Color in the United States: An Overview." Joint Center for Political and Economic Studies, Dec-2004.
[33] Q. Ang, W.-dong Wang, B.-ya Zhao, J. Li, and K.-yuan Li, "Application of data mining based on clinical medicine database," in *2010 2nd International Conference on Signal Processing Systems (ICSPS)*, 2010, vol. 3, pp. V3-719-V3-723.
[34] K.-L. Tsui, W. Chiu, P. Gierlich, D. Goldsman, X. Liu, and T. Maschek, "A Review of Healthcare, Public Health, and Syndromic Surveillance," *Quality Engineering*, vol. 20, no. 4, pp. 435-450, 2008.
[35] H. Chen, *Medical informatics: knowledge management and data mining in biomedicine*. Springer, 2005.

[36] S. Glover, C. G. Moore, J. C. Probst, and M. E. Samuels, "Disparities in Access to Care Among Rural Working-Age Adults," *Journal of Rural Health*, vol. 20, no. 3, pp. 193-205, 2004.
[37] S. Sudman, N. M. Bradburn, and N. Schwarz, *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA, US: Jossey-Bass, 1996.
[38] K. Swartz, "Dynamics of People Without Health Insurance: Don't Let the Numbers Fool You," *JAMA: The Journal of the American Medical Association*, vol. 271, no. 1, pp. 64 -66, Jan. 1994.
[39] K. Swartz, J. Marcotte, and T. D. McBride, "Spells without health insurance: the distribution of durations when left-censored spells are included," *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, vol. 30, no. 1, pp. 77-83, 1993.
[40] P. F. Short and V. A. Freedman, "Single women and the dynamics of Medicaid.," *Health Services Research*, vol. 33, no. 5 Pt 1, pp. 1309-1336, Dec. 1998.
[41] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2011.
[42] S. M. Weiss and C. A. Kulikowski, *Computer systems that learn:classification and prediction methods from statistics, neural nets,machine learning, and expert systems*. San Mateo, California: Morgan Kaufman Publishers Inc., 1991.
[43] "IBM SPSS Modeler v.14 Help Documentation." IBM SPSS.
[44] "MEPS-HC Panel Design and Data Collection Process," *Agency for Healthcare Research and Quality Website*. [Online]. Available: http://meps.ahrq.gov/mepsweb/survey_comp/hc_data_collection.jsp. [Accessed: 06-Sep-2011].
[45] "MEPS HC-121 2008 Full Year Consolidated Data File Documentation." Agency for Healthcare Research and Quality, Nov-2010.
[46] "2008 Full Year Consolidated Data File Codebook," *Agency for Healthcare Research and Quality Website*. [Online]. Available: http://meps.ahrq.gov/mepsweb/data_stats/download_data_files_codebook.jsp?PUFId=H121. [Accessed: 06-Sep-2011].
[47] "MEPS HC-112 2007 Medical Conditions Data File Documentation." Agency for Healthcare Research and Quality, Nov-2009.
[48] "Labor Force Statistics from the Current Population Survey," *Bureau of Labor Statistics*. [Online]. Available: http://www.bls.gov/cps/. [Accessed: 13-Sep-2011].
[49] W. Zhong, B. Raahemi, and J. Liu, "Learning on Class Imbalanced Data to Classify Peer-to-Peer Applications in IP Traffic using Resampling Techniques," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, 2009, pp. 3548-3554.
[50] S. T. Moturu, W. G. Johnson, and H. Liu, "Predicting Future High-Cost Patients: A Real-World Risk Modeling Application," in *Bioinformatics and Biomedicine, IEEE International Conference on*, Los Alamitos, CA, USA, 2007, vol. 0, pp. 202-208.
[51] C. Hertzman, J. Frank, and R. G. Evans, "Heterogeneities in health status and the determinants of population health," in *Why are some people healthy and others not?: The determinants of health of populations*, A. de Gruyter (New York), 1994.
[52] A. M. Yip, G. Kephart, and P. J. Veugelers, "Individual and neighbourhood determinants of health care utilization: Implications for health policy and resource allocation," *Canadian Journal of Public Health. Revue Canadienne De Santé Publique*, vol. 93, no. 4, pp. 303-307, Aug. 2002.

[53] D. P. Andrulis, "Access to Care Is the Centerpiece in the Elimination of Socioeconomic Disparities in Health," *Annals of Internal Medicine*, vol. 129, no. 5, pp. 412 -416, 1998.
[54] "Poverty Thresholds 2008," *U.S. Census Bureau*. [Online]. Available: http://www.census.gov/hhes/www/poverty/data/threshld/thresh08.html. [Accessed: 22-Oct-2011].

APPENDICES

Appendix I: Poverty Thresholds: United States, 2008

Size of Family Unit	Weighted Average Thresholds	Related children under 18 years								
		None	One	Two	Three	Four	Five	Six	Seven	Eight or more
One person (unrelated individual)	10,991									
Under 65 years	11,201	11,201								
65 years and over	10,326	10,326								
Two people	14,051									
Householder under 65 years	14,489	14,417	14,840							
Householder 65 years and over	13,030	13,014	14,784							
Three people	17,163	16,841	17,330	17,346						
Four people	22,025	22,207	22,570	21,834	21,910					
Five people	26,049	26,781	27,170	26,338	25,694	25,301				
Six people	29,456	30,803	30,925	30,288	29,677	28,769	28,230			
Seven people	33,529	35,442	35,664	34,901	34,369	33,379	32,223	30,955		
Eight people	37,220	39,640	39,990	39,270	38,639	37,744	36,608	35,426	35,125	
Nine people or more	44,346	47,684	47,915	47,278	46,743	45,864	44,656	43,563	43,292	41,624

