HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Information and Natural Science
Department of Information and Computer Science

Luis Gabriel De Alba Rivera

Modeling and profiling people's way of living: A data mining approach to a health survey

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, November 25, 2009

Supervisor: Professor Erkki Oja
Instructor: Jaakko Hollmén, D.Sc. (Tech.)

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Luis Gabriel De Alba Rivera	
Name of the Thesis:	Modeling and Profiling people's way of living: A data mining approach to a health survey	
Date:	November 25, 2009	**Number of pages:** ix + 84
Department:	Department of Information and Computer Science	
Professorship:	T-61 Computer and Information Science	
Supervisor:	Professor Erkki Oja	
Instructor:	Jaakko Hollmén, D.Sc. (Tech.)	

In this work a data set consisting of several variables and hundreds of thousands of records is analyzed using different methodologies. The data is the result of a survey that valuates different areas and characteristics that affect directly or indirectly the health and life styles of the Finnish population. The survey was organized by various institutions from Finland.

Firstly, the data set is analyzed, parsed and organized in an effective structure. This process consists in recovering the valid answers from the survey, cleaning the data, removing outliers and filling up the database. The database is structured in a way that facilitates access and analysis of complex queries.

Secondly, the following step consists on a data exploratory analysis where the data is studied by means of graphical visualization and basic statistics. During this process the variables are analyzed, alone and combined, finding relevant information in them. Depending on the involved variables specific plots are generated. Restrictive queries are used to extract significant information of relevant parameters.

Thirdly, after the exploratory analysis the variables are modeled using different approaches. The objective is to come up with the best set of regression parameters to represent each independent variable. Simple linear and polynomial regression, variable selection algorithms as SISAL, neural network structures and principal components transformations are used for this purpose. The attributes of the models are compared to mean predictors of all variables to confirm improvements. Special attention is put on variables with the most missing points.

In conclusion, the exploratory analysis shows how fast and simple it is to extract relevant information from a data set if it is structured correctly; interesting facts are discovered by this approach. The regression analysis is an appropriate methodology to model the variables. The models show a considerable improvement over the mean predictor; for some variable notably simple setups are sufficient. Results for all methods are displayed in easy to follow tables. Strategies for future work are also presented.

Keywords: modeling, profiling, data mining, exploratory analysis, visualization, machine learning, regression, variable selection, healthcare study, life style, survey.

Acknowledgments

My Master's thesis research project has been carried out in the Parsimonious Modelling group from the Helsinki Institute of Information Technology at TKK. More than everything it has been a learning experience. I want to thank my lab partners Mika, Mikko, Janne and Miguel for their willingness to help. Big thanks also go to my instructor Jaakko Hollmén for trusting in me for this project. His dedication, simple solutions and interesting comments made the work more fun.

Professor Erkki Oja and post-doc Tapani Raiko were always supportive, providing information, signing papers and helping me with the burdens of my scholarship. Each one of the professors and teachers that instructed me during this two years in the MACADAMIA program I thank you. I may not be able to see further than the giants but I am able to see a piece of what the giants see and that's amazing.

Special thanks also go to my master degree mates László, Dušan and Stevan, for all the ideas, experiences and *code* we shared. The new generation guys Li, Prem and Agha because the best way to learn something is by explaining it; I really enjoyed that you trusted me in so many things.

Finland, the great country I have the opportunity to study at. I learned a lot about its high standards always aiming for equality and prosperity.

My gratitude also goes to my two families, my mom Lucero, dad Luis and my *new* mom and dad Noemi and Armando. My brother Pepe, sisters Lucy, Lulu and Martha; because you all make the journey easier. Although your addiction to Finnish chocolate is impossible to maintain.

Finally, I would like from my hearth to thank my wife Aime who has always supported me with belief and love. Walking always together under the sun, under the moon; all this is for you.

Espoo, November 25th 2009

Luis Gabriel De Alba Rivera

Solo es digno de la felicidad quien todos los días se esfuerza por conquistarla.
Daniel El Viejo

Luis Gabriel De Alba Rivera was supported by the Programme AlBan, the European Union Programme of High Level Scholarships for Latin America, scholarship No. E07M402627MX

Contents

Abbreviations and Notations

AI	Artificial Intelligence
ANN	Artificial Neural Network
BMI	Body Mass Index
DEA	Data Exploratory Analysis
DM	Data Mining
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
RSS	Residual Sum of Squares
SISAL	Sequential Input Selection Algorithm
SQL	Structured Query Language
SSE	Sum of Squared Errors
THL	Terveyden ja Hyvinvoinnin Laitos
	(National Institute for Health and Welfare)
mmHg	Millimeters of Mercury (to measure blood pressure)
mmol/L	Millimoles per Liter (to measure elements in blood)
$\mathbf{X}$	$N \times d$ input data matrix
$\mathbf{y}$	$N \times 1$ output data matrix
β	Regression coefficients of a single response linear model
$\mathbf{x}$	Vector of observations of d variables $\mathbf{x} = [x_1, \ldots, x_d]$
y	An observation of an output variable
$\hat{y}$	An estimate of the observation y
ε	Random error

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays it is known that data is everywhere; not only bits of data but really large amounts of it. Nevertheless, the majority of the data observed today has always been there but we, as humans, were not able to collect it or realize it existed. Lately, however, the recollection systems have improved and we are *observing*, *handling* and *working* with these immense piles of information. Today we are able to record incredible amounts of data in just a few second, e.g. the new Synoptic Survey Telescope being constructed in northern Chile will be generating 340MB of data per second by 2015[1]; and saving the data in continuously-improving smaller and more secure devices as, e.g. flash memories, is even more possible.

Furthermore, in the industry, commerce, medical centers, etc., there are advanced machines with precise measurement equipment updating databases constantly. For a long time those measurements, or data, have been saved and used for many business-related purposes. But just recently has a further approach been taken with large datasets, a decision to further collect and analyze the extracted valuable information hiding behind the original purpose of the data. From astronomy to psychology there are new and bast datasets waiting to be analyzed. However, the old approaches are not good enough and new methodologies are needed. Each dataset is, at some aspects, different to others, and traditional technology and approaches are not good, fast or even viable enough.

Data mining is a new research discipline that took off just a few years ago with the aim of solving these problems. Data Mining (DM) is defined in [7] as:

> "The science of extracting useful information from large datasets or databases"

DM combines Computer Science, Statistics, Artificial Intelligence (AI) and more. From the given definition it is important to notice the concept of *useful information*. On the one hand there is the term *useful*, defined as "Having the character or quality to be of use or utility; suitable for use; advantageous, profitable, beneficial"[2]. On the other hand, the term *information* has been lately used, with some limitations,

[1] Large Synoptic Survey Telescope will generate 30TB of data per night, see [17].
[2] USEFUL at Oxford English Dictionary, see [23].

as a synonym of knowledge. Therefore, data mining can be thought as the science of extracting beneficial knowledge from data. Thus knowledge, in this context, is what is required to make informed decisions [22].

Let's rephrase this last statement with an example. If one knows that it is "cloudy" this information can be considered as limited knowledge. It is limited because there is no further information related with the fact that it is cloudy, by itself only an atmospheric state. Nevertheless, if one extracts more information, let's say that it may be rainy in the evening, the limited knowledge is complemented and it becomes valid and strong enough to be of use. For example, by deciding to do the grocery shopping in the morning instead of the evening. The same happens with data analysis; many times large analysis efforts lead to limited or, even, useless knowledge. However, at some point, the link to previous results may show up or better yet, a specific set of information pays off for all the previous work. Another comparison of data mining is the mining of precious metals, e.g. gold. It is necessary to extract and process large quantities of ore in order to obtain a few grams of gold [22]. Thus in data mining large quantities of data have to be *mined* in to find interesting results.

Although data mining is still largely seen as a new field, there is plenty of industries already benefiting from it such as: banks, finance institutions, telecommunications companies, etc. [22]. In the area of telecommunications, for example, there has been plenty of studies done on fraud detection. There is a high interest from companies to detect in advance those who are obtaining un-billable service and/or undeserved fees. With correct fraud detection systems the companies in the industry of telecommunication shall be able to reduce the large losses they suffer annually [11].

Machine learning (ML) is the discipline of programming computers to optimize a performance criterion using example data or past experience [2]. The application of machine learning methods to large dataset can be seen as data mining [2]. ML is a part of the AI field, as it is expected for the systems to learn and change with respect to the different environments they are presented with, in this case, the data. In the ML field, searching for patterns in the data has been a fundamental problem in history and the continuous search has lead to successful results and discoveries [3].

The required input for a DM/ML analysis is the data. The data may come from one of two fronts. It could be that the data is the result of a specialized study where it was strictly captured for the purpose of analysis, therefore, simplifying some parts of the DM/ML process. Or, the data is the result of a process where information was captured for other purposes than further analysis. For example, for a bank transaction data is collected during the operation of the system used to keep a log of the bank movements, i.e. a different purpose than a specialized data analysis. However, the dataset may turn out be large and interesting enough to be analyzed with a set of DM/ML steps.

Models, or patters, are some of the usual results of a DM/ML analysis [7]. The models may be predictive (used to make predictions of future inputs) or descriptive

(to gain knowledge from the data) or both [2].

This thesis is the result of a data mining and machine learning approach to a large dataset of life styles in the Finnish population. The project included many, if not all, necessary steps to accomplish a small-scale but complete data mining analysis where interesting and useful results were obtained. Collaterally to the data mining work the recovering of missing variables using the presence of others by means of linear regression was also studied.

R, see [27], has been used as the de facto software tool for all the data analysis within the thesis. Perl has been an auxiliary language mainly used for scripting, e.g. pre-processing, data parsing, simplifying the creation of consults, etc. All the data its being handled in a SQLite, see [10], database with a defined structure based on three tables.

1.1 Outline

The work done for this thesis followed a similar approach as the one stated in [7], consisting on multiple steps. Initially, the dataset was understood, reviewed, the different variables were selected and defined; afterwards the parsing process was established. The data-processing step continued with the parsing and clean-up of the data, generating a valid dataset in an easy-to-understand and fast-to-access format.

With a structured dataset the data mining analysis was performed. The initial step consisted in an exploratory visualization analysis of the data; some relationships were set and some findings were done. The exploratory analysis was followed by a modeling and regression analysis used to establish basic relationships between missing and available variables. The regression was used for all variables but focused on those with the most missing elements. The methods used were mean predictor, forward stepwise selection, SISAL, neural networks and principal component analysis.

The study concludes with the analysis of the different findings during the data mining approach, the visual exploratory analysis, the recovering of the missing variables and the quality and usefulness of the results.

Chapter 2

Background

In this thesis a dataset is analyzed the later being the result of a web test prepared by various institutions from Finland. The dataset contains information regarding the life styles or life's risk the Finnish population has at present time. The original test was presented as an interactive part of a television show called *Elämä Pelissä*[1]. The television show was used as an incentive mechanism for the population to go and answer the web test form where important factors were captured. The web test was answered by approximately half a million Finns. Although this is a large number, the test does not assure the independence of answers, i.e. one person could have answered the test several times. Nevertheless, half a million people account for almost 10% of the total Finnish population. The television show was a part of the Finnish broadcasting company YLE during the year 2007. It was created by YLE with the support of Duodecim and the Finnish National Institute for Health and Welfare (THL in Finnish).

Duodecim is a Finnish medical society playing an important role in the Finnish national health care they have cooperation with different organization and officials [4]. Duodecim has supported this research by providing the author of the thesis with the original dataset.

Another participant supporting the research is THL. A research unit inside THL has been doing the FINRISK study since 1972, this study takes place every 5 years. The study is based on a survey on risk factors of chronic noncommunicable diseases. It uses random and representative population samples from all Finland. Datasets resulting of the FINRISK surveys are used for a number of different research projects and for national health monitoring needs [29]. Besides the contribution of THL to the development of the on-line test they also contributed with the algorithm that calculates the life expectancy of a person given a set of input variables. The FINRISK life expectancy algorithm is based on 37+ inputs, i.e. answers to different questions. Thus the users filling up the *Elämä Pelissä* test are able to predict their life expectancy.

[1] Elämä Pelissä can be roughly translated in English as: *Life at stake.*

2.1 Elämä Pelissä : The test

In Finland the life expectancy 100 years ago was just above 40 years; now that expectancy has almost doubled [14]. Individuals are free to take their own decision and select their own choices on how their health and life expectancy is affected. Most of the Finnish people already know, or think that they know, how to improve their life and avoid health problems. However, many things that are considered important are not that important or have not been proved important, e.g. eat more vegetables instead of stop smoking [14].

The *Elämä Pelissä* test consists of 39 questions regarding the individual's living conditions, lifestyle, social relationships and experiences in different areas. Accurate answers to these questions will facilitate the prediction of "How long will the individual live?". The test can be taken as many times as wished; this openness is with the sole purpose of showing the individuals what changes in his/her actual life style could improve or worsen his/her health [14].

The test consists of 39 question considered key life factors. Of course the existence of more factors that could affect the prediction is known, e.g. hereditary diseases or genetic issues. Nevertheless, to keep the test accurate and at the same time simple to an average public, specific or personal factors were not taken into account. The impact these factors may play, regarding the length of a life is limited, additionally individuals do not have enough information at hand [14]. For example cancer, in some cases is hereditary while in other cases not; therefore asking a question related to this disease is irrelevant. Other cases, e.g. glucose or sodium content in the blood, are difficult questions to ask as the answers are hardly known by individuals and could not be calculated by them without a laboratory study.

A descriptive list of the questions is in Appendix A. More information about the questions and importance of each one is available in [14].

2.2 Elämä Pelissä : Life expectancy prediction

As previously noticed the *Elämä Pelissä* test is used to do a life forecast, i.e. predict how long a person is going to live according to the given input (answer to questions). The forecast takes into account a large number of factors. The test provides accurate forecasts of life for people between 25 and 74 years of age that do not have any *hard* illness affecting them, e.g. cancer or heart problems [14]. The forecast is calculated based on the average figure of similar responses based on health monitoring. However, the use of the average, although complete and simple, discards the well known individual variations. In general, the individual variations are considered insignificant compared to the majority of the cases; therefore, affecting little the results. It is worth mentioning that the major impact factors are Smoking and Alcohol consumption. Nevertheless, all the factors, even simple ones, contribute to the prediction.

The base average of life expectancy is reported each year by Statistics Finland[2].

[2]Statistics Finland is a Finnish government agency dedicated to produce statistics and informa-

Current age	Life expectancy	
	Women	Men
0	82	76
20	83	76
40	83	77
60	85	80
80	89	87
100	102	102

Table 2.1: Life expectancy for men and women in Finland, according to [14].

The life expectancy works so that the older a person is the older he is expected to live. Table 2.1 exhibits the life expectancy for men and women.

When gathering the information it is normal that some input variables are mistaken or omitted. This is due to typo mistakes or respondents lack of knowledge, e.g. it is not common for people to know their diastole pressure in mmHg. The algorithm will use the population average (mean predictor) or if possible some related questions to predict the erroneous/missing variable.

The problem of missing variables will be one of the topics later explored within this thesis. In the following chapters there are studies on how to do better than the mean predictor by means of other known variables. The mean predictor although simple and in many cases good enough is not the best approach to solve the missing variable problem, specially when a lot of information is available and DM/ML methods could be used. Replacing the missing data with the mean is known to produce biased estimates, see [1], especially if the replaced variable is important for the estimation; therefore, this method shall be avoided if other approaches are available.

A similar project predicting for how long one is going to live has been developed by the University of Pennsylvania in [5]. They proposed a short and a long version of what they call *life calculator*. The data and predictions are focused to Americans (USA); therefore, they claim, the forecast may not be accurate for other nationalities. A similar behavior shall be expected for the *Elämä Pelissä* predictor, i.e. it will work better with Finns than other nationalities. The questions used by the life calculator are similar to those in the *Elämä Pelissä* test. For example, they ask about education, smoking behavior, occupation activities and blood pressure. However, they have a complete different section where they ask questions related to possible hereditary diseases; like heart problems and colorectal cancer, for example.

Another interesting study has been done in [16]. Initially they comment that during the last century the average live expectancy has increased dramatically in the industrialized countries; they claim that in America (USA) the increase has been of almost four fold. This drastic change however, is not an evolutionary change because the number of generations coexisting in a century is too small to have pushed a

tion services for the needs of society `http://www.stat.fi`

genetic change. The claim is that there are other factors affecting the aging process. In their study they model the demographic changes in human population simulating the conditions that expanded humans life expectancy. Therefore, seeking to find which other variables play an important role in the prediction of life expectancy.

Contrary to the previous papers, about an increment in the expectancy of life, in the study presented by the authors of [24] the future may be different. In their study they find out that increasing obesity will play an important role by reducing the life expectancy of people from two to five years, specially in the US. Obesity is related to hearth problems, hypertension, diabetes, and some kinds of cancers. Obesity can be considered an epidemic, two-thirds of American adults are overweight or obese. Therefore, future generations will live less healthy and possibly even shorter lives than their parents. This, although not in-line with a life expectancy predictor, gives important information on how one variable, weight, may play a decisive role.

A parsimonious model of life expectancy is proposed in the research of [18]. In this case the model has been called "subjective life expectancy", i.e. it is based on the peoples belief and not on the objective survival rates (what has been studied and recorded in reality). Initially they noticed that on average young respondents underestimate their true survival probability whereas old subjects overestimate their survival probability. To generate a model based on these issues it is necessary that the Bayesian model should be prone to psychological attitudes. The model developed in their paper allows the possibility of subjective belief that express ambiguity. That is, the survival is downward biased at younger ages and upward biased at older ages. This model is considered as a Bayesian learning model with a psychological bias. The generated model was able to fit the subjective beliefs correctly.

2.3 Elämä Pelissä : Related research

The *Elämä Pelissä* dataset is a vast source of information. Many aspects of it should be of interest for doctors, insurance companies, the government, etc. Of course, the dataset shall not be considered a strict and confident set of answers. The limitations in security and confidentiality that a web survey has put a bias on the obtained results. However, the large number of respondents (half a million) should counter balance the possible existing bias, reducing it to a minimal problem. Some of the answers observed in the test could be compared to more strict studies to provide some confidence. Plenty of research have been done related to health issues in Finland. The following papers were chosen and analyzed because they, in some degree, use the same variables as the ones available in the dataset.

In [12] it is stated that there is a direct association between alcohol consumption and the increase of carotid Intima-Media Thickness (IMT); this independently of age, sex and other factors, e.g. blood pressure, cholesterol levels or BMI. The authors worked with 2,074 individuals for the analysis. The participants were asked to report their alcohol consumption in different categories (similar to those of the *Elämä Pelissä* test) to determinate their total alcohol consumption. To complement the test other lifestyle information was asked, for example, education, smoking,

physical activity, etc. Therefore, the questionnaire used was very similar to the one analyzed in this thesis. The correlations between the different risk factors and IMT were calculated using a stepwise linear regression model. In their findings they observe that for heavy drinkers a significant increase of arterial augmentation is noticeable; supporting the statement that increased alcohol consumption is related to vascular damage in young adults especially among men. Other findings include that the frequency of wine consumption was directly correlated with carotid IMT, independent of age, sex and other cardiovascular risk factors. Nevertheless, they conclude their study was limited due to the small number of subjects drinking wine daily, hence, limiting the previous conclusion.

In another study by the authors of [21] the connection between alcohol-related mortality to age and sex is analyzed. The study estimates the impact of excessive alcohol use on life expectancy by sex. The study assumes that the mortality risk to alcohol-related causes is independent of all other mortality risks. The study was done with the data available from the official death register maintained by Statistics Finland. Findings state that four out of ten alcohol-related deaths were directly attributable to alcohol; another four were accidental and violent deaths with alcohol-related contribution to the cause of death; the remaining two were deaths from disease related to alcohol consumption. Deaths from disease with an alcohol-related contribution were eight times more common among men than women. Alcohol-related accidental and violent deaths were almost 9 times more common among men. An interesting impact of alcohol found by the study is that alcohol reduced the life expectancy in women by 0.4 years and by 2.0 years in men. However, smoking reduces the life expectancy, at age 15, by 2.6 years among men and 0.3 years among women, making smoking a more important cause of life reduction in men and a similar cause in women.