Unit of amounts: US dollars. **Source:** U.S. Census Bureau [54]

Appendix II: ASCII to CSV program code

```
// ASCII2CSV.cpp : Convert a MEPS's ASCII file to a CSV file based on a
specific Codebook.
// INPUT: year, numOfCols, startColumns.txt, endColumns.txt,
meps_input.txt
// OUTPUT: meps_output.txt

#include <iostream>
#include <fstream>
#include <conio.h>
#include <string>

using namespace std;

const int year = 2006;  // year of survey
const int numOfCols = 2; // number of columns
int columns[numOfCols][2] = {0};

void main()
{
/******** READ CODEBOOK IN ARRAY ********/
      ifstream startColumns, endColumns;
      startColumns.open("startColumns.txt", ios::in); // start columns of
the codebook
      endColumns.open("endColumns.txt", ios::in); // end columns of the
codebook
      if(startColumns.is_open() && endColumns.is_open()){
            for(int i=0; i<numOfCols; i++){
                  string start, end;
                  getline(startColumns, start);
                  getline(endColumns, end);
                  columns[i][0] = atoi(start.c_str());
                  columns[i][1] = atoi(end.c_str());
            }
      }

/******** ADD COMMA BETWEEN DATA ********/
      ifstream meps_input;
      ofstream meps_output;
      meps_input.open("meps_input.txt",ios::in); // ASCII file as input
      meps_output.open("meps_output.txt"); // CSV file as output
      if(meps_input.is_open() && meps_output.is_open()){
            while(!meps_input.eof()){     // it breaks out, but it's
working! if you don't want this break error, replace the condition with
i<numOfSamples in a FOR loop
                  string record;
                  getline(meps_input, record);
                  for(int j=0; j<numOfCols; j++){
                        string data;
                        if(columns[j][0]==0 && columns[j][1]==0)
                              data = "-";
                        else
                              data = record.substr(columns[j][0]-1,
columns[j][1]-columns[j][0]+1);
```

```
                        if(j<numOfCols-1)
                              meps_output << data << ",";
                        else
                              meps_output << data << "," << year <<endl;
                  }
            }
      }
      else cout << "Unable to open file";

      meps_input.close();
      meps_output.close();

      std::cout << "\nPress any key to exit...";
    _getch();
}
```

Appendix III: Aggregate program code

```
// Aggregate.cpp : Aggregate some MEPS data files (CSV) which have one
codebook order to one data file.
// INPUT : numOfFiles, 0.txt, 1.txt, 2.txt, ...
// OUTPUT : meps_aggregated.txt

#include <iostream>
#include <fstream>
#include <conio.h>
#include <string>

using namespace std;

const int numOfFiles = 3; // number of data files
ifstream dataFile[numOfFiles];
ofstream meps_aggregated;

void main()
{
	//for(int i=0; i<numOfFiles; i++){
	//	char* fileName;
	//	itoa(i,fileName,10);
	//	strcat(fileName,".txt");
	//	dataFile[i].open(fileName, ios::in);	// read data files
	//}
	dataFile[0].open("0.txt", ios::in);
	dataFile[1].open("1.txt", ios::in);
	dataFile[2].open("2.txt", ios::in);

	bool inputFilesAreOpen = true;
	for(int i=0; i<numOfFiles; i++)
		if(!dataFile[i].is_open())
			inputFilesAreOpen = false;

	meps_aggregated.open("meps_aggregated.txt"); // CSV file as output
	if(inputFilesAreOpen && meps_aggregated.is_open()){
		for(int i=0; i<numOfFiles; i++){
			while(!dataFile[i].eof()){
				string record;
				getline(dataFile[i], record);
				meps_aggregated << record << endl;
			}
			dataFile[i].close();
		}
	}
	else cout << "Unable to open files";

	meps_aggregated.close();

	std::cout << "\nPress any key to exit...";
    _getch();
}
```