The paper of [9] reports on how smoking affects the quality of life of men and women in Finland. Previous studies have shown that tobacco consumption has a direct impact on the quality of life. The study divided the population into four categories depending on their smoking behavior: daily, occasional, ex-smoker and never-a-smoker. The previous classification question was complemented by asking the individuals an estimation of their quality of life by rating how good their present life has been. Fifteen additional questions related to discomforts were also asked, the answers ranged from 0 to 1. The study was performed separately for men and women. The influential variables used were age, marital status, education and income. The findings show that compared to never-smokers daily smokers were associated with lower scores, i.e. worst feelings; in 9 of the 15 variables tested among men, for example sleeping distress and vitality. Smoking women showed also low values. Consequently, daily smokers have a lower quality of live compared to never smokers among both, men and women. However, the paper concludes that the numerical differences, although existent, are too small to consider smoking as a valid influence on the quality of life.

In the area of physical activities the paper of [28] is a study about physical activity among the Finnish youth. The study initially states that the increase of obesity and

the decrease in cardiorespiratory fitness among young people can be attributed to the large amount of time spent watching TV (or computer screens or playing video games). The study comprises 7,344 children from the northern regions of Finland (Lapland and Oulu). The subjects were asked about their daily physical activities, in ranges of 20 minutes; from 0 minutes to at least 1 hour per day. Corresponding to their sedentary behavior they were asked about the amount spent on watching TV, reading books, at the computer and/or playing video games. The results showed that boys are more physically active than girls; approximately 20% of the subjects participate in moderate to vigorous physical activities. On the contrary, 50% of the subjects do watch TV for more than 2 hours per day. However, the assumption that TV incites inactivity showed some contradiction; in some individuals (43% of boys, 23% of girls) watching 4 or more hours TV also included high levels of physical activities.

Related to diet and food intake there is the study of [25]. In the study it is analyzed the consumption of *polyphenols* [3] in Finnish adults. The polyphenols are known of having a beneficial effect on human health and they also provides protection against chronic and cardiovascular diseases. The study was done to 2,007 adults between 25 and 64 years of age. The daily mean intake of polyphenos is 863mg/d with a higher intake by men than women. It was noticed that in the list of 20 foods with the highest total polyphenol concentrations 16 were berries; in beverages the highest concentrations was found in coffee (Finnish diet is well known for high consumption of berries and coffee).

In the area of work and stress there is the study of [15] where they analyze the relation between Body Mass Index (BMI) and stress at work. Obesity is known to be related to an interaction between biological and environmental factors. Therefore, the claim that the increasing proportion of obese people and the increase in stress related to a working life could be connected. The study covered more than 45,000 employees; besides weight, height and age other variables were studied. These variables included marital status, smoking status, alcohol consumption and physical activity. These variables were used as predictors of the BMI. In the results it was noticed that the BMI value was higher for men than for women; that the BMI mean increased with age and that the BMI was highly related to socioeconomic status as well as a lower job demand.

These formal studies have two things in common. Firstly, all of them were performed in Finland and among Finnish people. Secondly, the questions used to extract information are very similar within them and in comparison with the *Elämä Pelissä* test, e.g. age, weight, stress, blood pressure, etc. Nevertheless, each study focused on a different area like physical activity or smoking behavior. The interest of these studies, regarding the *Elämä Pelissä* dataset, is that they were mainly limited in the number of participants. With the large data available it may be probable to find new tendencies and/or complement the presented ones. The work of finding multidimensional profiles could be improved by the localization of

[3] Polyphenols are a group of chemical substances mainly found in fruits like berries and vegetables like cocoa or coffee.

individual but important factors as, e.g. alcohol or smoking behavior.

The study of this thesis will be, of course, not based in any scientific medical terms, but more in the machine learning data mining science of extracting useful and important information from a large dataset. Of what could be or should be interpreted from the analysis results will depend on the specialists that take a look at them. Some results are expected to be easily interpretable and some may require more specialization.

Chapter 3

Data pre-processing

A dataset is the resulting representation of the measurements taken from the environment or a given process [7]. This collection of measurements can be seen as a $n \times p$ matrix, where n represents the number of objects from where the measurements were taken and where p represents the different measurements recorded. The set of recorded variables describe the underlying process that generated them. In the case of the *Elämä Pelissä* dataset the variables represent the life style of the different people that answered the test. Usually datasets are captured for purpose other than data mining. Data mining is, in many cases, a secondary step; where the dataset is used to extract new and interesting information hiding behind.

3.1 The dataset

The purpose of the *Elämä Pelissä* test is to invite people to play the life prediction game, get to know their health status, life expectancy and act in consequence. All "plays" that have took place have been stored. The data was saved as it was served without any specification. The test transactions have been saved as URLs[1] including the transaction variables[2]. The test consist on 17 forms (web pages) and as each page is sent to be processed it is saved as a unique URL, carrying with it the previously recorded variables and the ones given in the actual page. Therefore, it is expected that in one play all saved URLs but the last one are incomplete because not all values from all variables have been given until the end. Besides the normal variables in the test extra control variables are used. These control variables indicate if the test is complete and/or if it is a replay.

The knowledge on how the dataset has been structured and how it shall be read is not a strict part of the DM/ML approach. However, it is a part of the complete data analysis process strategy. One section in the process is called "Collection of

[1] URL stands for Uniform Resource Locator. URLs are commonly known as web page addresses.

[2] Symbols & and = are used to separate and give variables an individual value, e.g. ?var1=3&var2=5. This approach is used to send and retrieve variables in scripting languages under GET transmission format. Therefore the variables are visible in the URL, contrary to POST format where the variables are send by other methods.

data". The collection of data includes the *How is it going to be collected* and *How is it going to be saved*. Moreover, this section can be extended to cover all the required pre-processing just before starting the strict analysis.

Duodecim provided 16 gzipped files. These files are from 8 servers which saved the test transactions for two months, September and October of 2007. The servers shared the load, therefore all files have similar number of transactions. For September, the initial month when the test was put on-line, each server saved 1,980,000 transactions approximately. For October there were only 70,000 transaction in each server. The uncompressed dataset uses more than 24GB and has more than 16 million transactions.

3.2 Parsing the data

One of the features that differentiates data mining from other types of data analysis is the large amounts of data to process [7]. With 24GB of information it is necessary to organize the data in an efficient way. There are two alternating phases in a data mining approach. In the first phase it is necessary to get the data from the *data source* and in the second phase it is necessary to run analysis algorithms on the extracted data. Because of the continuous switching between these two phases the correct organization of the data will play an important role on the speed with witch the data is analyzed.

As previously stated not all transactions in the provided dataset are *complete* transaction; actually only a small percentage are complete, therefore valid. Hence, a filtering is necessary. The dataset was parsed using a Perl script programmed for this purpose. The parsing program followed 8 steps to handle the data as best possible;

1. A transaction is selected only if it is a full transaction and if it represents the first play of the game, replays are discarded.
2. The valid transaction is *broken* into its different variables.
3. The date, UNIX epoch, of the transaction is calculated according to the available information.
4. A reference to the input file and the row where the transaction is coming from is extracted.
5. The recovered variables are cleaned-up, see sub section 3.2.1.
6. Extra informative variables Body Mass Index (BMI) and Total Alcohol Consumption are calculated.
7. A log of cleaned-up variables is kept as reference.
8. The variables of the transaction, the date, references and log are saved into a SQLite database, see sub section 3.2.2.

Question		Sub questions	
7.	Household size	7.1	Adults (18+ years)
		7.2	Minors (<18 years)
9.	Cholesterol levels	9.1	Total cholesterol
		9.2	HDL cholesterol
10.	Blood pressure	10.1	At systole
		10.2	At diastole
14.	Alcohol use	14.1	Beer III, IV (bottle 1/3 L)
		14.2	Long drink (bottle 1/3 L)
		14.3	Concentrated alcohol (shot 4cl)
		14.4	Wine (glass 12cl)
		14.5	Light wine (glass 12cl)

Table 3.1: Variable expansion; questions 7, 9, 10 and 14 represent 11 variables.

3.2.1 Data clean-up

The on-line test includes some data protection, e.g. limitation of the number of characters per input variable. Nevertheless, the web forms of the test do not handle all the possible inputs that may be introduced. As in any massive survey it is expected that the subjects answering commit mistakes. The most obvious and easy to detect errors are typos, extra information or mixed answers. Errors appear due to poorly structured questions or laziness from the users to read and answer carefully.

For the data analysis it is necessary to keep the data as clean and coherent as possible. Each URL contains 47 variables; there are 40 basic questions (as described in Appendix A) but some questions divide in sub-questions and each sub-question has to be saved as an independent variable, see Table 3.1. That is why the total number of variables is 47 and not the number of questions. All variables are quantitative, i.e. they represent a number. There are 18 continuous (nominal) variables and 29 discrete (categorical) variables. Discrete variables are represented by integers while continuous are represented by floats.

In the test it is possible to leave answers empty. If an answer is empty this usually means that the user did not know the answer or he/she did not want to provide it.

During the data clean-up four issues were handled;

1. All URL encoding were converted to normal ASCII characters, e.g. %30 to 0.

2. Categorical question in the test are limited in the web form by means of radio buttons. Each radio button is related to a unique integer value. However, it was noticed that because the variables were moved from web-form to web-form using the URL to carry them some *hacking* took place; some values were altered and set outside their valid range. Incorrect answers out of the valid range were set to an empty value.

3. For continuous variables non-numerical values are not valid answers. However, the web-forms do not prevent the users from introducing strings. Handling this problem was a big challenge, in many cases the given strings complemented the answer, e.g. 78k in the question "How much do you weigh?". Therefore, deleting these answers would have lead to a massive loss of information (on why it is not good to just eliminate problematic records, see [1] second chapter). To overpass this problem the script was programmed to look for ordinary strings and remove them. For example, the characters k(kilos), c(centimeters), t(hours), v(years)[3] were removed.

   Alcohol related sub questions (question No. 14) had plenty of unacceptable answers. The test asks for the number of bottles or glasses, being drunk per week; however, the questions and how to answer them is confusing. Some answers were given in litters, others in centiliters, others in bottles, etc. These problems were noticed because the strings users attached to their answers, e.g. 25p. The script is programmed to handle the majority of the cases and if not able it will ask *for help*. The script will *learn* the provided help and use it as a reference for similar problems.

4. Another important issue observed in the dataset is the *shifting* of variables. Questions 37 and 38 are not included in the URL if their answers are left empty, contrary to the rest of the questions that are included even if empty. The missing of such space holders shifts the order of the variables. Although not a serious issue it has to be considered and handled correctly in order not to save answers of one question in another question's section.

Informative variables BMI and Total Alcohol Consumption are calculated while parsing each valid transaction. The variable BMI is calculated as the weight of a person divided by his/her squared height, Equation 3.1. Total Alcohol Consumption variable is calculated as the sum of all the different alcohol consumption references, Beer, Long drink, Concentrated alcohol, Wine and Light wine. Although different beverages, it is possible to sum them together if they are registered in quantities containing similar amounts of pure alcohol. That is why the questions ask for doses in 1/3 of a liter, 4 centiliters and 12 centiliters. A similar approach was used in [12].

$$BMI = \frac{weight(kg)}{height^2(m^2)} \tag{3.1}$$

In total there are 49 variables being processed, 47 original from the test and 2 calculated ones. To these 49 variables 3 reference variables are added: time of transaction, file transaction identifier and the row number where the transaction was found. These reference variables though not of use in the DM/ML analysis will serve to backtrack errors to the original sources. The variables are all handled as numbers; this approach will provide an optimized database with a small footprint.

[3]The t and v are for the Finnish words Tuntia (hours) and Vuotta (years).

3.2.2 SQLite database

To perform the different phases of a DM/ML analysis it is necessary to have access to particular sub-sets of the original dataset and, with the retrieved data, compute the required analysis. One way to improve the access to the dataset is by organizing it in a structure that simplifies the burden of finding relevant points in it, i.e. be able to search for points quickly. Using a relational database that can be accessed by means of the Structured Query Language (SQL) is a convenient way. SQL is a super set implementation of what is known as relational algebra. The language is simple and intuitive, simplifying the access to sub-datasets. The strength of databases is that they provide fast execution for most of the queries; this is useful when the queries are not known in advance and are formulated as the DM/ML project progress [7].

To save and organize the *Elämä Pelissä* dataset the SQLite database engine is used [10]. This engine is based on extremely efficient R*Tree structures and the total engine size is less than 500 kilobytes; it also has a small memory footprint. Another benefit of using SQLite is the available plug-ins for Perl and R that permit interacting with the database directly from these programs. Three tables were created for the dataset:

vars It is the main table where all the collected variables are saved. It has 55 numerical fields plus the primary key. Each record has a key pointing to **files**.

info It is an informative table where original data, i.e. before processing, about Alcohol variables and Sleeping time is kept. This table can be considered as a log from the parsing script that processes the data. Each record has a key pointing to **vars**.

files This table contains information regarding the parsed files. Number of transaction processed and number of valid transactions.

The parsing process for the 16 input files lasted no more than 30 minutes. A total of 16,085,351 transactions were analyzed (15,555,584 for September and 529,767 for October) from those transaction only 457,097 were found as valid ones (445,278 for September and 11,819 for October). Hence, from the provided dataset less than 1 in 32 URLs was useful, only 3.29% of the recorded transactions. Moreover, a record was eliminated, setting the final number of transactions to 457,096. This record was reported by the logging mechanism as a "record with incomplete number of variables". An in depth analysis to row: 1,082,747 of file: ACCESS_LOG_LAHNA_SYYS showed a damaged record; the URL has *garbage* in many of the variables.

The parsing also generated 4,937 records for the **info** table, i.e. there were around 5,000 transactions (1% of the total) where the alcohol values or sleeping time information was detected as invalid and needed post processing by the parsing algorithm.

The process finished successfully with a completely populated database, in an easy to access format and with a size of just 65MB, a radical improvement from the original 24GB of data.

3.2.3 Outliers

Outliers are points that come outside the main body of the data. These points have the property of affecting the generation of data models. After parsing the data and having it loaded into the database the following step is to detect the outliers in it. Outliers have two possible sources; first, people made mistakes when answering the questions, e.g. giving a height of 17 meters instead of 1.7 meters. Second, people intentionally played with the system giving invalid values, e.g. in the dataset some age values are set to 299,830 years.

The categorical variables are limited to only accept a restricted number of answers, incorrect answers are set to empty as it was explained in the previous section. Therefore, no more processing is needed. However, for continuous variables it is necessary to calculate some basic statistics and limit their range accordingly. Initially the variables are limited to what could be called an "educated guess"; thereafter with a simple statistical analysis the variables are limited to a range of 3σ, i.e. 3 standard deviations from the mean. This approach considers 99.7% of the cases. This simple process is done to each continuous variables; including the calculated ones BMI and Total alcohol consumption.

The variables outside the statistical determined ranges are set to a **-2** value in the database[4], also the counter kept at column OUTLIERVALUE is increased. Otherwise the record as a whole is not altered. This technique was performed manually to each continuous variable in the dataset.

An example with variable Age is observable in Table 3.2. Initially the range of values was wide open, up to 300,000 years with an extremely large variance. The first filter, educated guess, was set to be < 500. This filter limited the age to go from 0 to 495 and normalized the original statistics values. However, the range of age is still not usable, 495 is still to much for a human being; therefore a new filter is established setting the maximum of years to three standard deviations above the mean. With this last filter the last column is obtained. Where the range of Age goes from 1 to 88, a realistic value for human beings. During the filtering process the mean had a small change, it went from 45 to 43 years. However, the variance changed drastically, from more than half a million to 217. The filtering did not affect either the Median (44 years) nor the Mode (50 years). A total of 801 Age variables were set as outliers (64 from the first filter, 737 from the second). A similar process was performed to all the continuous variables. Lets notice that not all variables went through a two-step filtering, a second filter was applied only if required, e.g. the range of values with only one filter was still too large. The complete results for all variables can be observed in Appendix B.

If a record has few variables as outliers, e.g. two or three, it may imply that some typos occurred while the user played the game. However, a record with multiple outliers may imply that somebody introduced the errors on purpose; hence, the full record shall be marked as an outlier, it should be discarded at once and its values not used. With the numbers registered at the column OUTLIERVALUE it is possible

[4]In the table **vars** two numerical codes are used to know why a value is not present or valid. A **-1** means it is a missing value. A **-2** means the value is an outlier.

Age			
	Initial		Final
Min	0	0	**1**
Max	299830	495	**88**
Mean	45.35	43.25	43.15
Variance	594993.9	226.37	217.24
Std Dev	771.36	15.05	14.74
Median	44	44	44
Filter	< 500	$\mu < 3\sigma$	
Outliers	64	737	**801**

Table 3.2: Variable Age; before and after applying the different filtering rules.

to know the number of outlier fields in each record. The table **vars** also includes a column called OUTLIER; this field is set to 1 when the record is an outlier and is kept as 0 when the record is valid.

Figure 3.1 depicts the number of outliers against the number of variables. It can be seen that mistakes are common, there are 74,437 records with 1 to 4 outliers, i.e. 16.28% of the answers in the test have a problem in up to 4 variables. In the figure is observable a *breaking* point at 5 variables. Therefore, five was decided as the number of outliers to separate valid records from invalid ones. Each record having 5 or more outliers will be discarded, i.e. set to 1 in the OUTLIER column. With this division 2,668, 0.58%, records were marked as outliers.

In brief; two kinds of outliers were pinpointed in the data: the variables alone and the complete records. When a variable is set as an outlier it is no longer usable; however, the record it belongs to is. On the other hand, when a record is tagged as an outlier none of its variables are usable, not even the few valid ones it may have.

The elements with values outside the valid ranges were labeled as outliers, elements with missing values were not changed. Table 3.3 presents the proportions of missing, outlier and valid elements for each column in the database.

From the table it is noticeable that the discrete variables do not have outliers, due to their limited set of answers. The only discrete variables outside this restriction are *House size adults* and *House size minors*. These questions are open questions, i.e. the domain of the answer is the positive integers, however some outliers were found and therefore the range was restricted. For any other case these two variables are treated as discrete during the analysis.

The variables with the most missing values are Cholesterol total, Cholesterol HDL, Blood pressure diastole and Blood pressure systole. The high missingness of these variables is expected because of the difficulty the questions impose; knowing the correct answers beforehand is improbable. Other variables with high missingness are House size minors and Number of cigarettes smoked per day. Although there is people with no children and people that does not smoke it is necessary to confirm

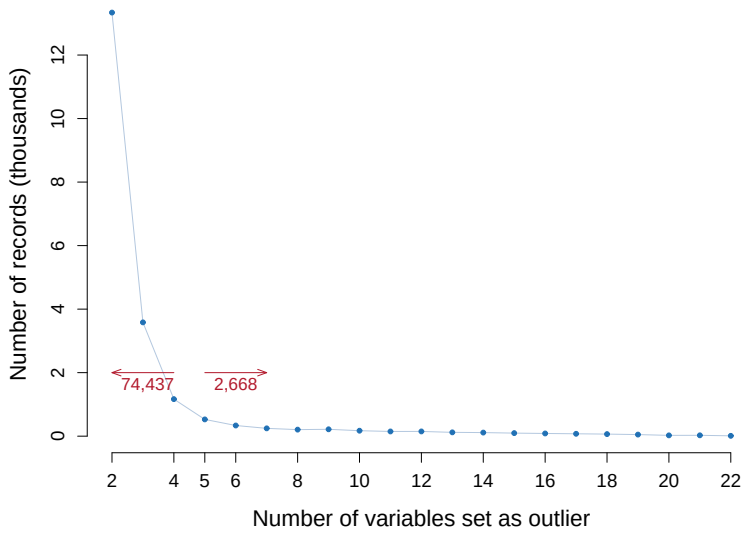

Figure 3.1: Plot showing the number of outliers against the number of records. The breaking point was set at 5. The arrows show the number of records in each set, less than 5 (Left) and 5 or more outliers (Right). The plot excluded the records with only 1 outlier due to their large number (56,351).

this by giving a zero value in the answer. An empty answer is not automatically translated to a zero-value answer. Therefore if the answers are not given the variables are treated as missing ones. On the contrary, for the Alcohol-related variables empty answers will be translated to a zero-value, meaning that if a variable is left blank its value will be set to zero automatically. This modification is done accordingly to the documentation of the test, see [14].