Appendix IV: Insurance Coverage Dataset Codebook

MEPS Household-Component 2006-2007-2008 Insurance Coverage Prediction Data File Codebook											
Variables		Values									
Name	Description	-1	0	1	2	3	4	5	6	7	8
DUPERSID	PERSON ID (DUID + PID)	30002019 to 89688102									
REGION0#	CENSUS REGION AS OF 12/31/0#			Northwest	Midwest	South	West				
MSA0#	METROPOLITAN STATISTICAL AREA AS OF 12/31/0#		Non-MSA	MSA							
AGE0#X	AGE AS OF 12/31/0# (EDITED/IMPUTED)			0-17	18-40	41-65	66-85 (top)				
SEX	SEX			Male	Female						
RACETHNX	RACE/ETHNICITY (EDITED/IMPUTED)			Hispanic	Black-No other race/Not hispanic	Asian-No other race/Not hispanic	Other race/ Not hispanic				
MARRY0#X	MARITAL STATUS-12/31/0# (EDITED/IMPUTED)			Married	Widowed	Divorced	Separated	Never Married	Under 16 - Inapplicable		
FTSTU0#X	STUDENT STATUS IF AGES 17-23 - 12/31/0#	Inapplicable (age)		Full-time	Part-time	Not a student					
EDUCYR	YEARS OF EDUCTION COMPLETED	Inapplicable (age)	No school/ Kindergarten only	1 to 17 (top-coded)							
HIDEG	HIGHEST DEGREE OF EDUCATION			No degree	GED	High school diploma	Bachelor's degree	Master's degree	Doctorate degree	Other degree	Under 16 - Inapplicable
ACTDTY53	MILITARY FULL-TIME ACTIVE DUTY - R5/3			Yes - Active duty	No - Not FT active duty	Under 16 - Inapplicable	Over 59 - Inapplicable				
HONRDC53	HONORABLY DISCHARGED FROM MILITARY			Yes - Honorably discharged	No - Not honorably discharged	16 or younger - Inapplicable	Now active duty				
WAGEP0#X	PERSON'S WAGE INCOME			- 25K	25K - 50K	50K - 75K	75K - 100K	+ 100K			
TTLP0#X	PERSON'S TOTAL INCOME EXCEPT REFUND & SALES INCOME			- 25K	25K - 50K	50K - 75K	75K - 100K	+ 100K			
POVCAT0#	FAMILY INCOME AS % OF POVERTY LINE - CATEGORICAL			Poor/ Negative	Near poor	Low income	Middle income	High income			
RTHLTH53	PERCEIVED HEALTH STATUS - RD 5/3			Excellent	Very good	Good	Fair	Poor			
MNHLTH53	PERCEIVED MENTAL HEALTH STATUS - RD 5/3			Excellent	Very good	Good	Fair	Poor			
PCOUNT	PRIORITY CONDITION COUNT	0 to 10									
IADLHP53	INSTRUMENTAL ACTIVITIES OF DAILY LIVING LIMITATIONS - RD 5/3		No	Yes							
ADLHLP53	ACTIVITIES OF DAILY LIVING LIMITATIONS - RD 5/3		No	Yes							
WLKLIM53	LIMITATION IN PHYSICAL FUNCTIONING-RD5/3		No	Yes							
HAVEUS42	HAVE ACCESS TO USUAL SOURCE OF CARE PROVIDER-R4/2		No	Yes							
EMPST53	EMPLOYMENT STATUS RD 5/3	Inapplicable (age)		Currently employed	Has a job to return to	Employed during the round, not now	Not employed				
INSCOV0#X	COVERED BY HOSP/MED INSURANCE IN 0# (ED)		0 month (without-coverage)	1-12 months (with-coverage)							
PERWT0#F	FINAL PERSON WEIGHT, 200#	423 to 69862									
YEAR	YEAR OF SURVEY	2006 to 2008									

Appendix V: Results of Whole Population Clustering

Input (Predictor) Importance

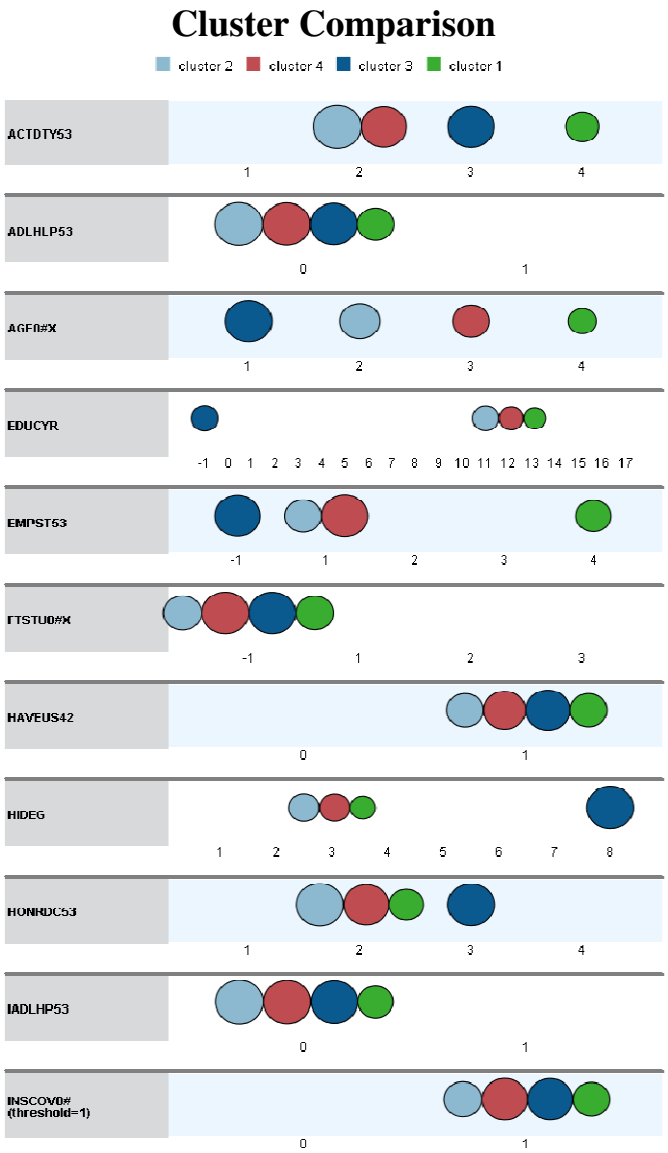
1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-2	cluster-4	cluster-3	cluster-1
Size	28.3% (24661)	27.5% (23981)	26.7% (23266)	17.6% (15342)
Inputs				
	ACTDTY53 2 (99.2%)	ACTDTY53 2 (90.3%)	ACTDTY53 3 (95.0%)	ACTDTY53 4 (74.3%)
	ADLHLP53 0 (99.6%)	ADLHLP53 0 (99.9%)	ADLHLP53 0 (99.7%)	ADLHLP53 0 (92.7%)
	AGE0#X 2 (68.5%)	AGE0#X 3 (59.6%)	AGE0#X 1 (100.0%)	AGE0#X 4 (55.5%)
	EDUCYR 12 (30.1%)	EDUCYR 12 (26.6%)	EDUCYR -1 (33.0%)	EDUCYR 12 (34.5%)
	EMPST53 1 (57.3%)	EMPST53 1 (95.7%)	EMPST53 -1 (95.0%)	EMPST53 4 (86.7%)
	FTSTU0#X -1 (63.6%)	FTSTU0#X -1 (97.2%)	FTSTU0#X -1 (100.0%)	FTSTU0#X -1 (100.0%)
	HAVEUS42 1 (58.6%)	HAVEUS42 1 (78.2%)	HAVEUS42 1 (89.9%)	HAVEUS42 1 (93.4%)
	HIDEG 3 (40.9%)	HIDEG 3 (41.4%)	HIDEG 8 (100.0%)	HIDEG 3 (45.0%)
	HONRDC53 2 (97.2%)	HONRDC53 2 (90.8%)	HONRDC53 3 (100.0%)	HONRDC53 2 (83.2%)
	IADLHP53 0 (99.0%)	IADLHP53 0 (99.6%)	IADLHP53 0 (99.8%)	IADLHP53 0 (86.4%)
	INSCOV0# (threshold=1) 1 (63.7%)	INSCOV0# (threshold=1) 1 (91.4%)	INSCOV0# (threshold=1) 1 (93.1%)	INSCOV0# (threshold=1) 1 (94.4%)

Part 1

Cluster	cluster-2	cluster-4	cluster-3	cluster-1
	MARRY0#X 5 (52.8%)	MARRY0#X 1 (68.5%)	MARRY0#X 6 (95.0%)	MARRY0#X 1 (55.1%)
	MNHLTH53 1 (37.0%)	MNHLTH53 1 (42.4%)	MNHLTH53 1 (52.2%)	MNHLTH53 3 (34.6%)
	PCCOUNT 0 (64.9%)	PCCOUNT 0 (39.0%)	PCCOUNT 0 (88.7%)	PCCOUNT 3 (20.5%)
	POVCAT0# 1 (28.9%)	POVCAT0# 5 (59.3%)	POVCAT0# 1 (29.8%)	POVCAT0# 4 (28.6%)
	RACETHNX 1 (39.3%)	RACETHNX 4 (64.9%)	RACETHNX 4 (40.7%)	RACETHNX 4 (66.2%)
	RTHLTH53 2 (31.7%)	RTHLTH53 2 (38.0%)	RTHLTH53 1 (48.1%)	RTHLTH53 3 (33.0%)
	SEX 2 (58.5%)	SEX 1 (55.4%)	SEX 1 (51.1%)	SEX 2 (62.1%)
	TTLP0#X 1 (97.8%)	TTLP0#X 2 (53.8%)	TTLP0#X 1 (100.0%)	TTLP0#X 1 (78.7%)
	WAGEP0#X 1 (99.1%)	WAGEP0#X 2 (54.4%)	WAGEP0#X 1 (100.0%)	WAGEP0#X 1 (97.9%)
	WLKLIM53 0 (97.6%)	WLKLIM53 0 (95.7%)	WLKLIM53 0 (99.9%)	WLKLIM53 0 (60.3%)
	REGION0# 3 (40.3%)	REGION0# 3 (33.5%)	REGION0# 3 (37.8%)	REGION0# 3 (40.8%)
	MSA0# 1 (84.6%)	MSA0# 1 (86.9%)	MSA0# 1 (84.6%)	MSA0# 1 (76.8%)

Part 2


Cluster Comparison
cluster 2
cluster 4
cluster 3
cluster 1
ACTDTY53
ADLHLP53
AGE0#X
EDUCYR
EMPST53
FTSTU0#X
HAVEUS42
HIDEG
HONRDC53
IADLHP53
INSCOV0#
(threshold=1)

Part 1


MARRY0#X
MNHLTH53
PCCOUNT
POVCAT0#
RACETHNX
RTHLTH53
SEX
TTLP0#X
WAGEP0#X
WKLIM53
REGION0#
MSA0#

Part 2

Appendix VI: Attribute Ranking of C5 Modeling

Stages 0 to 10:

	Attribute Ranking of C5 Modeling										
No of Removed Attributes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Recent Removed Attribute	-	TTLP0#X	IADLHP53	MSA0#	RTHLTH53	HONRDC53	WLKLIM53	EMPST53	HIDEG	MNHLTH53	EDUCYR
Accuracy (%)	**80.65**	**80.67**	**80.65**	**80.83**	**80.71**	**80.67**	**80.72**	**80.56**	**80.08**	**79.28**	**78.06**
Rank 1	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X
Rank 2	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42
Rank 3	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#
Rank 4	RACETHNX	RACETHNX	RACETHNX	RACETHNX	RACETHNX	WAGEP0#X	RACETHNX	EDUCYR	WAGEP0#X	WAGEP0#X	RACETHNX
Rank 5	EDUCYR	WAGEP0#X	WAGEP0#X	WAGEP0#X	WAGEP0#X	RACETHNX	WAGEP0#X	REGION0#	RACETHNX	RACETHNX	WAGEP0#X
Rank 6	WAGEP0#X	EDUCYR	EDUCYR	REGION0#	REGION0#	REGION0#	REGION0#	WAGEP0#X	REGION0#	MARRY0#X	MARRY0#X
Rank 7	REGION0#	MARRY0#X	MARRY0#X	EDUCYR	EDUCYR	EDUCYR	PCCOUNT	RACETHNX	PCCOUNT	REGION0#	REGION0#
Rank 8	TTLP0#X	REGION0#	REGION0#	PCCOUNT	PCCOUNT	PCCOUNT	EDUCYR	PCCOUNT	EDUCYR	PCCOUNT	PCCOUNT
Rank 9	MARRY0#X	EMPST53	EMPST53	MARRY0#X	MARRY0#X	SEX	ACTDTY53	MARRY0#X	SEX	ADLHLP53	ADLHLP53
Rank 10	EMPST53	RTHLTH53	PCCOUNT	FTSTU0#X	FTSTU0#X	EMPST53	FTSTU0#X	SEX	MARRY0#X	FTSTU0#X	ACTDTY53
Rank 11	SEX	SEX	HIDEG	SEX	SEX	ACTDTY53	SEX	ACTDTY53	ACTDTY53	ACTDTY53	FTSTU0#X
Rank 12	HIDEG	HIDEG	MNHLTH53	ACTDTY53	EMPST53	ADLHLP53	ADLHLP53	FTSTU0#X	FTSTU0#X	SEX	SEX
Rank 13	PCCOUNT	WLKLIM53	RTHLTH53	EMPST53	MNHLTH53	FTSTU0#X	MNHLTH53	ADLHLP53	ADLHLP53	EDUCYR	-
Rank 14	RTHLTH53	HONRDC53	FTSTU0#X	ADLHLP53	ACTDTY53	HIDEG	MARRY0#X	MNHLTH53	MNHLTH53	-	-
Rank 15	MNHLTH53	FTSTU0#X	WLKLIM53	WLKLIM53	ADLHLP53	MARRY0#X	HIDEG	HIDEG	-	-	-
Rank 16	IADLHP53	ADLHLP53	SEX	MNHLTH53	WLKLIM53	MNHLTH53	EMPST53	-	-	-	-
Rank 17	MSA0#	MNHLTH53	ADLHLP53	HONRDC53	HIDEG	WLKLIM53	-	-	-	-	-
Rank 18	ACTDTY53	PCCOUNT	HONRDC53	HIDEG	HONRDC53	-	-	-	-	-	-
Rank 19	ADLHLP53	ACTDTY53	ACTDTY53	RTHLTH53	-	-	-	-	-	-	-
Rank 20	HONRDC53	MSA0#	MSA0#	-	-	-	-	-	-	-	-
Rank 21	WLKLIM53	IADLHP53	-	-	-	-	-	-	-	-	-
Rank 22	FTSTU0#X	-	-	-	-	-	-	-	-	-	-