Regarding the outliers, the highest percentage are observable in the following questions: *Until what age will you live?* and *When is the best age to live?*. Although these were simple questions the high number of outliers found was not expected. More difficult questions, e.g. Blood pressure diastole, did not show as many outlier as these two variables.

Column	Type	Records		
		Valid	Missing	Outlier
Sex	D	453412(99.19%)	3684(0.81%)	–
Age	C	452584(99.01%)	3711(0.81%)	801(0.18%)
Height	C	449834(98.41%)	5779(1.26%)	1483(0.32%)
Weight	C	448692(98.16%)	4037(0.88%)	4367(0.96%)
Education	C	448173(98.05%)	5648(1.24%)	3275(0.72%)
Economic situation	D	452154(98.92%)	4942(1.08%)	–

Income	C	422703(92.48%)	27726(6.07%)	6667(1.46%)
House size adults	D	451762(98.83%)	4102(0.90%)	1232(0.27%)
House size minors	D	211093(46.18%)	244341(53.46%)	1662(0.36%)
Cholesterol total	C	243651(53.30%)	210739(46.10%)	2706(0.59%)
Cholesterol HDL	C	149737(32.76%)	304937(66.71%)	2422(0.53%)
Blood pressure systole	C	315379(69.00%)	135993(29.75%)	5724(1.25%)
Blood pressure diastole	C	311925(68.24%)	136960(29.96%)	8211(1.80%)
Diabetes status	D	448081(98.03%)	9015(1.97%)	–
Father infarct	D	449151(98.26%)	7945(1.74%)	–
Mother infarct	D	450130(98.48%)	6966(1.52%)	–
Alcohol beer	C	448281(98.07%)	0(0.00%)	8815(1.93%)
Alcohol drink	C	453227(99.15%)	0(0.00%)	3869(0.85%)
Alcohol concentrated	C	450223(98.50%)	0(0.00%)	6873(1.50%)
Alcohol wine	C	450368(98.53%)	0(0.00%)	6728(1.47%)
Alcohol light wine	C	449763(98.40%)	0(0.00%)	7333(1.60%)
Get drunk	D	404024(88.39%)	53072(11.61%)	–
Vegetables	D	451714(98.82%)	5382(1.18%)	–
Fruits and berries	D	451564(98.79%)	5532(1.21%)	–
Kind of butter	D	451699(98.82%)	5397(1.18%)	–
Kind of cooking oil	D	450683(98.60%)	6413(1.40%)	–
Kind of milk	D	448674(98.16%)	8422(1.84%)	–
Smoker for a year	D	451063(98.68%)	6033(1.32%)	–
Smoking now	D	448052(98.02%)	9044(1.98%)	–
Number of cigarettes	C	199507(43.65%)	253868(55.54%)	3721(0.81%)
Seat belt	D	451189(98.71%)	5907(1.29%)	–
Sleeping time	C	447291(97.85%)	4433(0.97%)	5372(1.18%)
Sports activities	D	451907(98.86%)	5189(1.14%)	–
Association activities	D	448925(98.21%)	8171(1.79%)	–
Movies and theater	D	451329(98.74%)	5767(1.26%)	–
Religious events	D	448256(98.07%)	8840(1.93%)	–
Reading and music	D	451765(98.83%)	5331(1.17%)	–
Hobbies	D	451118(98.69%)	5978(1.31%)	–
Issues with spouse	D	447882(97.98%)	9214(2.02%)	–
Issues with children	D	448002(98.01%)	9094(1.99%)	–
Stress	D	451709(98.82%)	5387(1.18%)	–
Life satisfaction	D	451930(98.87%)	5166(1.13%)	–
Work study load	D	448011(98.01%)	9085(1.99%)	–
Dreams impossible	D	450634(98.59%)	6462(1.41%)	–
Do not have good friends	D	450958(98.66%)	6138(1.34%)	–
What age will live	C	431032(94.30%)	15585(3.41%)	10479(2.29%)

Best time to live	C	406536(88.94%)	35884(7.85%)	14676(3.21%)
BMI	C	441894(96.67%)	6201(1.36%)	9001(1.97%)
Total alcohol consumption	C	448323(98.08%)	0(0.00%)	8773(1.92%)

Table 3.3: Table of variables with the proportion of records that are Valid, Missing or Outlier. The second column indicates if the variable is discrete (D) or continuous (C).

3.2.4 Grouping and Sampling

The database includes a total of 457,096 records. There are only 18,654 (4.08%) complete records, i.e. records where all the variables have a valid value. The complete records were grouped into identical ones to see if a *basic profile* could be identified. However, 18,489 independent groups were found, the largest one with 7 records; a really small number to call it a profile.

When using the original database 448,230 groups were found. The largest group has 1,835 records in it; nevertheless, instead of being an informative group, as the expected profile, the group resulted to be the one with the records where all the variables were left empty, i.e. no questions were answered. The next largest groups have only 62 and 61 records hence, they are really small. The difference with the first group is that in these ones the Sex value is given, the rest of the variables are still empty.

A database with complete valid records was generated. It has 18,489 records, its basic statistics were calculated and are observable in Appendix C. Because its size this database is easy and fast to handle and may be useful to speed up some analysis. However, a complete record may indicate two things; first, that it was answered by a person that knows him/herself really well and is willing to share the information or, second; that it was answered by a person that knows his/her numbers because he/she has some problem and has to be cheeking them constantly. The later will represent a problem for further analysis because the data will be biased.

Another approach is to create samples of the original database based on all the valid records. Nine databases were generated using random sampling selection from the original database. Only valid records, i.e. not outliers, were used; the pool to choose from had 454,428 elements. The databases were filled in sets of three; with 10,000 (2.2%), 50,000 (11.0%) and 200,000 (44.0%) records respectively. The basic statistics of each group were also calculated, see Appendix C. The observed values are, as expected, close to those of the original data. However as it will be explained in the next chapters the more records used the better results can be obtained.

Figure 3.2 illustrates a plot where different statistical information is compared between the databases: Original, Complete records and samples 10k, 50k, 200k. Two variables are used for comparison Education and Blood pressure systole. It is observable that all databases present similar basic statistics values. The largest

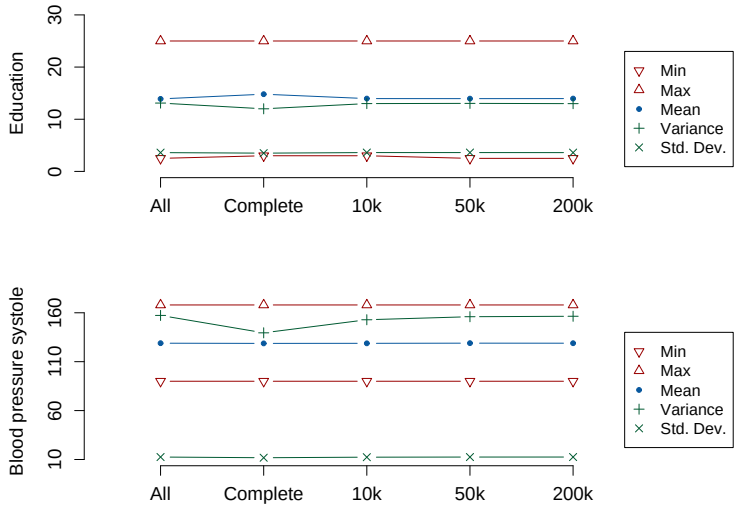

Figure 3.2: Plot where the distributions for variables Education and Blood pressure systole for the sampled, complete and original databases are presented. First column, ALL, shows the values for the original dataset. Column COMPLETE shows the corresponding values for the dataset with only full records. Remaining columns 10K 50K 200K show the statistics for the sampled datasets.

deviation is seen at the complete records dataset, where the variance is smaller. This difference may correspond to some of the issues stated in the previous paragraphs.

The generation of the smaller databases will be of use when analyzing the data and modeling the variables. Their size and completeness makes them easy to use and fast to analyze. The original database, however, has not been discarded. In the following section a data exploratory analysis is performed. The different variables are plotted in different ways and some interesting relations are found.

Chapter 4

Data exploration

When a new dataset is ready to be analyzed; one of the best, easiest and less costly ways of doing it is by performing a Data Exploratory Analysis (DEA). Data exploratory analysis represent the set of techniques used for displaying a collection of data so that the information within the data is easily available [19]. The data shall be displayed in a form that is simple, helpful and easy to interpret. Visual methods are important because the human brain's capability of analyzing images in a fast and accurate way. Correct visualization and careful analysis are ideal for finding unexpected relationships in the dataset [7]. The exploratory analysis will provide a way to check assumptions and reveal additional information that may have not been expected. Nevertheless, the limitations of the exploratory analysis are due to the high dimensionality of the data; how to display enough variables at once to make the visualization valid, reliable and useful is a challenge. The basic principles on how to display data for accurate and effective analysis were followed as recommend in [32].

The *Elämä Pelissä* dataset has 20 continuous and 29 discrete variables. Initially each variable is analyzed independently. Then discrete variables are analyzed against each other and continuous variables are done so too. Finally, discrete variables are analyzed against continuous variables.

4.1 Single plots

The *Mean* and *Median* are simple but interesting data summaries. Both are measures of location [7]. The specific values for each variable were calculated in the previous chapter for all continuous variables (see Appendix B and C). In the single plots these measurements will be easy to visualize. For each single variable a plot and a histogram is calculated. Histograms are simple tools useful for displaying the frequency distribution of a dataset. They could be used to observe data following different distribution, e.g. normal distribution. In a histogram the height of each rectangle is dependent on the number of observations that lay on the width of the same rectangle [19].

A total of 49 plots were generated, all graphics are available as a part of the

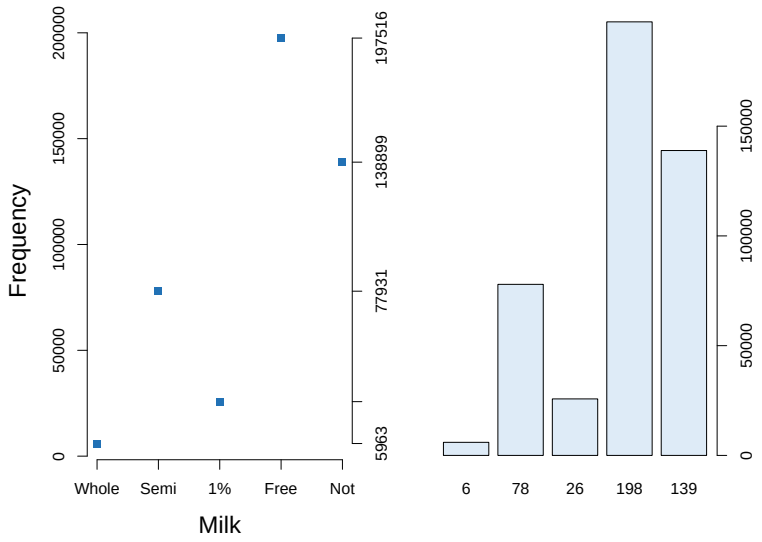

Figure 4.1: Single plot of discrete variable Milk. "Fat free milk" and "Do not drink milk" are the most common answers. With 197,516 (43%) and 138,899 (30%) respondents respectively.

research; however only some figures have been included in this thesis. For discrete variables Figure 4.1 illustrates the variable Milk. For continuous variables Figure 4.2 illustrates the BMI of the population. All 49 plots were visually analyzed, some interesting discoveries/observations are presented in Table D.1 in Appendix D. It is important to notice that these results come from a visual analysis therefore provided numerical values are not exact. Some plots returned valuable information, others presented obvious results and some few ones no information at all.

4.2 Discrete variables

The *Elämä Pelissä* dataset has 29 discrete variables. Each variable was plotted against each other in a matrix-like figure. The size of the matrix corresponds to the number of different option combinations between the two evaluated discrete variables. For example, there are five possible answers for Economic situation (Much better, Better, Equal, Worse, Much worse) and there are four answers for Life satisfaction (Very pleased, Pleased, Somehow, Dissatisfied). Thus when these variables are plotted together the resulting matrix will have 20 squares or options; each square will be labeled with the probability of the two answers shown together in the dataset.

A number and a color are associated to each square, the sum of all the numbers in the matrix is 1000. The smaller the number is the smaller the chance of having such case in the data, contrary to what a large number represents. The colors go from pale yellow to dark green, where the pale yellow is associated with low probabilities

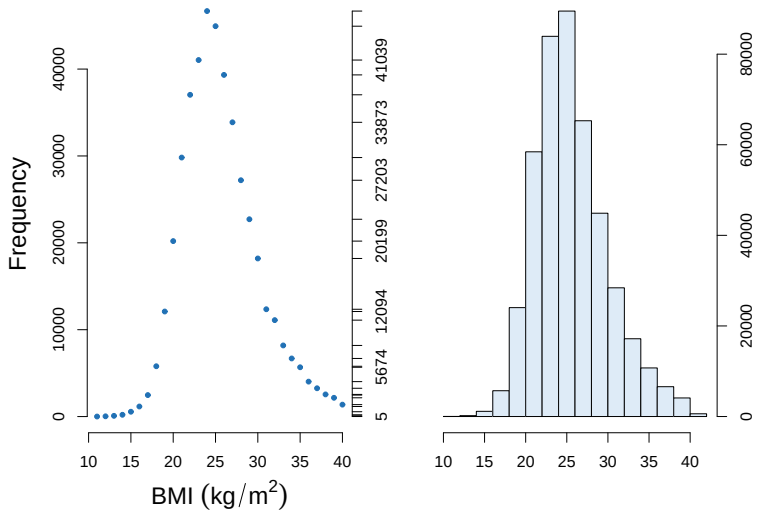

Figure 4.2: Single plot of continuous variable BMI. The plot resembles a normal distribution with a mean at 25.

and dark green with high probabilities. On the third and fourth axis of the plot there are two histograms showing the number of corresponding answers for each individual option.

Depending on the resulting plot some interpretation can be done. With Economic situation and Life satisfaction a possible linear relation is observed, i.e. a poor economic situation is related to a dissatisfied life. Figure 4.3 portrays the plot of Economic situation and Life satisfaction. Smoking and Stress are also correlated variables, Figure 4.4 depicts them.

With 29 discrete variables 406 combinations are possible; all combinations were plotted and saved as part of the research. Some interesting discoveries are described in Table D.2 in Appendix D. Not all plots showed important information but all cases were studied. The given percentages correspond to the pair of selected variables from the amount of respondents found in the total number of records available. No filters have been used, e.g. dividing the respondents by sex or age groups; meaning that the dataset was used as a whole.

4.3 Continuous variables

The *Elämä Pelissä* dataset has 18 continuous variables plus two calculated ones. Each one of the variables has been plotted against each other in a contour-like plot. The contour has different colors for each surface which depends on the number of elements it represents. Higher number of elements in the area are represented with a dark green color, while lower or none elements are drawn in pale gray.

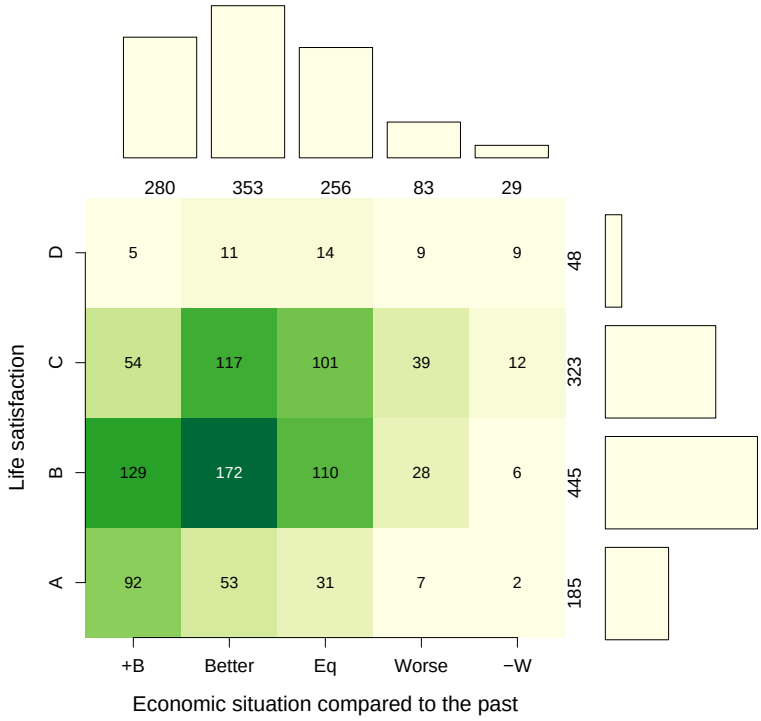

Figure 4.3: Mixed plot of discrete variables: Economic situation and Life satisfaction. The plot shows a possible linear relation between the two variables; very pleased people will have a much better economic situation contrary to dissatisfied people having a much worse economic situation.
Code: A very pleased; B pleased; C somewhat satisfied; D dissatisfied.

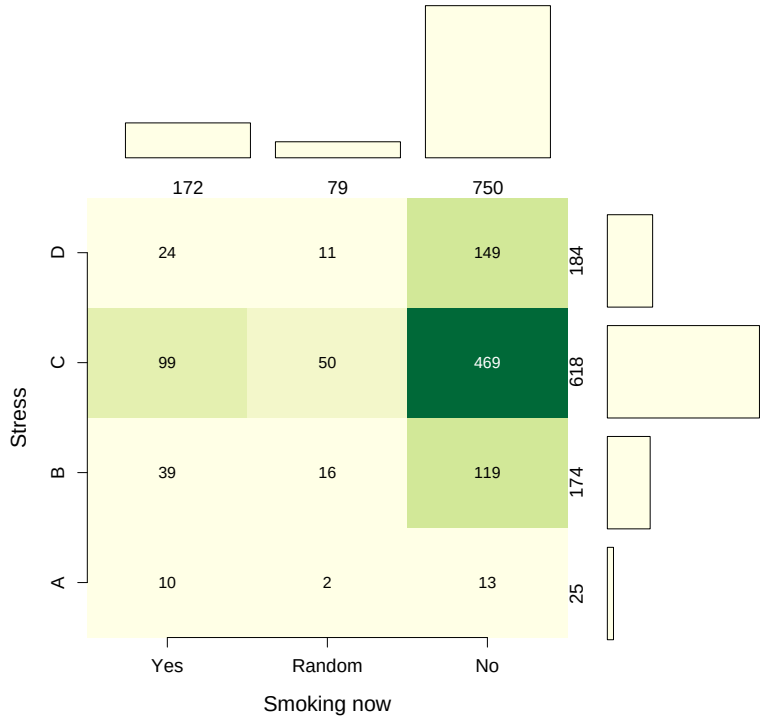

Figure 4.4: Mixed plot of discrete variables: Smoking now and Stress. The plot shows that smokers are somewhat or more stressed; around 15% of the records fail in these categories.
Code: A yes, life its hard; B more than people in general; C somewhat; D no.

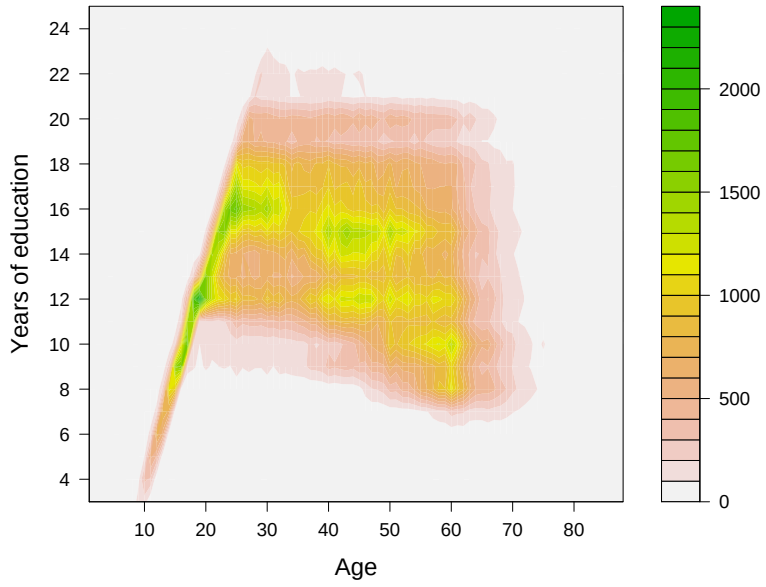

Figure 4.5: Contour plot of continuous variables Age and Education. The education level is limited by age, i.e. younger people do not have higher education levels. It is also observable older generations are less prepared, e.g. 40-50 years old people concentrate in 12 to 16 years of education while 50 to 60 years old people concentrate in 8 to 14 years of education.