Stages 11 to 21:

	Attribute Ranking of C5 Modeling										
No of Removed Attributes	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Recent Removed Attribute	SEX	FTSTU0#	MARRY0#X	ADLHLP53	REGION0#	ACTDTY53	PCCOUNT	WAGEP0#X	RACETHNX	POVCAT0#	AGE0#X
Accuracy (%)	**78.09**	**77.41**	**76.43**	**76.44**	**75.83**	**75.68**	**75.41**	**75.02**	**74.16**	**69.48**	**69.19**
Rank 1	AGE0#X	AGE0#X	AGE0#X	HAVEUS42	HAVEUS42	AGE0#X	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42
Rank 2	HAVEUS42	HAVEUS42	HAVEUS42	AGE0#X	AGE0#X	HAVEUS42	AGE0#X	AGE0#X	AGE0#X	AGE0#X	-
Rank 3	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	POVCAT0#	-	-
Rank 4	RACETHNX	RACETHNX	RACETHNX	RACETHNX	RACETHNX	RACETHNX	RACETHNX	RACETHNX	-	-	-
Rank 5	WAGEP0#X	WAGEP0#X	WAGEP0#X	WAGEP0#X	WAGEP0#X	WAGEP0#X	WAGEP0#X	-	-	-	-
Rank 6	MARRY0#X	REGION0#	PCCOUNT	PCCOUNT	PCCOUNT	PCCOUNT	-	-	-	-	-
Rank 7	REGION0#	PCCOUNT	ACTDTY53	ACTDTY53	ACTDTY53	-	-	-	-	-	-
Rank 8	PCCOUNT	ADLHLP53	REGION0#	REGION0#	-	-	-	-	-	-	-
Rank 9	ADLHLP53	ACTDTY53	ADLHLP53	-	-	-	-	-	-	-	-
Rank 10	ACTDTY53	MARRY0#X	-	-	-	-	-	-	-	-	-
Rank 11	FTSTU0#X	-	-	-	-	-	-	-	-	-	-
Rank 12	-	-	-	-	-	-	-	-	-	-	-
Rank 13	-	-	-	-	-	-	-	-	-	-	-
Rank 14	-	-	-	-	-	-	-	-	-	-	-
Rank 15	-	-	-	-	-	-	-	-	-	-	-
Rank 16	-	-	-	-	-	-	-	-	-	-	-
Rank 17	-	-	-	-	-	-	-	-	-	-	-
Rank 18	-	-	-	-	-	-	-	-	-	-	-
Rank 19	-	-	-	-	-	-	-	-	-	-	-
Rank 20	-	-	-	-	-	-	-	-	-	-	-
Rank 21	-	-	-	-	-	-	-	-	-	-	-
Rank 22	-	-	-	-	-	-	-	-	-	-	-

* IBM SPSS Modeler has shown the very important and important variables using dark blue and light blue highlights respectively.

Appendix VII: Attribute Ranking of MLP Modeling

Stages 0 to 10:

	Attribute Ranking of MLP Modeling										
No of Removed Attributes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Recent Removed Attribute	-	TTLP0#X	MSA0#	IADLHP53	ADLHLP53	SEX	WLKLIM53	RTHLTH53	MNHLTH53	EMPST53	HONRDC53
Accuracy (%)	**76.2**	**76.3**	**76.3**	**76.4**	**75.9**	**75.5**	**76.4**	**75.8**	**76.6**	**76.7**	**76.4**
Rank 1	HAVEUS42	AGE0#X	AGE0#X	AGE0#X	AGE0#X	EDUCYR	AGE0#X	HAVEUS42	AGE0#X	AGE0#X	AGE0#X
Rank 2	AGE0#X	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	HAVEUS42	AGE0#X	HAVEUS42	HAVEUS42	PCCOUNT
Rank 3	RACETHNX	POVCAT0#	EDUCYR	RACETHNX	POVCAT0#	AGE0#X	EDUCYR	POVCAT0#	FTSTU0#X	EDUCYR	EDUCYR
Rank 4	IADLHP53	EDUCYR	RACETHNX	WAGEP0#X	WAGEP0#X	HIDEG	PCCOUNT	PCCOUNT	EDUCYR	WAGEP0#X	HAVEUS42
Rank 5	WAGEP0#X	HIDEG	ACTDTY53	PCCOUNT	HIDEG	POVCAT0#	RACETHNX	ACTDTY53	POVCAT0#	PCCOUNT	POVCAT0#
Rank 6	EDUCYR	WAGEP0#X	WAGEP0#X	HIDEG	EDUCYR	PCCOUNT	WAGEP0#X	WAGEP0#X	WAGEP0#X	RACETHNX	RACETHNX
Rank 7	HIDEG	RACETHNX	HIDEG	EDUCYR	RACETHNX	RACETHNX	FTSTU0#X	EDUCYR	RACETHNX	HIDEG	HIDEG
Rank 8	POVCAT0#	MARRY0#X	PCCOUNT	FTSTU0#X	PCCOUNT	WAGEP0#X	POVCAT0#	RACETHNX	PCCOUNT	POVCAT0#	WAGEP0#X
Rank 9	MARRY0#X	ACTDTY53	POVCAT0#	POVCAT0#	ACTDTY53	EMPST53	HIDEG	MARRY0#X	ACTDTY53	MARRY0#X	FTSTU0#X
Rank 10	PCCOUNT	FTSTU0#X	FTSTU0#X	ACTDTY53	REGION0#	FTSTU0#X	ACTDTY53	HIDEG	HIDEG	REGION0#	REGION0#
Rank 11	ACTDTY53	PCCOUNT	MARRY0#X	REGION0#	EMPST53	ACTDTY53	REGION0#	EMPST53	MARRY0#X	FTSTU0#X	MARRY0#X
Rank 12	EMPST53	REGION0#	REGION0#	MARRY0#X	MARRY0#X	MARRY0#X	MARRY0#X	HONRDC53	REGION0#	ACTDTY53	ACTDTY53
Rank 13	TTLP0#X	MNHLTH53	HONRDC53	EMPST53	FTSTU0#X	REGION0#	MNHLTH53	REGION0#	HONRDC53	HONRDC53	-
Rank 14	REGION0#	EMPST53	MNHLTH53	WLKLIM53	HONRDC53	HONRDC53	EMPST53	FTSTU0#X	EMPST53	-	-
Rank 15	FTSTU0#X	IADLHP53	RTHLTH53	RTHLTH53	RTHLTH53	MNHLTH53	HONRDC53	MNHLTH53	-	-	-
Rank 16	MNHLTH53	HONRDC53	WLKLIM53	HONRDC53	MNHLTH53	RTHLTH53	RTHLTH53	-	-	-	-
Rank 17	HONRDC53	RTHLTH53	EMPST53	MNHLTH53	WLKLIM53	WLKLIM53	-	-	-	-	-
Rank 18	RTHLTH53	WLKLIM53	SEX	SEX	SEX	-	-	-	-	-	-
Rank 19	WLKLIM53	ADLHLP53	ADLHLP53	ADLHLP53	-	-	-	-	-	-	-
Rank 20	ADLHLP53	SEX	IADLHP53	-	-	-	-	-	-	-	-
Rank 21	MSA0#	MSA0#	-	-	-	-	-	-	-	-	-
Rank 22	SEX	-	-	-	-	-	-	-	-	-	-

Stages 11 to 21:

	Attribute Ranking of MLP Modeling										
No of Removed Attributes	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Recent Removed Attribute	ACTDTY53	MARRY0#X	REGION0#	FTSTU0#	HIDEG	WAGEP0#X	PCCOUNT	RACETHNX	EDUCYR	HAVEUS42	POVCAT0#
Accuracy (%)	**76.5**	**76.1**	**75.7**	**75.7**	**76.2**	**75.7**	**75.3**	**74.75**	**74.2**	**71.3**	**64.5**
Rank 1	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	AGE0#X	-	AGE0#X	AGE0#X	AGE0#X
Rank 2	EDUCYR	EDUCYR	HAVEUS42	HAVEUS42	HAVEUS42	POVCAT0#	HAVEUS42	-	POVCAT0#	POVCAT0#	-
Rank 3	PCCOUNT	HAVEUS42	PCCOUNT	EDUCYR	EDUCYR	EDUCYR	POVCAT0#	-	HAVEUS42	-	-
Rank 4	HAVEUS42	WAGEP0#X	EDUCYR	PCCOUNT	PCCOUNT	HAVEUS42	EDUCYR	-	-	-	-
Rank 5	HIDEG	PCCOUNT	WAGEP0#X	POVCAT0#	POVCAT0#	RACETHNX	RACETHNX	-	-	-	-
Rank 6	WAGEP0#X	POVCAT0#	HIDEG	WAGEP0#X	RACETHNX	PCCOUNT	-	-	-	-	-
Rank 7	RACETHNX	RACETHNX	POVCAT0#	RACETHNX	WAGEP0#X	-	-	-	-	-	-
Rank 8	POVCAT0#	HIDEG	RACETHNX	HIDEG	-	-	-	-	-	-	-
Rank 9	FTSTU0#X	FTSTU0#X	FTSTU0#X	-	-	-	-	-	-	-	-
Rank 10	REGION0#	REGION0#	-	-	-	-	-	-	-	-	-
Rank 11	MARRY0#X	-	-	-	-	-	-	-	-	-	-
Rank 12	-	-	-	-	-	-	-	-	-	-	-
Rank 13	-	-	-	-	-	-	-	-	-	-	-
Rank 14	-	-	-	-	-	-	-	-	-	-	-
Rank 15	-	-	-	-	-	-	-	-	-	-	-
Rank 16	-	-	-	-	-	-	-	-	-	-	-
Rank 17	-	-	-	-	-	-	-	-	-	-	-
Rank 18	-	-	-	-	-	-	-	-	-	-	-
Rank 19	-	-	-	-	-	-	-	-	-	-	-
Rank 20	-	-	-	-	-	-	-	-	-	-	-
Rank 21	-	-	-	-	-	-	-	-	-	-	-
Rank 22	-	-	-	-	-	-	-	-	-	-	-

* IBM SPSS Modeler has shown the very important and important variables using dark blue and light blue highlights respectively.