The number of elements available for the plot depends on the selected variables, e.g. for Age and Education there are 445,281 (97.4%) elements. Figure 4.5 illustrates a person's years of education along with the age. Figure 4.6 depicts the consumption of wine depending on the years of education. For this later plot there are 440,817 (97.0%) elements. It is noticeable how educated people drink more wine, specially with 14 to 18 years of education.

With 20 continuous variables 190 different plots were generated. All plots were saved as part of the research. Some interesting discoveries are described in Table D.3 in Appendix D. Only few plots showed important information but all cases were studied. No filters were used during the generation of the plots.

4.4 Mixed variables

When working with discrete and continuous variables it is important to decide how to visualize pairs that come from both sources, i.e. how to have continuous and discrete variables on the same plot. For the analysis of the *Elämä Pelissä* dataset the discrete variables are set in the x-axis with their corresponding number of limited

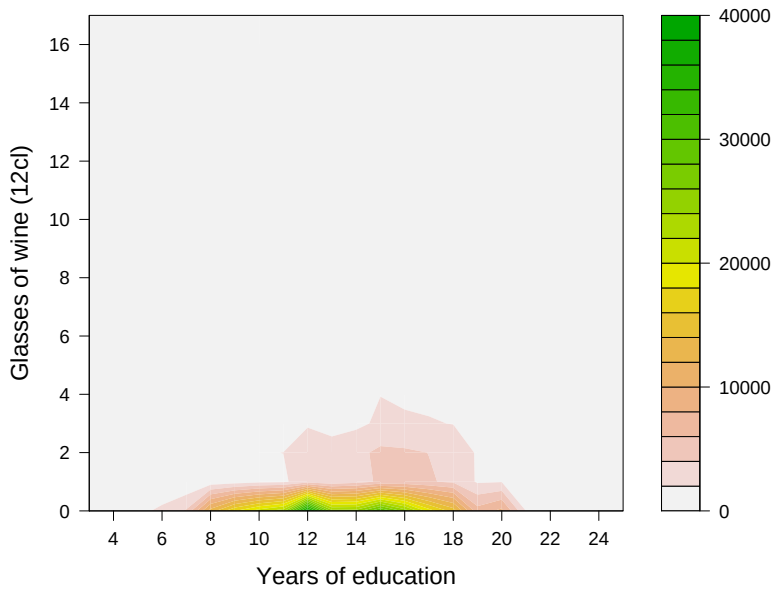

Figure 4.6: Contour plot of continuous variables Education and Wine consumption per day in glasses. The years of education go from 6 to 21. The number of glasses of wine have a peak at four when education is between 14 and 16 years. The majority of the population drink less than one glass of wine per day; however, people with an education ranging between 12 to 18 years are the ones drinking more wine.

options; the continuous variables are set in the y-axis with their full range of values. The query applied to the database selects all the valid elements corresponding to the pair of variables being processed. With the resulting dataset a matrix-like structure is generated where the columns are normalized. Each column represents the various options of the selected discrete variable. The normalization is done with the objective of better observing how the continuous variables are *distributed* over the discrete variables in each one of their different options. At the third axis the number of respondents per column is indicated. A color palette is added at the fourth axis, to facilitate the interpretation. The colors go from pale white to dark red; where pale white indicates low proportion of elements and dark red a high proportion. Let's not forget that in these plots the values are given in proportions and not in number of elements.

Figure 4.7 illustrates the plot of Economic situation and Education. The five different columns in the figure correspond to each one of the five different options to select from Economic situation. For all the columns dark red colors are observed between 12 and 16 years, specially 12 years. This means that independent of the economic situation a high proportion of people study for 12 years at least.

Figure 4.8 depicts the plot of Issues with children and Sleeping time; an interesting variation is observable when problematic children show up. Another example is Figure 4.9 where variables Type of milk and Age are portrayed.

With a total of 49 variables (29 discrete and 20 continuous) the number of different plots is 580. Some interesting discoveries are described in Table D.4 in Appendix D. Only few plots showed important information but all cases were studied. As in the previous cases no filters were used during the generation of the plots.

4.5 Specific consults

In the previous sections the *Elämä Pelissä* dataset has been analyzed using a visual exploratory approach; different kinds of plots were used depending on the type of variables. However, for all cases the full dataset has been used, i.e. all available and valid records were used without filtering. A filter is a mechanism used to generate a valid subset of the full valid dataset. A filter can be used to select records limited to the male or female gender, records that have w weight or y age, for example. With a filter correctly applied in a DEA it is possible to find interesting relations inside subsets of data. The filter, or filters, can be as simple or complex as required. Initially some basic but useful filters are applied, e.g. sex, age group, etc. More in-depth analysis may require a more advanced filter or sequence of filters.

To facilitate the filtering, plotting and analysis a Perl script was programmed. The script simplifies the generation of the SQL queries and it generates an R file used to plot the resulting figure (if stated the script will handle the plotting also). The script can be run in batch mode or using a cgi web interface. However, the web interface is not fully polished; the selection of variables and the setting of ranges have to be done manually making it error prone.

The following four queries have been generated with the help of the script. The

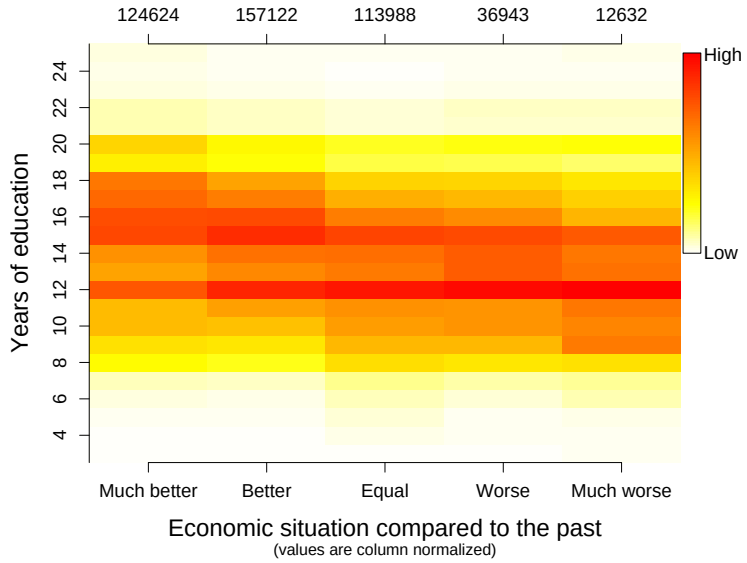

Figure 4.7: Plot of discrete and continuous variables: Economic situation and Years of education. A *step* is observable at the different columns showing a possible relationship between the two variables. The people with a "Much better" economic situation have an education level between 14-18 years; the people with a "Much worse" economic situation have an education level between 8 and 14 years. While not conclusive, it is noticeable that the education level directly affects the actual economic situation.

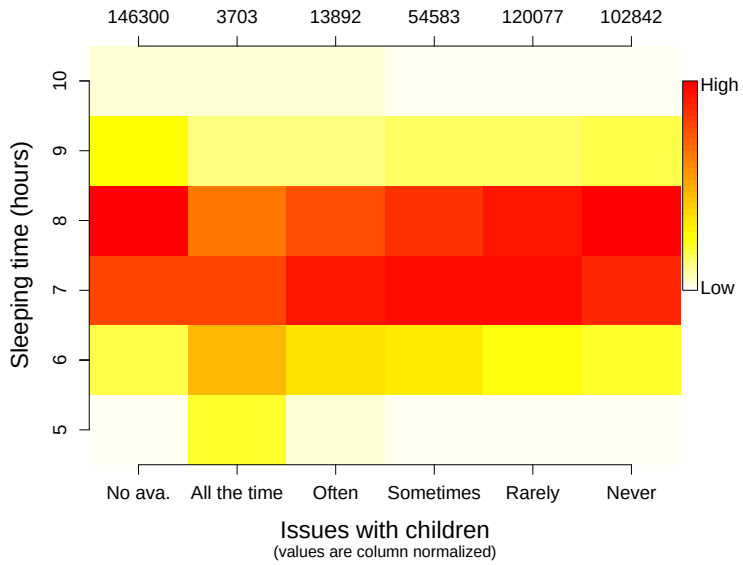

Figure 4.8: Plot of discrete and continuous variables: Issues with children and Sleeping time. The number sleeping hours for the population in general concentrate between 7 to 8. The variation of hours of sleeping is similar in all the options of variable Issues with children; however at option "Issues all the time" a difference is noticeable. For this case the sleeping time range goes down to 6 to 7 hours per day and, in some cases, to 5 hours. Therefore, its a fact that problematic children do affect the resting time of the parents.

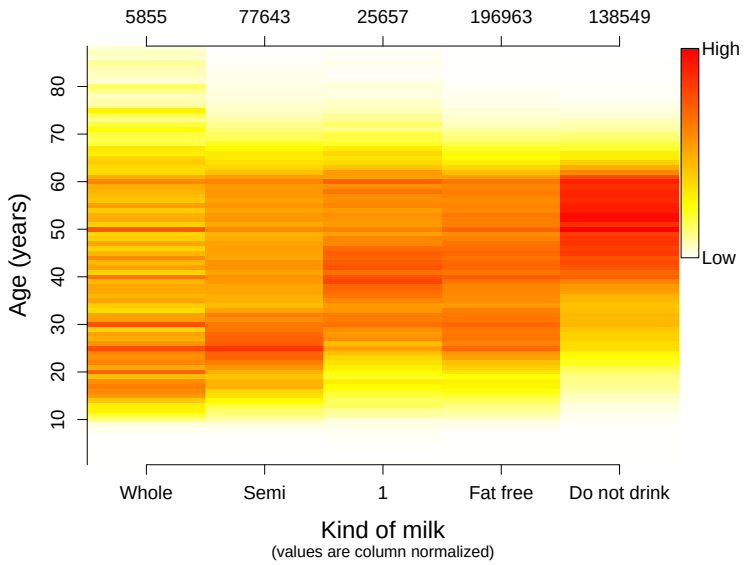

Figure 4.9: Plot of discrete and continuous variables: Kind of milk and Age. Two interesting patters can be observed in this plot. Firstly; it is observed that the age range of people not consuming milk goes from 40 to 60 years. Secondly; whole milk is consumed in all the range of ages, but it is specifically consumed by older people between 60 to 80 years. This can be explained with the fact that some time ago only the natural form of milk existed. The new milk categories are of recent times; therefore, old generations seem to be consuming what they grew up with.

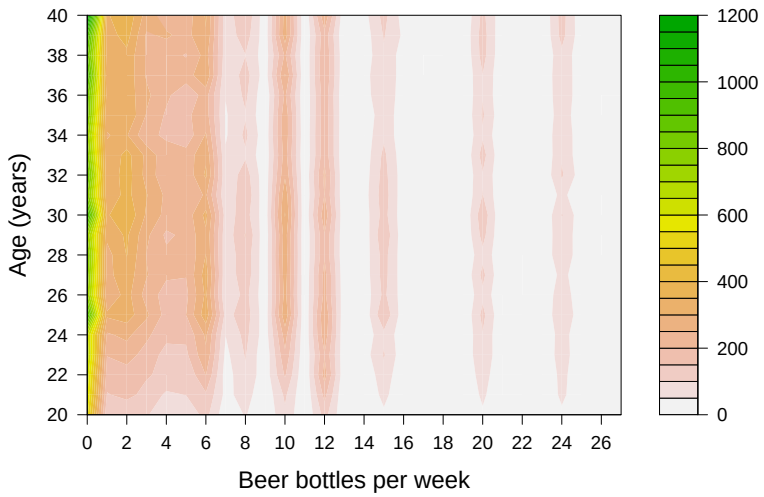

Figure 4.10: Beer consumption per week for males between 20 to 40 years old; 68,694 records (15.0%) found in this category. The gross of the population concentrates between 0 to 6 beers per week. The consumption is higher for people older than 25 years. There is a visible consumption, 200 to 300 people, at 10, 12 and 15 beers per week. Between 20 to 25 beers per week the visibility is less and a smaller population of around only 100 people exists. Therefore, in average, it is highly probable that a male between 20 to 40 years old drinks one beer a day.

combination of filters that could be applied to the *Elämä Pelissä* dataset is extremely large (if not infinite); therefore, the analysis was limited to four consults only. The questions shown below are the result of the literature review and my personal interest.

1. What is the beer consumption per week for males between 20-40 years? Figure 4.10.

2. What is the average income for people, male and females, with high stress levels and a worse economic situation than the past? Figure 4.11.

3. What are the different life satisfaction levels for females that do normal to hard sports and have one or more minors living at home? and for those not having minors? Figure 4.12.

4. What are the cholesterol total levels for males that sleep seven or more hours, do not practice sports and are smokers? and for those doing sports and not smoking? Figure 4.13.

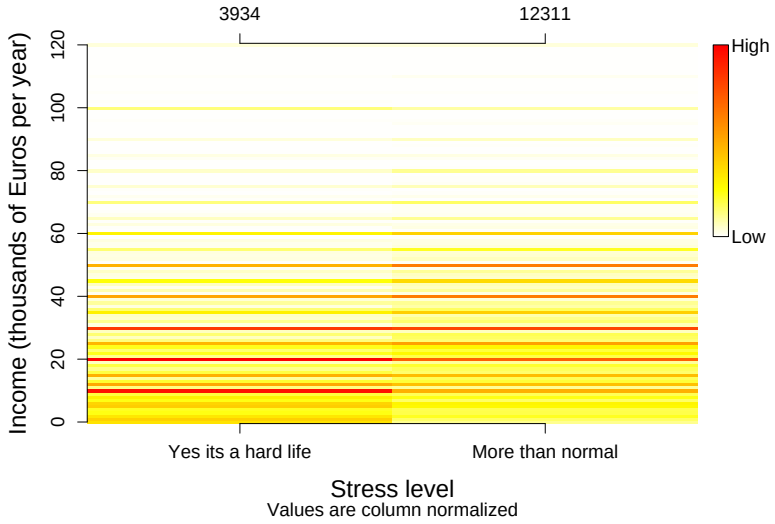

Figure 4.11: Income per year for people with high levels of stress and a "Worse" or "Much worse" economic situation that in the past; 16,456 records (3.6%) found in this category. The results show a population earning mainly between 0 to 60 thousands Euros per year. The highest levels of stress are seen at 30 thousand Euros. It is also noticeable that very stressed people, "Yes, its a hard life", have in average a lower income than people stressed "More than normal". However, the former has only 3,987 elements, compared to 12,469 elements of the later, a ratio of almost 1:3. As could have been expected, the economic situation plays an in important role in people's stress level.

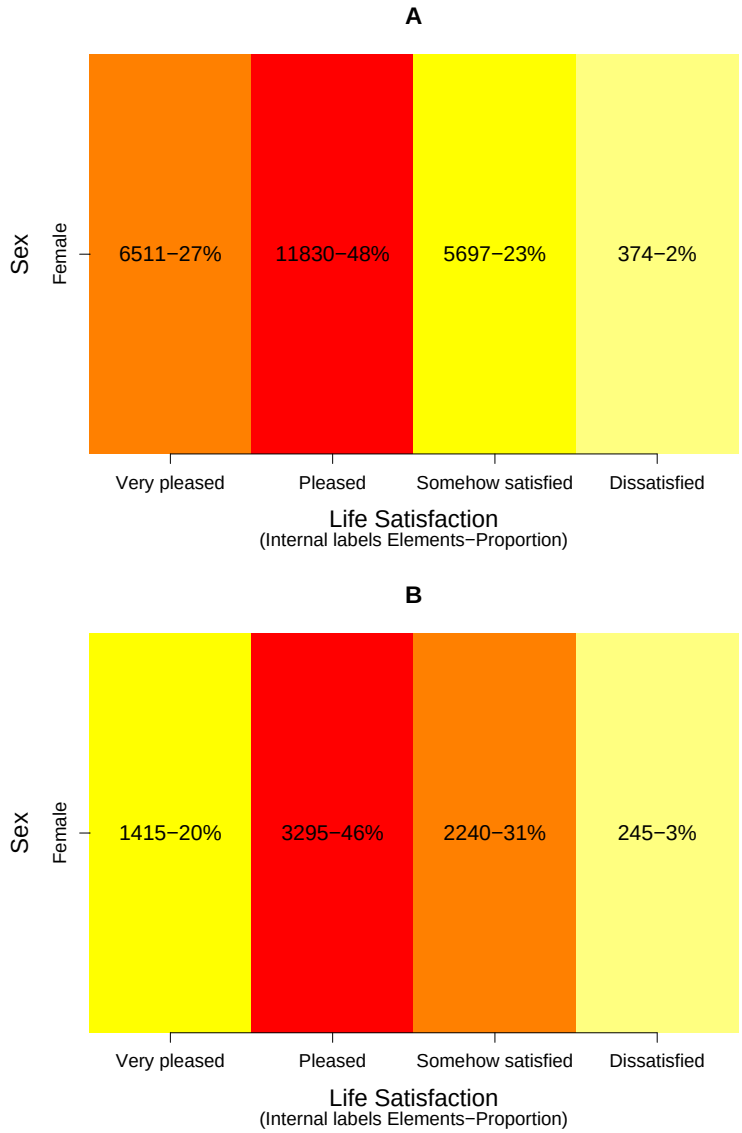

Figure 4.12: Life Satisfaction of females that do heavy and normal sports having one or more minor (A) or no minors (B) at home. In this plot it is noticeable how much children influence the life satisfaction of females. While the proportion of answers for "Pleased" and "Dissatisfied" do not change between plots, "Somehow satisfied" and "Very pleased" are switched. In plot A the really satisfied people represent 27% of the population and 23% represent the quite satisfied. When children are not present there is a reduction in the really satisfied people proportion to 20% and an increase in the quite satisfied people to 31%. Therefore, females with children are more satisfied with their lives, hence, children play an importance role in female life satisfaction.