Appendix VIII: Results of Uninsured Population Clustering

Input (Predictor) Importance

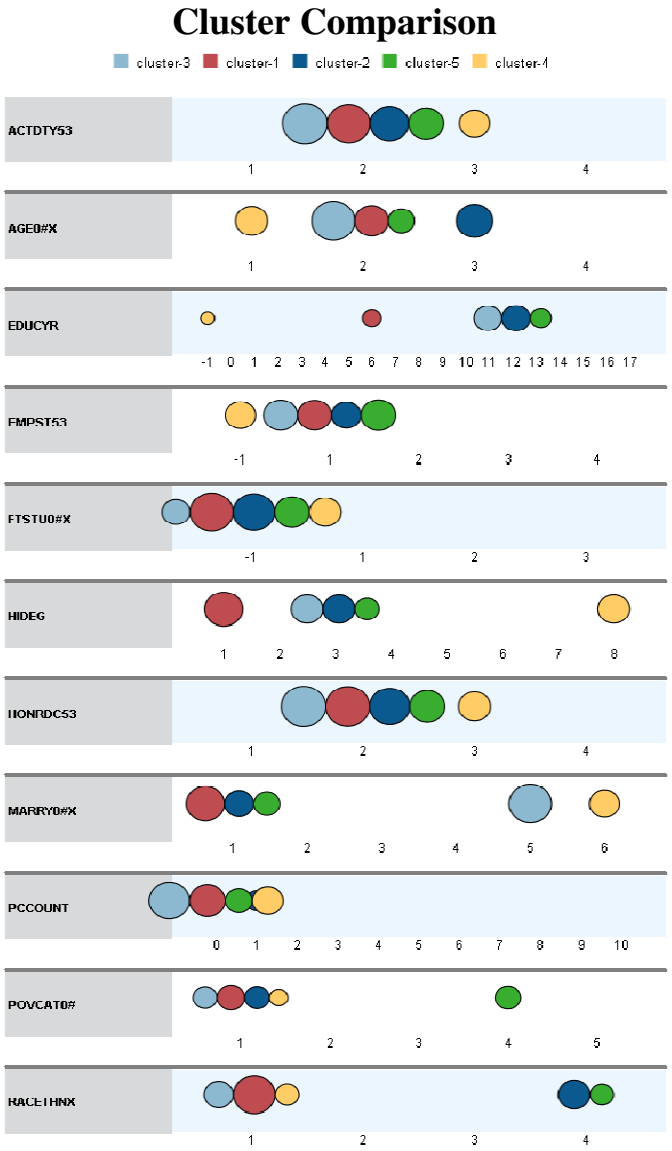
1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-3	cluster-1	cluster-2	cluster-5	cluster-4
Size	25.2% (3398)	24.2% (3254)	21.7% (2927)	16.1% (2170)	12.8% (1717)
Inputs	ACTDTY53 2 (99.8%)	ACTDTY53 2 (94.7%)	ACTDTY53 2 (85.4%)	ACTDTY53 2 (94.9%)	ACTDTY53 3 (90.5%)
	AGE0#X 2 (92.0%)	AGE0#X 2 (50.1%)	AGE0#X 3 (75.0%)	AGE0#X 2 (52.7%)	AGE0#X 1 (100.0%)
	EDUCYR 12 (38.2%)	EDUCYR 6 (18.0%)	EDUCYR 12 (46.9%)	EDUCYR 12 (34.3%)	EDUCYR -1 (19.0%)
	EMPST53 1 (58.1%)	EMPST53 1 (59.0%)	EMPST53 1 (50.1%)	EMPST53 1 (93.5%)	EMPST53 -1 (90.5%)
	FTSTU0#X -1 (39.0%)	FTSTU0#X -1 (96.4%)	FTSTU0#X -1 (99.0%)	FTSTU0#X -1 (93.5%)	FTSTU0#X -1 (100.0%)
	HIDEG 3 (50.2%)	HIDEG 1 (75.8%)	HIDEG 3 (58.3%)	HIDEG 3 (46.0%)	HIDEG 8 (100.0%)
	HONRDC53 2 (98.1%)	HONRDC53 2 (99.6%)	HONRDC53 2 (92.7%)	HONRDC53 2 (95.1%)	HONRDC53 3 (100.0%)
	MARRY0#X 5 (90.4%)	MARRY0#X 1 (75.4%)	MARRY0#X 1 (48.5%)	MARRY0#X 1 (54.0%)	MARRY0#X 5 (90.5%)
	PCCOUNT 0 (81.6%)	PCCOUNT 0 (66.0%)	PCCOUNT 1 (29.4%)	PCCOUNT 0 (56.4%)	PCCOUNT 0 (94.7%)
	POVCAT0# 1 (29.0%)	POVCAT0# 1 (39.2%)	POVCAT0# 1 (35.4%)	POVCAT0# 4 (50.7%)	POVCAT0# 1 (33.3%)
	RACETHNX 1 (42.8%)	RACETHNX 1 (90.3%)	RACETHNX 4 (58.3%)	RACETHNX 4 (43.0%)	RACETHNX 1 (56.3%)

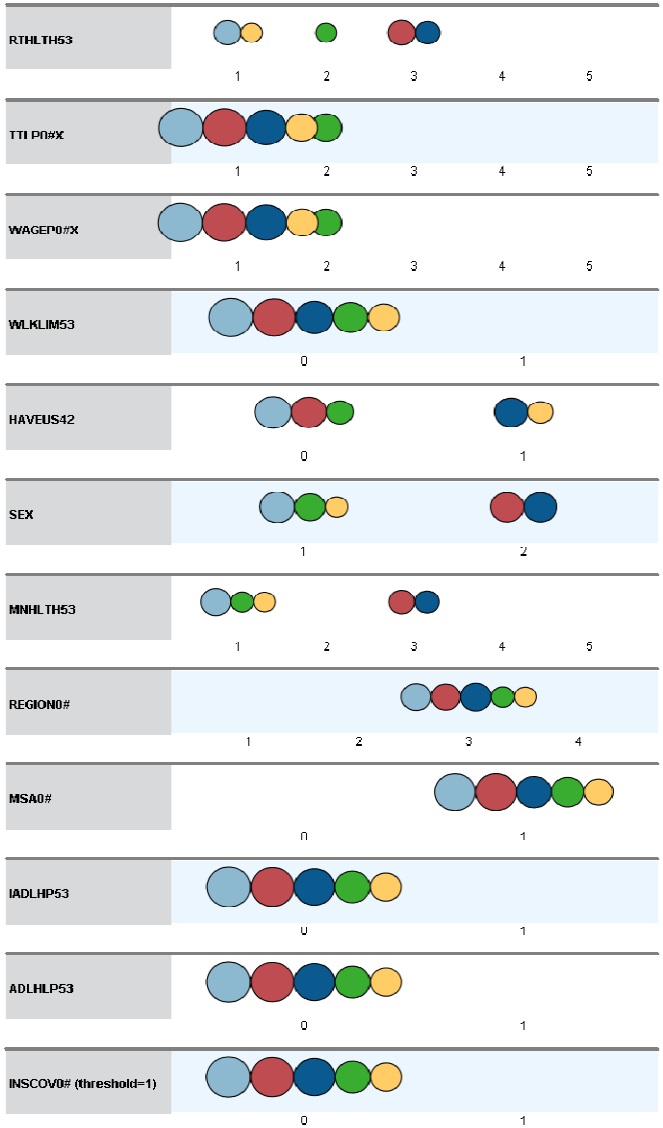
Part 1

Cluster	cluster-3	cluster-1	cluster-2	cluster-5	cluster-4
	RTHLTH53 1 (36.9%)	RTHLTH53 3 (40.1%)	RTHLTH53 3 (36.7%)	RTHLTH53 2 (35.0%)	RTHLTH53 1 (44.8%)
	TTLP0#X 1 (99.1%)	TTLP0#X 1 (98.6%)	TTLP0#X 1 (93.6%)	TTLP0#X 2 (79.0%)	TTLP0#X 1 (100.0%)
	WAGEP0#X 1 (99.7%)	WAGEP0#X 1 (99.7%)	WAGEP0#X 1 (98.2%)	WAGEP0#X 2 (77.0%)	WAGEP0#X 1 (100.0%)
	WLKLIM53 0 (98.9%)	WLKLIM53 0 (98.3%)	WLKLIM53 0 (92.9%)	WLKLIM53 0 (96.9%)	WLKLIM53 0 (99.8%)
	HAVEUS42 0 (67.6%)	HAVEUS42 0 (65.9%)	HAVEUS42 1 (66.2%)	HAVEUS42 0 (57.9%)	HAVEUS42 1 (63.6%)
	SEX 1 (52.1%)	SEX 2 (59.0%)	SEX 2 (65.6%)	SEX 1 (74.1%)	SEX 1 (52.6%)
	MNHLTH53 1 (46.5%)	MNHLTH53 3 (36.7%)	MNHLTH53 3 (35.7%)	MNHLTH53 1 (40.5%)	MNHLTH53 1 (49.0%)
	REGION0# 3 (45.6%)	REGION0# 3 (45.1%)	REGION0# 3 (56.7%)	REGION0# 3 (42.3%)	REGION0# 3 (50.0%)
	MSA0# 1 (85.6%)	MSA0# 1 (89.8%)	MSA0# 1 (71.7%)	MSA0# 1 (86.2%)	MSA0# 1 (84.3%)
	IADLHP53 0 (99.7%)	IADLHP53 0 (99.7%)	IADLHP53 0 (97.2%)	IADLHP53 0 (99.7%)	IADLHP53 0 (99.9%)
	ADLHLP53 0 (99.9%)	ADLHLP53 0 (99.8%)	ADLHLP53 0 (98.8%)	ADLHLP53 0 (100.0%)	ADLHLP53 0 (99.7%)
	INSCOV0# (threshold=1) 0 (100.0%)	INSCOV0# (threshold=1) 0 (100.0%)	INSCOV0# (threshold=1) 0 (100.0%)	INSCOV0# (threshold=1) 0 (100.0%)	INSCOV0# (threshold=1) 0 (100.0%)

Part 2


Cluster Comparison
cluster-3
cluster-1
cluster-2
cluster-5
cluster-4
ACTDTY53
AGE0#X
EDUCYR
FMPST53
FTSTU0#X
HIDEG
IIONRDC53
MARRY0#X
PCCOUNT
POVCAT0#
RACETHNX

Part 1


RTHLTH53
TTLP0#X
WAGEP0#X
WLKLIM53
HAVEUS42
SEX
MNHLTH53
REGION0#
MSA0#
IADLHP53
ADLHLP53
INSCOV0# (threshold=1)

Part 2