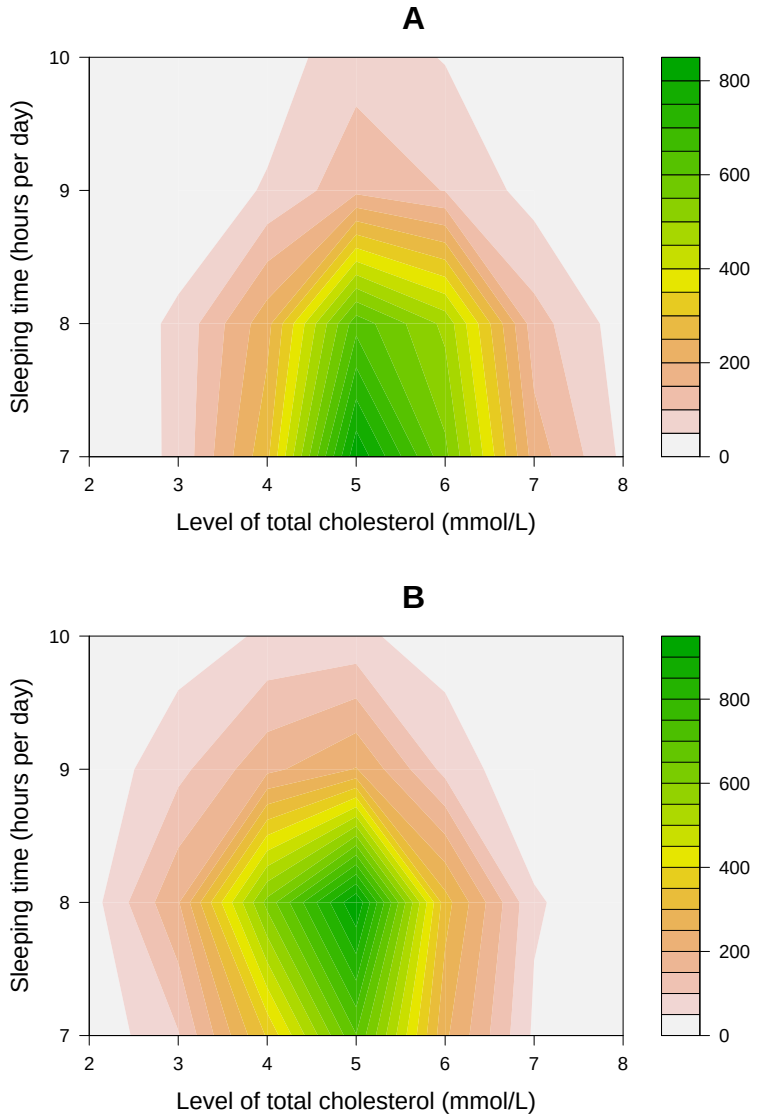

Figure 4.13: Cholesterol total levels for males sleeping seven or more hours that smoke and do not practice sports at all (A) and when not smoking and doing hard sports (B). Initially, it is observable that the mean level of cholesterol is at 5 mmol/L and that the average sleeping time variates from 7 (A) to 8 (B) hours. This variation may be due to the hard sports performed. Another observable issue is that the cholesterol contour levels are higher in plot A than in plot B. In the first one there are noticeable peaks (yellow-green) above 6 mmol/L while in the second one the levels go below 4 mmol/L. Therefore, there is an indication that a sedentary life complemented with smoking increases the levels of cholesterol. The desirable level of total cholesterol is below 5.2 mmol/L according to [20]. Cholesterol total is one of the variables that will be studied in the following chapter due to its high level of missingness.

The majority of the results obtained during the exploratory analysis could be considered new, interesting or reaffirmation of previous knowledge. In some cases the results are *exotic*, e.g. the infarct of father increasing the BMI of a person. Some questions that may arise after the analysis as: why, how and who, shall be studied by professionals from different fields as doctors, economist, sociologist, etc. Notwithstanding, this was the result of a data exploratory analysis; everything that was considered interesting has been pointed out throughout this thesis.

The exploratory data analysis showed the importance of an initial visualization stage of the data and its basic features. Different types of plots used with different pairs of variables showed important information within the data; without the need of an in-depth analysis. The filtering options and visualization tools facilitated the burden of preparing the queries and generating the plots, thus speeding up the exploratory process. Nevertheless, the analysis was delimited. The vast amount of available data and the different aspects of it would have required a full project focused to visualization and exploratory analysis; only to extract all the available information.

Chapter 5

Regression models

5.1 Introduction

The analyzed dataset contains 47 variables of which 29 are discrete and 18 are continuous. As explained before, the dataset is the result of a questionnaire answered by a large population. The application of a questionnaire to a large population has consequences; on the positive side the number of responses is very high, this gives a wide range of different answers and enough correct responses to minimize the effect of possible errors. On the negative side large datasets are difficult to handle and usually require pre-processing; computational expensive tasks turn tricky to apply.

The dataset's size also contributes to its porosity, i.e. there are unanswered questions. The variables with missing values, as well as the outliers were previously marked in the pre-processing of the data. The percentage of availability of the variables can be seen in Table 3.3. The missingness of each variable is highly related and dependent on the difficulty of the question related to it. If the respondent does not know the exact answer he/she may try to give the most used value, i.e. "the value everybody says is the correct one". For example, in the case of Blood pressure levels, people who had it checked a few days before answering the questionnaire may know the correct value, the same for people who have to check it constantly. However, the rest of the population may just try an average value or leave the field empty.

With enough information the missing variables could be recovered by means of other available variables, one approach is the use of regression methods. In a linear regression problem the objective is to estimate the response variable y by a linear combination of the input variables $\mathbf{x} = [x_1, \ldots, x_d]$ [31]. The regression model is given by

$$\mathbf{y} \;=\; \mathbf{X}\beta + \varepsilon\,, \tag{5.1}$$

where $\mathbf{y}$ is a $N \times 1$ vector of responses and $\mathbf{X}$ is a $N \times d$ matrix of input values. The regression coefficients are given by $\beta = [\beta_1, \ldots, \beta_d]^T$ and are unknown. The noisy

elements $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_d]^T$ are independent and normally distributed with zero mean and a given variance $\varepsilon_n \sim N(0, \sigma^2)$ [31].

The linear regression model given in [8] is

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \,, \tag{5.2}$$

the only major difference is in how the formula is represented. The term β_0 is the intercept coefficient of the regression model.

To measure the quality of the regression model the Residual Sum of Squares (RSS) [8] also known as Sum of Squared Errors (SSE) [31] could be used. However, during this thesis the Mean Squared Error (MSE) is used. The sole difference is that the error produced by the whole model is averaged by the number of elements used. The Root Mean Squared Error (RMSE) is the root of the MSE; by extracting the root the measurement of the error is in the exact same units as the variable being estimated.

$$RSS = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{5.3}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{5.4}$$

For the regression methods it is common to work only with continuous variables. Nevertheless, the idea of discarding all the discrete variables and only using the continuous ones did not seem correct. Specially, when more than half of the variables belong to this category; some useful information may be lying within them. It was decided to use the 7 most available discrete variables. They were *transformed* into a valid continuous domain to be used in the regression problem. These selected variables and their numerical transformation are shown in Table 5.1.

Various regression methodologies were applied for this thesis. First, a regression model was generated for each variable using its best predictor with a single variable. Then the previous approach was improved using a polynomial regression. The third and forth approaches included multiple input variables, in both cases different algorithms were used to select a relevant subset of predictor variables.

From previous analysis 5 variables were detected as the most missing ones, with a missingness between 30% to more than 50%. Four of these variables are related to topics which a given person has a high probability of not knowing the answer; the variables are Blood pressure for diastole and systole, Cholesterol total and HDL. The fifth variable is Number of cigarettes being smoked per day; although, it is an easy to know answer, the response is only available for 43% of the records. The missing answers could just mean zeroes, i.e. not smoking anything or people hiding their smoking habits, thus affecting the result of the test. In an effort to improve

Variable	Options (numerical value)
Sex	Male(-1) Female(+1)
Economic situation	Much better(+2) Better(+1) Equal(0) Worse(-1) Much Worse(-2)
House size adults	1(1) 2(2) ...
Smoker for a year	No(0) Yes(1)
Smoking now	No(0) Randomly(1) Yes(2)
Stress	Yes life is hard(3) More than average people(2) Somewhat(1) No(0)
Life satisfaction	Very pleased(+3) Pleased(+2) Somewhat(+1) Dissatisfied(-1)

Table 5.1: Discrete variable used for regression. The different options of each variable were transformed to an appropriate numerical value.

the regression of these variables a group of artificial neural networks were trained and a PCA regression approach applied.

It has been stated many times before the size of the dataset. However, the question of how many elements to use for regression is still open. As stated in [3] increasing the size of the data reduces the over-fitting problem, this is observed in Figure 5.1. The number of points used affect the variance of the error for training and testing; the more points are used the more stable is the model. Therefore, it was decided that for all the regression approaches all the available points were going to be used. In the case of k-fold cross-validation only 20,000 elements are used because the approach is always repeated, at least 100 times, to increase the reliability of the results.

For each method studied in this thesis all the continuous variables were modeled using the previously mentioned discrete variables, plus all the remaining continuous variables with an availability of at least 95% and both calculated variables, BMI and Total Alcohol Consumption. As a baseline to measure each one of the regression models the errors given by the Mean predictor were used. The Mean predictor is based on the mean, or most common answer, for each variable; an improvement over the mean predictor can be considered a gain, or learning, by the model. The Table E.1 in Appendix E has the mean predictor for all the variables. Each variable has a different measure unit. To facilitate comparison and detection of importance within variables they were normalized to have a zero mean and unit variance.

5.2 Single predictor

In a linear regression model it is assumed that the regression function is linear in the inputs, or that the generated model is a good approximation [8]. By convention, in any model, linear or not, the x-variables has the role of explanatory variable while

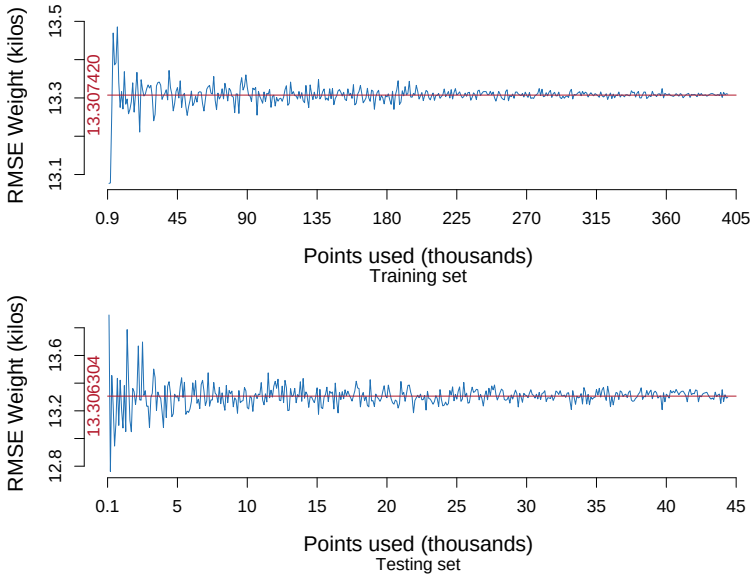

Figure 5.1: Variance of error in the Training and Testing phases depending on the number of samples used.

the y-variable is the response or result [19]. A linear regression model has the form

$$y = \alpha + \beta x + \varepsilon\,, \tag{5.5}$$

where the noise ε is independently and identically distributed with mean zero and variance σ^2. The coefficient x may come from different sources as its explained in [8]. It can be a quantitative input, a transformation of the quantitative input, an expansion of the input (e.g. $X_2 = X_1^2$), etc. The estimation of the variable $\hat{y}$ is done as

$$\hat{y} = a + bx\,, \tag{5.6}$$

where a is the intercept value, b is the slope and x the explanatory variable.

For this thesis two linear approaches were followed. In the first approach the best linear fitting variables are found, i.e. input variables are used, one by one, to find the explanatory variable x that best fits the response variable y.

This first approach was performed as follows; for each pair of different variables, where one is the explanatory and the other one is the response, the complete valid dataset was retrieved. With 90% of the retrieved dataset the linear model was generated, the remaining 10% was used for testing the model. The variable that presented the smallest testing MSE was the selected explanatory variable of the actual variable.

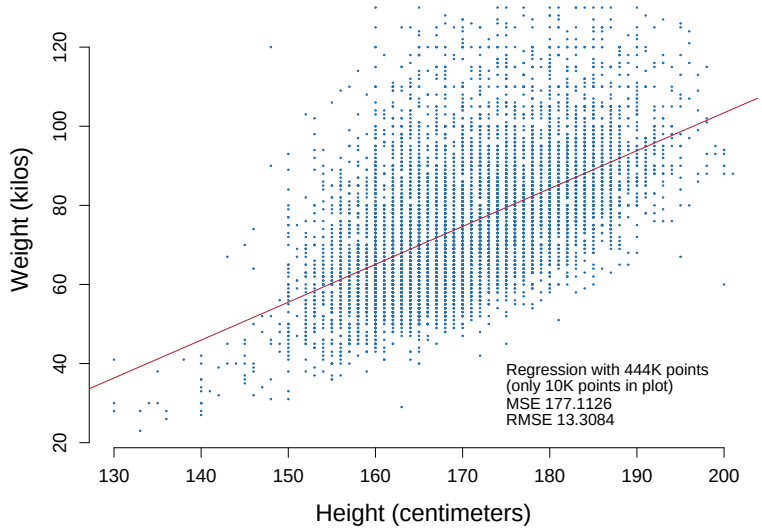

Figure 5.2: Plot showing the linear model of variable Weight when using Height as explanatory variable.

With the selected variable a 10-fold cross-validation was repeated 100 times and used to estimate the coefficients α and β as in Equation 5.6. The regression process returned different results with interesting information. The results and error reduction compared against the mean predictor are available in Table E.2 in Appendix E. As an example is the Height of a person; by using Sex as a explanatory variable the MSE was reduced by 50% compared to the mean predictor, see Equation 5.7. For Weight a reduction of 30% is achieved when using Height as the explanatory variable, see Equation 5.8. Figure 5.2 shows the regression of variable Weight by means of Height.

$$H = 171.75 - 6.76\,S \tag{5.7}$$

$$W = -88.12 + 0.96\,H \tag{5.8}$$

Not all models improved over the mean predictor, e.g. Alcohol Light Wine; in this case the error increased by 1%. In other cases the gain was too small, for example Cholesterol Total and HDL had a gain below 5%.

In the second approach the models were set to have two fixed variables Sex and Age plus a third one selected as the best explanatory variable. Moreover, the model was trained with different polynomial degrees of the explanatory variable; the degrees went from 1 to 5. All the different polynomials degrees were trained with all variables and the best polynomial with the best explanatory variable was selected. From there followed a 10-fold cross-validation repeated 100 times to estimate the coefficients α

and $[\beta_1 \dots \beta_d]$ where d is the degree of the selected polynomial.

Sex and Age are selected as basic variables and they both have a linear influence on the regression. The selection of these specific variables is because they can be seen as discriminants for the data, i.e. a way to cluster the data in a "natural" form (separation by sex and age groups). It is expected that, for example, height, weight, habits, stress levels, etc. differ between males and females. The same happens with Age, factors vary over time and affect different, e.g. 60 year old people have different habits than 30 year olds'. In many of the studies presented in Section 2.3 this separation is noticed and used; therefore, this same regression approach is also used.

Sex and Age played an important role for the regression. For example, Age contributed with more than 20% of the regression coefficient for Education, Income, Cholesterol total, Sleeping time and Best time to live; for this last one having more than 80% contribution. Sex influenced the Alcohol concentrated and Light wine, with a participation of more than 10%; the same for Education and Blood pressure systole but not diastole. The use of a polynomial regression model also contributed to an improvement compared to the previous single variable approach. There were improvements on the prediction of almost all variables; Height went from 51% to 63%, Education from 0.02% to 0.05%, Number of cigarettes from 66% to 71%. The difficult variables: Blood pressure systole and diastole from 0.09% to 0.17% and 0.09% to 0.13% respectively. The complete results can be seen in Tables E.3 and E.4 in Appendix E.

Figure 5.3 shows the regression of variable Weight using Height, the algorithm selected a polynomial of degree 5 as the best predictor. Height prediction improved by 5 points, from 31% to 36%, using this approach. The regression formula can be seen at Equation 5.9 where S is Sex, A is Age and H is Height.

$$\begin{aligned} W &= 35301.200 - 1.239\,S + 0.213\,A - 1105.050\,H^1 + 13.734\,H^2 \\ &\quad -0.085\,H^3 + 0.0003\,H^4 - 3.2E^{-7}\,H^5 \end{aligned} \tag{5.9}$$

5.3 Feature selection

When there are plenty of variables available for regression it may seem wise to use all of them at once to predict the missing variable y. However, when solving the linear models it is normal that all the input factors turn out to have coefficients different from zero regardless of its real importance [31]. Hence the problem of selecting a relevant subset of features upon which to focus attention. Machine learning algorithms including decision trees are known to degrade in performance, i.e. accuracy; when given as input features that are not necessary for predicting the required output [13]. A feature X can be considered strongly relevant if the removal of X alone will result in a deterioration of the classifier [13]. An approach to filter different features independently can be observed in Figure 5.4. All the set

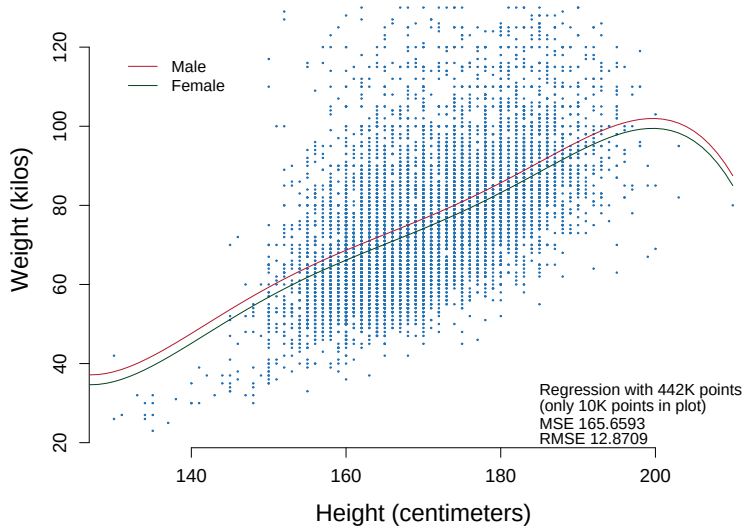

Figure 5.3: Plot showing the model of variable Weight when using Height in a polynomial form as explanatory variable. Sex was separated to appreciate the different curves for males and females. Variable Age, was kept fixed to its mean value.

Figure 5.4: Feature filter approach according to [13].

of features are given as initial inputs then a subset is selected based on a selection algorithm (Backward Selection, Forward Selection, SISAL, etc.) this selected subset is later used as an input baseline for any Induction Algorithm, e.g. Linear regression, Neural networks, etc.

For feature selection the Stepwise selection methods are well known strategies used to obtain candidate subsets from a set of input variables [31]. For this thesis two methodologies are investigated, Forward Stepwise Selection and SISAL [30]. Both methodologies were used as a way to select the best subset of features that provide the best regression in a parsimonious model.

5.3.1 Forward stepwise selection

The Forward Stepwise Selection algorithm starts with an empty set of inputs and adds the input that most decreases the RSS (or most improves the fit) in each subsequent step. The improvement in fit is often based on the F-statistics, Equation 5.10, where $\hat{\beta}$ is the model with k features, $\tilde{\beta}$ is the model with $k+1$ features and N is

the number of elements in the dataset. The algorithm sequentially adds the predictor producing the largest value of F; the forward selection algorithm is stopped when the F ratio is not greater than the 95th percentile of the $F_{(1,N-k-2)}$ distribution [8, 31].

$$F = \frac{RSS(\hat{\beta}) - RSS(\tilde{\beta})}{RSS(\tilde{\beta})/(N-k-2)} \tag{5.10}$$

Each variable ran the forward selection algorithm in order to select the best set of features that models the variable being analyzed. When the best features were selected a 10-fold cross-validation repeated 100 times was run to calculate the coefficients of the selected features.

The use of forward selection resulted in improvements against the mean predictor at almost all variables but Alcohol Drink. The variables with the largest improvements were Number of cigarettes, Height, Weight, Best time to live and Income with a 67%, 61%, 35%, 28% and 18% improvement respectively. The most important factors and their weights for these variables were Smoking now (0.83), Sex (0.51), Height (0.64), Age (0.94) and House size adults (0.25). The number of selected factors per variable was 5/15, 6/14, 4/14, 2/15 and 7/15; neither Height nor Weight can be used for modeling themselves. Let's notice that variable Number of cigarettes is one of the five most missing variables.

The forward selection algorithm selected less than 50% of the number of features available for each variable; moreover, the few selected variables contributed to a significant improvement in the prediction.

For the four remaining most missing variables the gains were as follows: Cholesterol Total 4%; Cholesterol HDL 3%; Blood pressure systole 15%; Blood pressure diastole 12%. The number of features selected was 4, 1, 3, 3 of 14 possible. One interesting thing to notice is that the algorithm selected Age and Weight as the most important variables for Blood pressure diastole and systole. For Cholesterol total the variable with the most influence was Age. The complete set of results for the forward selection algorithm can be seen in Tables E.5 and E.6 in Appendix E.

5.3.2 SISAL

Sequential Input Selection Algorithm (SISAL) is a simple and efficient backward elimination algorithm proposed by [30]. In SISAL the removal of features is based on the median and an empirically estimated width of parameter distributions sampled with a cross-validation re-sampling process [30]. Initially all features are used in the algorithm. The regression coefficient β_i of each feature is sampled by a k-fold cross-validation repeated M times, i.e. a total of Mk coefficients are generated for each feature i. The median m_{β_i} is calculated for each feature using the Mk estimates. The width of the sampled distribution is calculated by

$$\Delta_{\beta_i} = \hat{\beta}_i^{high} - \hat{\beta}_i^{low}, \tag{5.11}$$

where $\hat{\beta}_i^{high}$ and $\hat{\beta}_i^{low}$ are the $Mk(1-q)$ th and Mkq th values in the ordered list of the Mk estimates of $\hat{\beta}_i$. The variable q is a constant with a predefined value of $q = 0.165$. Using the median and the width the least significant feature is removed from the input set. This is done by using the ratio $|m_{\beta_i}|/\Delta_{\beta_i}$; the feature with the smallest ratio is dropped from the set of inputs [30]. This sequence, of cross-validation and pruning, is repeated until there are no more features to remove. The output is a number of models equal as the number of features available, due to the fact that only one feature is removed at a time. The model with the minimum validation error is selected and returned as the best model.

Each variable ran the SISAL algorithm to select the best model that represents it. A 10-fold cross-validation repeated 100 times was used to select the best features and models. The best results were obtained for variables Number of cigarettes, Height, Weight, Best time to live and Income, with respective improvements of 69% 61% 37% 29% and 22% against the mean predictor. As can be seen, the same best variables were observed while using the Forward stepwise selection algorithm; therefore, these variables may be easily predicted with a linear model. The most important factors for these variables were Smoking now (0.68), Sex (0.49), Height (0.51), Age (0.66) and House size adults (0.19). Similar to those selected by forward selection; however, the weight of the variables variates because SISAL selected, in general, more features for each model. The number of selected features per variable was 12/15, 9/14, 11/14, 12/15 and 13/15; neither Height nor Weight can be used for modeling themselves. As in forward selection variable Number of cigarettes shows up in the list of the best modeled variables.

The relatively high number of selected features provided a better regression compared to the single variable regression and a slightly improvement over forward selection. However, the use of more than 80% of the available features per variable may not be considered a simple and parsimonious model.

For the most missing variables the improvement was as follows: Cholesterol Total 7%; Cholesterol HDL 4%; Blood pressure systole 19%; Blood pressure diastole 15%. The number of selected features were 9, 7, 4, 11 of 14 respectively. A noticeable improvement in the regression can be observed, specially for Cholesterol HDL with an improvement of 4%. For all these four variables Age has a high influence on the prediction, specially in Blood pressure diastole and systole. Another important factor selected by SISAL and with a relative high influence in the regression is Total Alcohol Consumption, a calculated variable; it specially plays an important role for Cholesterol HDL. The complete set of results for the SISAL algorithm can be seen in Tables E.7 and E.8 in Appendix E.

5.4 Regression improvement

The results of the previous methods show that some variables are easily defined by a linear model while others do not show an improvement over the mean predictor. If the gain is minute and/or it requires plenty of calculations to obtain it, it may turn infeasible to use the model, specially if the mean predictor is so simple. From the five most missing variables only one presented an improvement of more than 50%, this was Number of cigarettes. For both Blood pressures the improvement was above 15% and for both Cholesterol levels it was below 10%. For the later the gain levels by means of forward selection and SISAL can be considered small.

It is well known that the linear models are easy to interpret and fast to calculate; unfortunately they are not accurate enough for some problems [30]. The dependencies between variables may not be linear, therefore a non-linear model would explain them better. It may also be that the features may work better if they are transformed before the linear modeling is done. Two different methodologies are investigated; firstly, a feed-forward neural network is used for non-linear models. Secondly, regression on Principal Components (PCA) is analyzed.

The five most missing variables are used in this section; the available features are determined by the results of the previously executed forward selection and SISAL algorithms, i.e. the variables selected by these methods are used as input for the neural network and PCA approaches in the first run of experiments.

The importance of a variable in a model is given by how substantial is its coefficient in the regression model. Because all variables are mean centered and have unit variance it is possible to compare them with each other. Neural networks and PCA are run, in a second set of experiments, using only enough features to represent at least 80% of the weight available. These TOP features are also selected from the results of the forward selection and SISAL algorithms.

5.4.1 Artificial neural networks

The artificial neural networks (ANN) are powerful processing systems originally inspired by biological neurons [3]. The ANN are used because of its highly adaptive nonlinear models and high capability of adaptation based on the data. They have been used to solve many different problems, e.g. financial and biochemical systems [31]. Feed-forward ANN provide an adaptable way to generalize linear regression functions, therefore the purpose of using them in this thesis. Generally a neural network consist of an input layer, a hidden layer and an output layer which are fully connected, i.e. each node in any layer is connected to all the nodes in the previous layer. For how a ANN works, its variations and applications refer to [3, 8, 33].

The configurable neural network model available in R will be used for this thesis. The function is based on the one presented in [33]; the feed-forward neural network is of generic form with a single hidden layer. For the purpose of modeling the given variables, the ANN are set to have a linear output instead of a logical one and the number of hidden neurons varies according to the number of input features. The number of neurons in the hidden layer is changed from a minimum of half the

number of input features to twice plus one the number of input features, e.g. if there are 5 input features the number of hidden neurons will vary from 3 to 11. With the network configured a 10-fold cross-validation process is repeated 100 times to increase the reliability of the results. From the resulting networks the one with the smallest error is selected.

In Table E.9, Appendix E, the results for the ANN when using the forward selection features ALL and TOP can be observed. Compared to the results obtained by only using forward selection a small positive improvement in the prediction is observed for both cases. Therefore, in the case of TOP, the model can be constricted to only the most important variables without affecting the results when using the ANN approach. Besides these facts the most difficult variable, Cholesterol HDL, did not present any improvement in the prediction.

In Table E.10, Appendix E, the results for the ANN when using SISAL are presented. No improvement is visible in the predictions compared to the use of SISAL alone. The only variable that presented a considerable gain is Number of cigarettes. This variable went from a 69% gain to 75% in the ALL variables test and to 73% in the TOP test.

The use of Neural Networks proved beneficial when using the features given by forward selection; nevertheless the gain was minute compared to the burden of training, testing and analyzing the ANN models. Neural Networks have proved to be beneficial in an extensive number of different areas, however they are not the best approach in all cases.

5.4.2 Regression on principal components

Methods for dimensionality reduction can be applied in the context of regression and data analysis [19]. It may be useful to transform the set of explanatory variables in a way that better suits the regression approach. Principal Component Analysis (PCA) is an orthogonal linear transformation that positions the data in a new coordinate system. Each component in PCA represents a share of the variance; the first component has the largest sample variance; the second component has the second largest variance and so on [8]. The PCA transformation is based on the decomposition of the data in its eigenvectors and eigenvalues, where each resulting vectors is orthogonal to each other. Further explanations and applications can be found in [2, 7, 8].

PCA is applied to the features selected by forward selection and SISAL. The data is used “raw”, i.e. there is no previous normalization, because the provided function performs this step. After calculating the PCA, enough components are selected to represent at least 95% of the variance. In some cases only one variable was available, nevertheless the method was applied although no effect could be expected; this was done because the test ran systematically.

In Table E.11, Appendix E, the results for regression on principal components for the features selected by forward selection are observable. Compared to the results when only using forward selection no improvement is observed nor an important decrease noticed either. The only variable with a noticeable change is Number of

cigarettes; when using ALL the features it goes down to an improvement of only 12%, on the other hand when using only the TOP features the gain goes up to 66%, matching that of the original prediction by forward selection.

In Table E.12, Appendix E, one can observe the results when using the SISAL features with PCA. Compared to the use of SISAL alone a reduction on the gain by the predictor is noticeable; the most affected variables are Cholesterol HDL, Blood pressure systole and Number of cigarettes. The tendency of no gain is also observed when using the TOP variables. Hence, the regression on principal components did not improve the models while using the SISAL features as input.

PCA, although a well known methodology, did not show any gain on the quality of the models. A future approach may include an in-depth study of the selected variables and their relations, as well as the visualization of the different components to observe, e.g. clusters.

When modeling the most missing variables, Cholesterol levels, Pressure levels and Number of cigarettes it was noticed that nonlinear models, as feed-forward neural networks, did not improve the quality of the prediction over other methods. In fact, the results were similar to those of simpler linear models and, in some cases, the results were even worse. Similar results were observed while using regression on principal components, no improvements were seen at all. Moreover, some variables have a noticeable reduction in the regression quality compared to the simpler models, e.g. Number of cigarettes. The results observed on Number of cigarettes provide a strong confidence of the linear model representation, as it could be calculated with an improvement of more than 60%, against the mean predictor, with a linear model with as few as 3 explanatory variables.

Chapter 6

Summary and conclusions

This thesis consisted on three parts. In the first part the dataset, *Elämä Pelissä*, was presented and studied. The data was cleaned, pre-processed and inserted into a custom-made database structure. All this with the purpose of facilitating the exploratory analysis and prediction of health related variables based on individuals' profile values. During this phase the basic statistics of each variable were calculated and outlier elements were removed. The result of this phase facilitated an Online Analytical Processing (OLAP), i.e. an approach to answer analytical queries in a fast and reliable way.

For the second part a data exploratory analysis was performed. It was based mainly on the basic statistics of the data and the visualization of the information in different plots. The information extracted during the exploratory analysis present an interesting baseline for further analysis. All the variables were analyzed with different approaches. An advanced script able to work with specific queries and plots was programmed with the sole purpose of finding interesting features that may not be easily visible.

In the third part of the thesis a data modeling approach was performed. Various methodologies were used for regression. The dataset has various elements as empty fields, i.e. missing information. The objective behind the regression methods were to create a model robust enough to predict these missing variables by using the available ones. Different configuration of linear models were tested; to complement them Forward stepwise selection and SISAL algorithms were used. The information obtained by these methodologies was fed into different multilayer perceptron networks; another test consisted on doing regression by means of principal component discriminant scores. The employment of fixed variables Sex and Age did improve the regression for single variable predictor models. Sex and Age showed to be relevant discriminant variables that shall be kept in mind for further analysis.

When using the variable selection algorithms, Forward stepwise and SISAL, interesting things were found. Both methodologies aim to reduce the number of necessary variables as much as possible while improving the quality of the model, i.e they follow a parsimonious approach. SISAL demonstrated to be better than Forward stepwise selection, the improvement difference is noticeable for some of the variables. How-

ever, SISAL was more dense in the number of variables selected, therefore providing a more complex model. The algorithms' different selection mechanism may be the cause of this difference. Forward stepwise increases the number of variables as a gain in the fit of the data while SISAL, on the other hand, reduces the number of variables based on the least significant variable.

In general, the work done provided large quantities of information and hints that may require further and deeper analysis. As consequence the work provided a structured baseline in which future work could be done over the dataset. The dataset provided detailed information that may not be available in smaller-scale questionnaires. The findings, related to the Finnish way of living and its population, may be of interest to the involved entities Duodecim and THL; they may provide further hints to focus the research. This thesis presented the initial round of findings and could be easily extended into more in-depth areas as required. On the technical side this thesis presented a complete data mining approach to a real life dataset, including detailed information on how to work and what to do with the data at each phase.

6.1 Future research

The research done for this thesis had a limited scope. In the regression analysis only continuous variables were used, for example. The analysis of discrete variables could be part of future research. Different approaches could be used for this purpose, e.g. Classification and Regression Trees (CART) for regression purpose [8]. Continuous variables could be integrated in the regression if correctly discretized. Within the available discretization methods the one investigated was Histogram Analysis, in which the histogram distribution of the corresponding variable is used as a way to discretize it [6].

A large dataset, as the one used for this thesis, could return interesting information if it is clustered correctly. Because of the data content there are some variables that could be used for "natural" clustering, i.e. dividing the dataset in different subsets using only the different categories of the selected variables. Sex and Age are important variables to keep in mind for this case; they played an important role when fixed for the regression models for single variable predictor as well as the polynomial approach. More advanced clustering methods could be useful to find different groups of people with similar profiles characteristics. The use of k-means and x-means could be an efficient way to find these clusters [26]. Mixture models and Bayesian networks could also be used as methodologies for the analysis of the data.

Bibliography

[1] Paul Allison. *Missing Data*. Quantitative Applications in the Social Sciences. Sage University Publications, 1st edition, 2001.

[2] Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 1st edition, 2004.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st edition, 2006.

[4] The finnish medical society duodecim. Internet: `http://www.duodecim.fi`, 2009. Accessed July 17th, 2009.

[5] Dean P. Foster, Choong Tze Chua, and Lyle H. Ungar. How long will you live? `http://gosset.wharton.upenn.edu/mortality`.

[6] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 1st edition, 2000.

[7] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 1st edition, 2001.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning — Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer, 2001.

[9] H. Heikkinen, P. Jallinoja, S. Saarni, and K. Patja. The impact of smoking on health-related and overall quality of life: A general population survey in finland. *Nicotine and Tobacco Research*, 10(7):1199–1207, 2008.

[10] Richard Hipp et al. Sqlite. `http://www.sqlite.org`, 2009.

[11] Jaakko Hollmén. *User Profiling and Classification for Fraud Detection in Mobile Communications Networks*. PhD thesis, Helsinki University of Technology, 2000.

[12] M. Juonala, J.S.A. Viikari, M. Kähönen, T. Laitinen, L. Taittonen, B.-M. Loo, A. Jula, J. Marniemi, L. Räsänen, T. Rönnemaa, and O.T. Raitakari. Alcohol consumption is directly associated with carotid intima-media thickness in finnish young adults. the cardiovascular risk in young finns study. *Atherosclerosis*, 204(2):e93–e98, 2009.

[13] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[14] Seppo Koskinen, Tommi Härkänen, Pekka Jousilahti, Kari Kuulasmaa, Tuija Martelin, Ari-Pekka Sihvonen, and Ninni Vanhalakka. Pelataan elämän peliä. Internet: `http://elamapelissa.yle.fi/sites/elamapelissa.yle.fi/pelataan_elaman_pelia.pdf`, 2007. In Finnish.

[15] A. Kouvonen, M. Kivimäki, S.J. Cox, T. Cox, and J. Vahtera. Relationship between work stress and body mass index among 45,810 female and male employees. *Psychosomatic Medicine*, 67(4):577–583, 2005.

[16] A. Laszkiewicz, Sz. Szymczak, and S. Cebrat. Prediction of the human life expectancy. `http://arxiv.org/abs/cond-mat/0305277`, 2003.

[17] Large synoptic survey telescope. Internet: `http://www.lsst.org/lsst/about/technology`, 2009. Accessed July 15th, 2009.

[18] Alexander Ludwig and Alexander Zimper. A parsimonious model of subjective life expectancy. `http://ideas.repec.org/p/mea/meawpa/07154.html`, 2007.

[19] John Maindonald and John Braun. *Data Analysis and Graphics using R*. The Cambridge University Press, 2nd edition, 2007.

[20] Mayo Clinic. Cholesterol levels: What numbers should you aim for? `http://www.mayoclinic.com/health/cholesterol-levels/CL00001`, 2009. Accessed September 1st, 2009.

[21] Pia Mäkelä. Alcohol-related mortality by age and sex and its impact on life expectancy: Estimates based on the finnish death register. *European Journal of Public Health*, 8(1):43–51, 1998.

[22] Michael Negnevitsky. *Artificial Intelligence*. Addison Wesley, 1st edition, 2002.

[23] Oxford english dictionary. Internet: `http://dictionary.oed.com`, 2009.

[24] S.J. Olshansky et al. Will obesity dramatically lower life expectancy in the united states? Internet: `http://healthfullife.umdnj.edu/archives/obesity_archive.htm`, 2005. Accessed July 27th, 2009.

[25] M.-L. Ovaskainen, R. Törrönen, J.M. Koponen, H. Sinkko, J. Hellström, H. Reinivuo, and P. Mattila. Dietary intake and major food sources of polyphenols in finnish adults. *Journal of Nutrition*, 138(3):562–566, 2008.

[26] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, 2000.

[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. `http://www.R-project.org`.

[28] T. Tammelin, U. Ekelund, J. Remes, and S. Näyhä. Physical activity and sedentary behaviors among finnish youth. *Medicine and Science in Sports and Exercise*, 39(7):1067–1074, 2007.

[29] Terveyden ja hyvinvoinnin laitos. Internet: `http://www.thl.fi`, 2009. Accessed July 17th, 2009.

[30] J. Tikka and J. Hollmén. Sequential input selection algorithm for long-term prediction of time series. *Neurocomputing*, 71(13-15):2604–2615, 2008.

[31] Jarkko Tikka. *Input Variable Selection Methods for Construction of Interpretable Regression Models*. PhD thesis, Helsinki University of Technology, 2008.

[32] Edward Tufte. *The visual display of quantitative information*. Graphics Press, 2nd edition, 2001.

[33] W.N. Venables and B.D. Ripley. *Modern applied Statistics with S-PLUS*. Springer, 2001.

Appendix A

Elämä Pelissä test

The following table describes the questions asked in the *Elämä Pelissä* test [14]. Discrete answers are marked as [ ]. Continuous answers are marked as __. Question 40 is only available in the test on the web.

Table A.1: Questions asked in the *Elämä Pelissä* test.

	Variables	Question given Possible answer options
1.	Gender	What is your sex? [ ]Male [ ]Female
2.	Age	How old are you? __Years
3.	Length	How tall are you? __Centimeters
4.	Weight	How much do you weigh? __Kilos
5.	Education	For how many years have you gone to school? __Years
6.	Income	What was your total income last year (after taxes)? __Thousands of euros
7.	Household size	Number of people living in your house (including yourself) __Adults (18+ years) __Minors (<18 years)
8.	Economic situation	How is your actual financial status compared to the past? [ ]Much better [ ]Slightly better [ ]The same [ ]Slightly worse [ ]Much worse
9.	Cholesterol levels	What is your cholesterol level (mmol/L) for ? __Total Cholesterol __HDL Cholesterol
10.	Blood pressure	What is your blood pressure (mmHg)? __At Systole __At Diastole

	Variables	Question given Possible answer options
11.	Diabetes status	Do you have diabetes? [ ]No [ ]Yes, Treatment is diet [ ]Yes, Treatment is medicine/insulin
12.	Father infarct	Has your father have a myocardial infarct? [ ]Yes [ ]No
13.	Mother infarct	Has your mother have a myocardial infarct? [ ]Yes [ ]No
14.	Alcohol use	How many glasses or bottles do you drink in a week of? __Beer III, IV (bottle 1/3l) __Long drink (bottle 1/3l) __Concentrated alcohol (shot 4cl) __Wine (glass 12cl) __Light wine (glass 12cl)
15.	Drunkenness	How often do you drink so that you get drunk? [ ]2+/Week [ ]1+/Week [ ]1+/Month [ ]<1/Month
16.	Fresh vegetables	How often do you eat fresh vegetables? [ ]<1/Week [ ]1-2/Week [ ]3-5/Week [ ]Daily
17.	Fruits and berries	How often do you eat fresh fruits or berries? [ ]<1/Week [ ]1-2/Week [ ]3-5/Week [ ]Daily
18.	Butter type	What do you use as *fat* for bread? [ ]Butter [ ]Margarine [ ]Nothing
19.	Cooking oil	What kind of cooking oil do you use? [ ]Vegetable oil [ ]No fats oil [ ]Butter/Margarine
20.	Milk	What kind of milk do you drink? [ ]Whole [ ]Semi-skimmed [ ]Number one [ ]Skimmed [ ]Do not drink
21.	Smoking	Have you ever smoked daily for at least one year? [ ]Yes [ ]No
22.	Smoking now	Do you smoke now? [ ]Yes [ ]Randomly [ ]No
23.	Number of cigarettes	How many cigarettes do you smoke daily? __Cigarettes
24.	Sports	How much physical activity do you do per week? [ ]Nothing hard (house chores) [ ]Light activities (walking) [ ]Normal activities (jogging) [ ]Hard activities (competition)
25.	Seat belt	Do you use a seat belt while traveling in a car? [ ]Yes [ ]No
26.	Sleeping	How many hours in average do you sleep per day? __Hours
27.	Association activities	How often you go to a club or association? [ ]1+/Week [ ]1+/Mont [ ]1+/Year [ ]Almost never
28.	Theater and movies	How often do you go to the theater or movies?

	Variables	Question given Possible answer options
		[ ]1+/Week [ ]1+/Mont [ ]1+/Year [ ]Almost never
29.	Religious events	How often do you assist to church or religious events? [ ]1+/Week [ ]1+/Mont [ ]1+/Year [ ]Almost never
30.	Reading and music	How often do you read books or listen to music? [ ]1+/Week [ ]1+/Mont [ ]1+/Year [ ]Almost never
31.	Hobbies	How often do you engage in hobbies-like activities? [ ]1+/Week [ ]1+/Mont [ ]1+/Year [ ]Almost never
32.	Spouse	Do you have issues with your spouse? [ ]Not married [ ]Almost all the time [ ]Sometimes [ ]Never
33.	Children	Do you have issues with your children? [ ]Not have children [ ]Almost all the time [ ]Quite often [ ]Sometimes [ ]Rarely [ ]Never
34.	Stress	Have you felt stressed the last months? [ ]Yes, life is unbearable [ ]Yes, more than average people [ ]Yes, somewhat [ ]Not at all
35.	Satisfaction	How satisfied are you with your achievements in your life? [ ]Very pleased [ ]Pleased [ ]Somewhat satisfied [ ]Dissatisfied
36.	Work and study load	Do you feel your work and/or study load is high? [ ]Do not have paid work [ ]All the time [ ]Quite often [ ]Sometimes [ ]Rarely
37.	Dreams	What do you think about: "It seem to me impossible to achieve dreams" [ ]Agree [ ]Somewhat agree [ ]Hard to say [ ]Slightly disagree [ ]Disagree
38.	Friends	What do you think about: "I do not have any good friend" [ ]Agree [ ]Disagree
39.	Age to live	Until what age will live? ( )Years
40.	Best age to live	When is the best age to be alive? ( )Years

Appendix B

Outliers search & labeling

The following table depicts the summary information of the continuous variables during the filtering process in the *Elämä Pelissä* data base.
† Stands for a filter that covers the elements from 0 to $\mu < +3\sigma$
‡ Stands for a filter that covers the elements within the range $-3\sigma < \mu < +3\sigma$

Table B.1: Filtering out outliers from original database.

	Age			**Height**		
	Initial	Temp.	Final	Initial	Temp.	Final
Min	0	0	**1**	0		**130**
Max	299830	495	**88**	999		**211**
Mean	45.35	43.25	43.15	170.5		170.67
Variance	594993.9	226.37	217.24	186.86		88.93
Std Dev	771.36	15.05	14.74	13.67		9.43
Median	44	44	44	170		170
Filter	$x < 500$	†		‡		
Outliers	64	737	**801**	1483		**1483**
	Weight			**Education**		
Min	0	0	**22**	0	0	**2.5**
Max	999	499	**130**	999	89	**25**
Mean	76.24	76	75.31	14.05	13.97	13.93
Variance	523.01	332.95	260.16	69.66	15.58	13.1
Std Dev	22.87	18.25	16.13	8.35	3.95	3.62
Median	74	74	74	14	14	14
Filter	$x < 500$	‡		$x < 90$	‡	
Outliers	144	4223	**4367**	134	3141	**3275**
	BMI			**Income**		
Min	0	10	**11.34**	0		**0**
Max	$9.99\text{x}10^{8}$	50	**40.35**	5472		**278**
Mean	4905.44	25.85	25.6	55.93		48.53
Variance	$4.43\text{x}10^{12}$	23.45	19.18	5494.75		1202.75
Std Dev	2105225	4.84	4.38	74.13		34.68

Median	25.1	25.06	25	45		45
Filter	$10 \leq x \leq 50$	‡		†		
Outliers	2765	6236	**9001**	6667		**6667**
	Cholesterol total			**Cholesterol HDL**		
Min	0	0	**2**	0	0	**0**
Max	999	19	**8**	999	19	**4.5**
Mean	5.67	5	4.98	2.52	1.8	1.73
Variance	424.97	1.03	0.85	531.64	0.84	0.44
Std Dev	20.61	1.02	0.92	23.06	0.92	0.66
Median	5	5	5	1.6	1.6	1.5
Filter	$x < 20$	‡		$x < 20$	†	
Outliers	670	2036	**2706**	342	2080	**2422**
	Blood pressure systole			**Blood pressure diastole**		
Min	0	43	**90**	0	41	**54**
Max	999	179	**168**	999	99	**99**
Mean	129.79	129.1	128.92	79.98	79.08	79.18
Variance	441.44	174.46	157.33	368.8	72.93	70.34
Std Dev	21.01	13.21	12.54	19.2	8.54	8.39
Median	130	130	130	80	80	80
Filter	$40 < x < 180$	‡		$40 < x < 100$	‡	
Outliers	2798	2926	**5724**	7211	1000	**8211**
	Alcohol beer			**Alcohol long drink**		
Min	0	0	**0**	0	0	**0**
Max	999	140	**27**	999	140	**9**
Mean	4.08	3.69	2.89	0.79	0.42	0.24
Variance	382.03	64.19	23.48	331.62	9.54	0.82
Std Dev	19.55	8.01	4.85	18.21	3.09	0.9
Median	1	1	1	0	0	0
Filter	$x \leq 140$	†		$x \leq 140$	†	
Outliers	320	8495	**8815**	268	3601	**3869**
	Alcohol concentrated			**Alcohol wine**		
Min	0	0	**0**	0	0	**0**
Max	999	140	**17**	999	140	**17**
Mean	1.71	1.29	0.81	2.18	1.8	1.36
Variance	368.27	29.02	5.04	342.31	26.04	6.89
Std Dev	19.19	5.39	2.24	18.5	5.1	2.62
Median	0	0	0	0	0	0
Filter	$x \leq 140$	†		$x \leq 140$	†	
Outliers	351	6522	**6873**	322	6406	**6728**
	Alcohol light wine			**Alcohol total**		
Min	0	0	**0**	0	0	**0**
Max	999	14	**1.5**	4495	250	**42**
Mean	0.49	0.07	0.01	9.26	7.03	5.95
Variance	312.93	0.36	0.01	7842.6	137.93	58.8
Std Dev	17.69	0.6	0.11	88.56	11.74	7.67

Median	0	0	0	3	3	3
Filter	$x \leq 140$	†		$x \leq 200$	†	
Outliers	1066	6267	**7333**	810	7963	**8773**
	Cigarettes per day			**Sleeping hours**		
Min	0	0	**0**	0	0.1	**4.5**
Max	999	150	**38**	999	23	**10**
Mean	8.21	6.52	5.74	7.52	7.48	7.48
Variance	1532.05	110.82	67.07	30.01	1.14	0.9
Std Dev	39.14	10.53	8.19	5.48	1.07	0.95
Median	0	0	0	0	7.5	7.5
Filter	$x \leq 150$	†		$0 < x < 24$	‡	
Outliers	494	3227	**3721**	320	5052	**5372**
	Age to live			**Best time to live**		
Min	0	1.6	**57**	0	1.5	**1.5**
Max	999	119	**111**	3050	89	**78**
Mean	86.89	84.41	84.89	36.75	37.48	37.43
Variance	1742.21	84.37	60.97	516.48	189.95	187.7
Std Dev	41.74	9.19	7.81	22.73	13.78	13.71
Median	85	85	85	38	40	40
Filter	$1 < x < 120$	‡		$1 < x < 90$	†	
Outliers	5249	5230	**10479**	14221	455	**14676**
	Household size adults			**Household size minors**		
Min	0	0	**0**	0	0	**0**
Max	9787	50	**4**	999	50	**5**
Mean	1.83	1.68	1.66	1.77	1.44	1.38
Variance	313.47	0.68	0.38	276.34	1.81	1.12
Std Dev	17.71	0.82	0.61	16.62	1.35	1.06
Median	2	2	2	1	1	1
Filter	$x \leq 50$	†		$x \leq 50$	†	
Outliers	125	1107	**1232**	143	1519	**1662**

Appendix C

Selection of samples

Based on the original database ten smaller databases were generated. The first one, called Complete (C), includes only the complete records, i.e., the records where all the variables have a value. The other nine databases are divided into three sets of three. Each set has a similar number of elements per database, 10k, 50k and 200k records; i.e., there are three databases with 10k elements, three with 50k and three with 200k. These databases were populated with random records selected from the original database. Only valid records, not outliers, where available to be selected. The random selection was without repetition.

The sampling was done with the purpose of showing that there is not a compulsory necessity of using the full database; a large enough sample presents similar characteristics. This characteristic is mainly used during the regression analysis. In the k-fold cross-validation approach repeated M times a sample of only 20k elements is used at each iteration. The reduction of the number of elements speeds-up the process without a significant influence in the result.

The following table depicts the statistics of the artificially generated databases; the average of values of each set of three is the one used. These values are, in general, really close to the ones observed at Appendix B (last column, FINAL).

Table C.1: Statistics from the sampled databases.

	Age				**Height**			
	C	10k	50k	200k	C	10k	50k	200k
Min	5	5	1	1	135	130	130	130
Max	87	87.33	88	88	202	204	210	211
Mean	45.36	43.16	43.27	43.21	172.4	170.67	170.68	170.67
Variance	127.37	215.96	215.55	215.71	85.69	88.43	88.15	88.01
Std Dev	11.29	14.7	14.68	14.69	9.26	9.4	9.39	9.38
Median	45	44	44	44	172	170	170	170
	Weight				**Education**			
Min	24	25.33	22.33	22	3	3	2.5	2.5
Max	130	130	130	130	25	25	25	25
Mean	77.83	74.12	75.26	75.24	14.75	13.96	13.95	13.96

Variance	230.59	256.21	257.3	257.23	11.98	13	13.04	12.99
Std Dev	15.19	16.01	16.04	16.04	3.46	3.61	3.61	3.6
Median	77	73.67	74	74	15	14	14	14
	BMI				**Income**			
Min	12.6	12	11.57	11.37	0	0	0	0
Max	40.35	40.34	40.35	40.35	277	276.33	277.33	277.67
Mean	26.06	25.57	25.6	25.59	58.99	48.48	48.72	48.73
Variance	16.81	17.73	19.13	19.06	1333.41	1176.64	1204.11	1200.75
Std Dev	4.1	4.37	4.37	4.37	36.52	34.44	34.7	34.65
Median	25.47	24.97	24.99	24.98	55	45	45	45
	Cholesterol total				**Cholesterol HDL**			
Min	2	2	2	2	0	0	0	0
Max	8	8	8	8	4.5	4.5	4.5	4.5
Mean	4.91	4.96	4.98	4.98	1.68	1.72	1.72	1.73
Variance	0.84	0.84	0.84	0.84	0.43	0.43	0.43	0.43
Std Dev	0.92	0.92	0.92	0.92	0.66	0.65	0.65	0.66
Median	5	5	5	5	1.5	1.5	1.5	1.5
	Blood pressure systole				**Blood pressure diastole**			
Min	90	90	90	90	54	54	54	54
Max	168	168	168	168	99	99	99	99
Mean	128.7	128.77	128.94	128.88	79.41	79.09	79.16	79.16
Variance	139.48	152.87	155.98	156.43	66.3	70.24	69.94	70.19
Std Dev	11.81	12.36	12.49	12.51	8.14	8.38	8.36	8.38
Median	130	130	130	130	80	80	80	80
	Alcohol beer				**Alcohol long drink**			
Min	0	0	0	0	0	0	0	0
Max	27	26.33	27	27	9	8.67	9	9
Mean	3.34	2.86	2.86	2.86	0.24	0.23	0.23	0.24
Variance	24.34	23.19	23.25	23.19	0.76	0.76	0.78	0.79
Std Dev	4.93	4.82	4.82	4.82	0.87	0.87	0.88	0.89
Median	1	0.67	1	1	0	0	0	0
	Alcohol concentrated				**Alcohol wine**			
Min	0	0	0	0	0	0	0	0
Max	17	17	17	17	17	16.67	17	17
Mean	1.01	0.76	0.78	0.79	1.72	1.34	1.35	1.35
Variance	6.23	4.63	4.82	4.9	7.89	6.76	6.82	6.81
Std Dev	2.5	2.15	2.2	2.21	2.81	2.6	2.61	2.61
Median	0	0	0	0	0	0	0	0
	Alcohol light wine				**Alcohol total**			
Min	0	0	0	0	0	0	0	0
Max	1	1	1	1.33	42	42	42	42
Mean	0.02	0.01	0.01	0.01	6.32	5.91	5.92	5.94
Variance	0.02	0.05	0.01	0.01	46.09	58.73	58.14	58.47
Std Dev	0.12	0.11	0.1	0.11	6.79	7.66	7.63	7.65
Median	0	0	0	0	4	3	3	3

	Cigarettes per day				**Sleeping hours**			
Min	0	0	0	0	4.5	4.67	4.5	4.5
Max	35	36.33	37.67	38	10	10	10	10
Mean	3.27	5.78	5.66	5.69	7.42	7.48	7.48	7.48
Variance	42.58	66.81	66.18	66.08	0.77	0.89	0.89	0.89
Std Dev	6.53	8.17	8.14	8.13	0.88	0.95	0.95	0.94
Median	0	0	0	0	7	7.5	7.5	7.5
	Age to live				**Best time to live**			
Min	57	57.33	57	57	2	2	1.83	1.5
Max	111	110.67	111	111	77	75.67	77.67	78
Mean	85.06	84.92	84.91	84.88	38.65	37.43	37.52	37.47
Variance	57.14	60.39	60.19	60.62	164.74	186.91	186.44	187.15
Std Dev	7.56	7.77	7.76	7.79	12.84	13.67	13.65	13.68
Median	85	85	85	85	40	40	40	40
	Household size adults				**Household size minors**			
Min	0	0	0	0	0	0	0	0
Max	4	4	4	4	5	5	5	5
Mean	1.7	1.66	1.66	1.66	1.15	1.38	1.38	1.38
Variance	0.34	0.37	0.38	0.37	1.13	1.13	1.12	1.11
Std Dev	0.59	0.61	0.62	0.61	1.06	1.06	1.06	1.06
Median	2	2	2	2	1	1	1	1

Appendix D

Data exploration findings

The following tables depict the most interesting findings during the data exploration analysis. The exploration consisted on simple queries used for displaying interesting information that is easily available. The variables were analyzed alone and mixed; discrete vs discrete, continuous vs continuous and discrete vs continuous. The letter k is used to indicate thousands, e.g., 40k equals 40,000.

Table D.1: Single variable exploration findings.

Age
There are two possible normal distributions with means around 25 and 45 years.
Alcohol, beer
One to five bottles per week is what the majority (29%) of the people drinks.
Alcohol, light wine
Few people drinks light wine (1%).
Alcohol, wine
Between 20k to 40k respondents drink 1 to 4 glasses per week.
Alcohol, drink to get drunk
22k (5%) respondents get drunk more than once a week.
Association activities
Around 241k (50%) respondents almost never participates.
Blood pressure diastole
Almost everybody (70%) between 75 and 90 as expected.
Blood pressure systole
Almost everybody (71%) between 120 and 140 as expected.
BMI
Normal distribution with mean at 25.
Cholesterol HDL
Highest proportion of respondents with levels at 1 and 2 mmol/L.
Cholesterol total
Normal distribution with mean at 5 mmol/L.
Cooking oil
Vegetable oil is the most used one with 282k (62%) respondents using it.

Diabetes

21k cases of diabetes (4.5%), 14k (4%) are treated with insulin.

Do not have good friends

52k (11%) respondents claim not having good friends.

Economic situation

Better and *Much better* situations are the most answered (63%) options.

Education

Dependent on the age of the people with a peak at 12 years.

Height

Normal distribution with mean at 170 cms.

Hobbies

More than once a week has 167k (37%) respondents.

Infarct father

Around 84k (18.8%) respondents with a *Yes* answer.

Infarct mother

Around 37k (8.3%) respondents with a *Yes* answer.

Income

Resembles a Poisson distribution. The higher point at 50k euros.

Issues children

147k (33%) respondents do not have children.

Issues spouse

Sometimes answer is the highest with 216k (48%);

91k respondants (20%) are single.

Milk

138k (30%) respondents do not consume milk, the highest proportion consume fat free 197k (43%).

Seat belt

There are 13k (2.8%) respondents that does not use seat belt.

Sex

There are 70k more woman than man in the data set.

Size of the house adults

One and two account for the majority of the answers (94%).

Sleeping time

7 and 8 hours are the average, as expected.

Smoking cigarettes per day

There is a tendency to round the answers, 10, 15 and 20.

Twenty being the most common answer.

Smoking daily for a year

197k (43%) respondents have been smokers for a year.

Smoking now

76k (17%) respondents are smoking now and 35k (8%) respondents tend to smoke randomly.

Stress

277k (61%) respondents are *somehow* stressed.

Weight

Normal distribution with mean at 70kg.

What age will live

Increased age to 91 years when respondants are between 80 to 90 years old.

Work and study load

56k (12%) respondents do not have a paid job.

Table D.2: Discrete vs Discrete exploration findings.

Association activities	Do not have good friends
Number of answers of not having good friends increase when the association activities occur almost never.	
Association activities	Religious events
Highly correlated, 35% respondents have a *Never* answer in both.	
Economic situation	Association activities
A better economic situation equals higher association activities.	
Economic situation	Dreams impossible
With a worse economic situation it is more difficult to achieve dreams.	
Economic situation	Life satisfaction
There is linear correlation, good economic situation equals to a satisfaction with life.	
Economic situation	Movies and theater
A better economic situation means people assist more often to these events.	
Economic situation	Stress
In a worse economic situation people is more stressed.	
Father infarct	Mother infarct
3.0% respondents have both as *Yes*.	
Get Drunk	Smoking now
People that get drunk is no more propense to be a smoker, there are a similar number of smokers (4-6%) in all drinking categories.	
Get Drunk	Sports activities
People that like to get drunk less is more sporty.	
House size adults	Dot not have good friends
The proportion of not having good friends is high, 6%, when there are only two adults in the house.	
House size adults	Get drunk
In cases of 1 or 2 adults the same drinking patterns are observed. As if couples are going out together.	
House size adults	Hobbies
In cases of 1 or 2 adults the same hobbies responses are observed.	
House size adults	House size minors
There is a high presence of minors when there are 2 adults; 1 minor 19%, 2 minors 21%.	
House size minors	Life satisfaction
The presence of one or more minor have a positive effect in life satisfaction.	

Life satisfaction	Dreams impossible
Dissatisfied people tends to agree with the impossibility of achieving dreams.	
Religious events	Dreams impossible
People that go to religious events tend to be more positive about achieving dreams.	
Sex	Diabetes status
Males have a higher probability of having diabetes 6.1% vs 4.2% of females.	
Sex	Economic situation
Worse economic situation answer occurs more with females, 12.3%, than males, 9.4%.	
Sex	Fruits and berries
Males eat fruit in an uniformly distributed way, 24% fore each option.	
42% of females tend to eat fruits and berries daily.	
Sex	Get drunk
Females are less propense to get drunk, males get drunk more often.	
Once per week answered by 20% of males vs 9% of females.	
Sex	Kind of milk
44% of the consumed milk is fat free and females intake 61% of it.	
Sex	Seat belt
Of those not wearing a seat belt 70% are males and 30% females.	
Sex	Vegetables
32% or respondents are females eating vegetables daily.	
Smoking now	Stress
There is a tendency for smokers to consider themselves as stressed,	
15% of the respondents fall in this category.	
Sport activities	Dreams impossible
People doing *light* and *normal* sports usually agree that achieving dreams is possible.	
Sport activities	Life satisfaction
People that do sports is usually satisfied with their lives.	
Sport activities	Stress
Stress can be related to people doing no sports at all.	
Doing nothing and having more than normal stress levels account for 6% of respondents.	
Stress	Dreams impossible
No stressed people agreeing with the possibility of achieving dreams accounts for 9.8%	
of the respondents.	
Vegetables	Fruits and berries
Eating Vegetables implies the consumption of fruits and berries.	
Similar consumption patterns account for 51% of the answers.	
Vegetables	Sports activities
Both activities seem to be associated. Daily consumption is related with	
light and *normal* sports activities for 38.5% of respondents.	

Table D.3: Continuous vs Continuous exploration findings.

Age	Best time to live
People tend to answer with their own age.	
Age	Blood pressure diastole/systole
Older people have a tendency for a higher blood pressure.	
Age	BMI
Older people have a tendency for a higher BMI.	
Age	Cholesterol Total/HDL
Older people have a tendency for a higher cholesterol level.	
Age	Education
The majority of the people starts education at age 6 and continues up to age 24, new generations are more educated than older ones.	
Age	Total alcohol consumption
A noticeable increase in consumption after the age 18.	
Blood pressure diastole/systole	Total alcohol consumption
High pressure seems to be correlated with heavy alcohol consumption.	
Cholesterol HDL	Blood pressure diastole/systole
When the HDL is lower the blood pressure is slightly higher.	
Cholesterol HDL	BMI
It is noticeable that for lower levels of HDL the BMI is higher.	
Education	Alcohol beer
Less educated people tend to drink more beer.	
Education	Alcohol wine
More educated people tend to drink wine.	
Education	Total alcohol consumption
Less educated people drink more alcohol.	
Education	Income
Highly educated people have a better income, the average income in Finland is seen between 40k to 60k euros.	
Height	Cigarettes per day
Taller people seems to smoke more.	
Height	Weight
Variables are correlated. Taller people weights more, as expected.	
Income	Total alcohol consumption
With an income of more than 20k the alcohol consumption increases to 5 or more doses per week. Between 20k to 60k euros the consumption can be more than 10 doses.	
Number of cigarettes	What age will live
Heavy smokers (20+ cigarettes per day) reduce the number of years they will live by 5 to 10 years compared to the average.	
Weight	Blood pressure diastole
Heavier people have a higher blood pressure, as expected.	
Weight	Cholesterol HDL
Heavier people have a lower level of HDL.	
Weight	Number of cigarettes
Heavier people seems to smoke more (as taller people do).	

Table D.4: Discrete vs Continuous exploration findings.

Diabetes status	Age
People with diabetes controlled by diet have an age range between 50-65 years, people using insulin have an age range between 55-70 years.	
Diabetes status	Blood pressure diastole
People with diabetes show higher levels of blood pressure.	
Diabetes	BMI
People without diabetes have a range between 20-30, while diabetics have a range between 23-35.	
Diabetes	Education
People with diabetes have a lower number of years of education.	
Diabetes	Weight
Diabetics tend to be heavier, with a range between 70-100 kilos.	
Dreams impossible	Education
People with higher education disagree less with the impossibility of achieving dreams compared to people with less education.	
Economic situation	Age
The *Much better* economic situation shows between 50-70 years. The *Much worse* shows from 40-60 years and *Worse* shows in old people 60+ years.	
Economic situation	Education
Educated people have a much better situation that less educated people.	
Economic situation	Sleeping time
A much worse economic situation has people sleeping less than 6.5 or more than 9.5 hours.	
Economic situation	What age will live
People with much worse economic situation expect to live less.	
Father infarct	Age
It occurs when one is between 40-60 years old.	
Father infarct	BMI
With a father infarct people have a higher BMI.	
Father infarct	Education
A father infarct is related, strangely, to a lower level of education.	
Fruits and berries	Age
Daily consumption concentrates between people 50-60 years old.	
Fruits and berries	Blood pressure diastole
People with a daily consumption have a slightly lower lever of blood pressure.	
Fruits and berries	Education
People with a higher education level consumes fruits and berries more often.	
Get drunk	Alcohol, beer
Getting drunk more than once per week is correlated to high levels of beer consumption.	
Get drunk	Age
People getting drunk more than once per week are, mainly, between 40-60 years old.	
Get drunk	Education
Educated people tends to get drunk less often.	
House size minors	Age

Minors usually cohabitate with parents when parents are between 30-50 years old.

Issues children Age

The conflicts usually occur when parents are between 40-60 years old, specially at theirs 50's.

Issues children BMI

When there are issues all the time the BMI range is higher, 24-32; (the average is 23-28), with not children the BMI is below 20-25.

Issues children Cholesterol total

When there are issues all the time the cholesterol level range is higher (5-7mmol/L) compared to the average of (5mmol/L).

Issues children Education

When there are issues all the time the parents' education level is below average compared to other groups.

Issues children Income

Not having children is related to a lower income.

Issues children Number of cigarettes

Heavy smokers have a tendency for issues all the time.

Issues children Sleeping time

With issues all time the sleeping time is lower, 5-7 hours; than the average of 7-8 hours.

Issues children Total alcohol consumption

Issues all the time is related to heavy alcohol consumption.

Issues children Weight

Issues all the time is related to heavier people.

Issues spouse Blood pressure diastole

Higher levels of blood pressure is observed when there are issues all the time.

Issues spouse Number of cigarettes

Consumption of 20 to 30 cigarettes is related to issues all the time with spouse.

Issues spouse Total alcohol consumption

Issues all the time is related to high levels of alcohol consumption.

Cooking oil Total cholesterol

Higher levels of cholesterol are present when using butter/margarine.

Cooking oil Education

More educated people use vegetable oil, less educated people use butter/margarine.

Kind of milk Age

Older people (60-90 years) drink whole milk, people not drinking milk is between 40-60 years.

Kind of milk Alcohol, beer

Whole milk drinkers tend to drink more beers than the average.

Kind of milk Blood pressure diastole/systole

Whole milk drinkers have higher levels of pressure.

Kind of milk BMI

Whole and semi milk drinkers have a lower BMI, 20-27; compared to the average of 23-27.

Kind of milk Cholesterol total

Whole milk drinkers have a higher level of cholesterol.

Kind of milk Education

Whole milk drinkers have a lower level of education.

Kind of milk Income

Whole milk drinkers have a lower income.

Kind of milk Cigarettes per day

Whole milk drinkers smoke more than the average.

Life satisfaction Alcohol, beer

Dissatisfied people drink more beer.

Life satisfaction Age

Dissatisfied people cover all the range of ages, from 20 to 60 years old.

Life satisfaction Best time to live

Dissatisfied people set the best time to live around 20-30 (young ages).

Life satisfaction Blood pressure diastole

Dissatisfied people have a higher pressure level.

Life satisfaction Education

Very pleased people are more educated compared to dissatisfied people.

Life Satisfaction Height

Taller people is slightly more dissatisfied than average-height people.

Life Satisfaction Income

Satisfaction is highly correlated with the income.

Life Satisfaction Number of cigarettes

Dissatisfied people tend to smoke more.

Life Satisfaction What age will live

Dissatisfied people have a lower expectancy, 60-80 years; compared to the average of 80-90 years.

Mother infarct Age

It occurs when people is between 45-65 years.

Mother infarct BMI

With a mother infarct people have a higher BMI.

Mother infarct Education

A mother infarct is related, strangely, to a lower level of education.

Movies and theater Age

Younger people, 20-30 years, go more often, *1+/month*, to these events.

Movies and theater Education

More educated people go more often, *1+/month*, to these events compared to less educated people that goes *almost never*.

Movies and theater Income

With a higher income people assist more often to these events.

Movies and theater Cigarettes per day

Heavy smokers *Almost never* go to movies/theater.

Reading and music Alcohol beer

Heavy beer drinkers *Almost never* read or listen to music.

Reading and music Education

Reading and listening to music is positively correlated to the education level.

Reading and music Cigarettes per day

Heavy smokers almost never read or listen to music.

Reading and music Total alcohol consumption

High alcohol consumption is related to no reading nor listening to music.

Religious events Age

Assistance of *1+/month* shows a higher level between 50-70 years old people,
1+/week shows a high level between 50-60 years and 25-30 years.

Religious events — Total alcohol consumption

There is a decrease in alcohol consumption as the assistance to religious events increases.

Seat belt — Alcohol, beer and concentrated

Higher beer/alcohol concentrated consumption is related to no using seat belt.

Seat belt — Age

Younger people, 20-40 years, are the ones less likely to use the seat belt.

Seat belt — Education

Less educated people tend to use the seat belt less.

Seat belt — Income

Lower income is related to not using the seat belt (Is due to public transportation?).

Seat belt — Total alcohol consumption

Higher levels of alcohol are related to not using the seat belt.

Seat belt — What age will live

People not using the seat belt gave a lower range, 70-90 years; compared to the average of 80-90 years.

Sex — Alcohol, beer

Males have a higher maximum of beer consumption with 25 bottles per week compared to
10 bottles per week by females.

Sex — Age

The range of ages for both sexes is similar, 25-60 years.

Sex — Blood pressure diastole/systole

Higher pressure levels show more in males than females.

Sex — BMI

Males have a range between 23-30 while female have a range between 20-27.

Sex — Education

Females are more educated, 12-18 years, compared to the 9-18 years of males.

Sex — Height

The range for males is between 170-190cm, for female the range is between 155-170cm.

Sex — Number of cigarettes

Males smoke more; 25 and 30 cigarettes per day is mainly done by males.

Sex — Total alcohol consumption

Males have a higher maximum level of alcohol consumption, 35 doses vs 20 of females.

Sex — Weight

The average of males is between 70-100kg, for females the average is between 50-80kg.

Smoker for a year — Alcohol, beer

Smokers are related to higher levels of beer consumption.

Smoker for a year — BMI

Smokers have a higher BMI.

Smoker for a year — Education

Smokers have a lower education level.

Smoker for a year — Total alcohol consumption

High consumption of alcohol is related to a *Yes* in smoking for a year.

Smoker for a year — Weight

Heavier people tend to have been smokers for a year.

Smoking now Alcohol, beer

A *Yes* answer is related to heavy beer consumption, no smokers are related to lower beer consumption.

Smoking now Age

Random smokers are between 20-30 years old.

Smoking now Education

Smokers tend to have a lower level of education.

Smoking now Number of cigarettes

High presence at 10, 15, 20 (up to 30) cigarettes per day, random smokers are between 1-5 cigarettes per day.

Smoking now Total alcohol consumption

Follow the same tendency as beer consumption.

Smoking now What age will live

Smokers tend to have a lower life expectancy, 70-90 years, compared to the average of 80-90 years.

Sports activities Age

Hard sports are being done by people between 15-30 years.

Sports activities Blood pressure diastole/systole

The blood pressure is lower for people doing hard sports.

Sports activities BMI

People not doing sports have a higher BMI, 23-27, compared to the BMI of people doing *hard* sports, 20-25.

Sports activities Sleeping time

People doing hard sports tend to sleep more, 7-10 hours, than the average of 7-8 hours.

Sports activities Weight

People doing nothing is heavier than people doing some kind of sports. There is abundance of people weighting 90+ kilos.

Stress Age

The majority of the people with *no stress at all* are between 55-65 years old (28%).

Stress Blood pressure diastole/systole

When answer is *Yes it's a hard life* people have a higher pressure level.

Stress Income

The answer *Yes it's a hard life* is related to a lower income.

Stress Sleeping time

Stressed people sleep less, or more, than the average (5-10 hours).

No stressed people sleeps more, 7-10 hours.

Stress What age will live

Stressed people responded with a lower life expectancy, 60-90 years, than the average of 80-90 years.

Vegetables Age

People eating less than once per week are between 20-30 years, people with daily consumption are between 40-60 years.

Vegetables Education

Less educated people tend to consume less vegetables.

Vegetables Income

A higher income is related to daily consumption of vegetables.

Vegetables Number of cigarettes per day

Heavy smokers tend to eat vegetables more sporadically.

Work/Study load Age

Not paid work is usually present within people between 60-70 years old.

Work/Study load Best time to live

People with *not paid work* set an older best time to live range 40-60 years, compared to the average of 30-50 years.

Work/Study load Blood pressure diastole

People with an answer of *Always* show a higher pressure level.

Work/Study load Education

Not paid work is related to lower education level.

Work/Study load Sleeping time

High load is related to shorter sleeping time, 5-8 hours, while *not paid work* and *rarely* load have a longer sleeping time 7-10 hours.

Appendix E

Regression tables

The following tables depict the different regression strategies used for this thesis. Depending on the approach the tables also show the selected variables, their correspondent coefficient the Mean Squared Errors for the training and testing sets. All strategies were compared to the Mean predictor at Table E.1, using this last one as baseline. The improvement is also depicted in the tables at the last two columns. Long tables were divided in two; the first section shows the regression parameters and values, the second section shows the errors and improvements.

Variables Y	Mean predictor	MSE	RMSE
age	43.2196	215.7516	14.6885
height	170.6765	88.1681	9.3898
weight	75.2627	257.8829	16.0587
education	13.9488	13.0000	3.6055
income	48.6815	1197.5406	34.6055
cholesterolTotal	4.9799	0.8375	0.9152
cholesterolHDL	1.7279	0.4342	0.6590
bloodPressureSystole	128.8930	156.4849	12.5094
bloodPressureDiastole	79.1710	70.1752	8.3771
alcoholBeer	2.8687	23.1699	4.8135
alcoholDrink	0.2362	0.7912	0.8895
alcoholConcentrated	0.7914	4.9086	2.2155
alcoholWine	1.3528	6.8010	2.6079
alcoholLightWine	0.0113	0.0111	0.1055
numberOfCigarettes	5.6978	66.2813	8.1413
sleepingTime	7.4834	0.8906	0.9437
whatAgeWillLive	84.8797	60.5785	7.7832
bestTimeToLive	37.4743	186.9253	13.6721
BMI	25.5954	19.0848	4.3686
totalAlcoholConsumption	5.9380	58.4402	7.6446

Table E.1: Mean predictors for continuous variables.

Variables Y	Predictor X	Intercept	Slope	MSE training	MSE testing	Improvement training	Improvement testing
age	alcoholWine	41.6643	1.0531	208.4756	208.4796	0.03	0.03
height	sex	171.7531	-6.7647	43.5831	43.5841	0.51	0.51
weight	height	-88.1171	0.9576	177.2027	177.2066	0.31	0.31
education	alcoholWine	13.6960	0.1776	12.7673	12.7676	0.02	0.02
income	economicSituation	41.5295	9.1205	1108.3357	1108.3588	0.07	0.07
alcoholBeer	sex	3.1394	-1.4752	21.1359	21.1364	0.09	0.09
alcoholDrink	sleepingTime	0.4336	-0.0263	0.7879	0.7879	0.00	0.00
alcoholConcentrated	alcoholBeer	0.5072	0.0962	4.5860	4.5862	0.07	0.07
alcoholWine	age	-0.0736	0.0333	6.5961	6.5963	0.03	0.03
alcoholLightWine	alcoholWine	0.0131	-0.0013	0.0112	0.0112	-0.01	-0.01
sleepingTime	stress	7.7120	-0.2215	0.8675	0.8676	0.03	0.03
whatAgeWillLive	lifeSatisfaction	82.2874	1.4929	58.7030	58.7045	0.03	0.03
bestTimeToLive	age	16.1769	0.4940	134.5867	134.5901	0.28	0.28
cholesterolTotal	age	4.3413	0.0129	0.8127	0.8128	0.03	0.03
cholesterolHDL	weight	2.1231	-0.0051	0.4275	0.4275	0.02	0.02
bloodPressureSystole	weight	110.7684	0.2372	142.0263	142.0389	0.09	0.09
bloodPressureDiastole	weight	66.9990	0.1595	63.8639	63.8696	0.09	0.09
numberOfCigarettes	smokingNow	-0.4573	7.1280	22.2687	22.2710	0.66	0.66

Table E.2: Regression with a single predictor.

Variables Y	Predictor X	Intercept	Sex	Age	X^1	X^2	X^3	X^4	X^5
height	weight	0.06994	-0.56846	-0.13737	0.28045	-0.03010	0.03187	-0.02252	0.00360
weight	height	-0.02820	-0.07411	0.19492	0.48191	0.04013	0.01739	-0.00507	-0.00138
education	height	0.02059	0.21028	-0.06333	0.18610	-0.00404	0.02665	-0.00744	-0.00118
income	houseSizeAdults	0.06133	-0.06043	0.13944	0.28243	-0.07002	-0.01128	0.00417	–
alcoholBeer	smokingNow	0.05284	-0.28839	-0.03349	0.29764	-0.05284	–	–	–
alcoholDrink	stress	-0.00446	-0.03148	-0.07110	0.01809	0.00167	0.00638	–	–
alcoholConcentrated	alcoholBeer	0.01822	-0.17771	0.08990	0.38770	0.09320	-0.21520	0.07343	-0.00749
alcoholWine	economicSituation	-0.01754	0.02635	0.18520	0.11005	0.02790	-0.02641	-0.00929	–
alcoholLightWine	sleepingTime	-0.00428	0.03192	0.01840	0.00598	0.00598	-0.00531	-0.00054	0.00061
sleepingTime	stress	-0.01682	0.06533	-0.15721	-0.18479	0.01682	–	–	–
whatAgeWillLive	lifeSatisfaction	-0.01459	0.06966	-0.00946	0.18267	0.01531	0.00087	–	–
bestTimeToLive	education	0.01718	0.05188	0.52700	-0.03711	-0.01775	0.00795	–	–
cholesterolTotal	alcoholBeer	0.015992	0.028146	0.184171	0.116005	-0.019072	-0.001036	0.000572	–
cholesterolHDL	weight	-0.025641	0.063274	0.036840	-0.132331	0.031040	0.011866	-0.003826	–
bloodPressureSystole	BMI	0.034368	-0.178180	0.212861	0.261529	-0.044149	-0.006727	0.006189	-0.001128
bloodPressureDiastole	BMI	0.028733	-0.140312	0.139877	0.295540	-0.031146	-0.015967	0.004042	–
numberOfCigarettes	smokingNow	-0.531695	-0.080528	0.035459	0.673183	0.531724	–	–	–

Table E.3: Polynomial regression with extra variables Sex and Age.

Variables Y	Predictor X	MSE training	MSE testing	Improvement training	Improvement testing
height	weight	0.3763	0.3838	0.62	0.62
weight	height	0.6450	0.6488	0.36	0.35
education	height	0.9524	0.9577	0.05	0.04
income	houseSizeAdults	0.9057	0.8910	0.09	0.11
alcoholBeer	smokingNow	0.8558	0.8437	0.14	0.16
alcoholDrink	stress	0.9911	1.0101	0.01	-0.01
alcoholConcentrated	alcoholBeer	0.9090	0.9241	0.09	0.08
alcoholWine	economicSituation	0.9593	0.9917	0.04	0.01
alcoholLightWine	sleepingTime	0.9981	1.0179	0.00	-0.02
sleepingTime	stress	0.9465	0.9453	0.05	0.05
whatAgeWillLive	lifeSatisfaction	0.9637	0.9846	0.04	0.02
bestTimeToLive	education	0.7166	0.7169	0.28	0.28
cholesterolTotal	alcoholBeer	0.9633	0.9506	0.04	0.05
cholesterolHDL	weight	0.9783	0.9570	0.02	0.04
bloodPressureSystole	BMI	0.8367	0.8229	0.16	0.18
bloodPressureDiastole	BMI	0.8743	0.8641	0.13	0.14
numberOfCigarettes	smokingNow	0.2898	0.2937	0.71	0.71

Table E.4: Mean squared errors and improvements for the training and testing data sets used in Polynomial regression with extra variables Sex and Age.

Variables Y	sex	age	height	weight	education	economicSituation	houseSizeAdults	alcoholBeer	alcoholWine	smokerForAYear	smokingNow	sleepingTime	stress	lifeSatisfaction	totalAlcoholConsumption
age	–	–	-0.1859	0.2571	-0.0776	–	–	-0.0578	0.1849	0.1589	-0.2065	-0.1419	-0.1223	–	–
height	-0.5810	-0.1007	–	0.3164	0.1110	–	–	–	0.0182	–	0.0084	–	–	–	–
weight	–	0.2032	0.5544	–	–	–	–	0.0494	–	–	–	–	–	-0.0515	–
education	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
income	–	0.1658	–	–	0.2022	0.2319	0.2324	0.0337	–	-0.1001	-0.0456	–	–	–	–
alcoholBeer	-0.2832	–	–	–	–	–	–	–	0.0476	-0.0074	-0.0278	–	–	–	–
alcoholDrink	–	–	–	0.0373	–	–	–	0.0354	–	0.0791	0.1878	–	–	–	–
alcoholConcent.	–	–	0.1399	–	-0.0766	–	–	0.1754	–	–	–	–	–	–	–
alcoholWine	–	0.1903	–	–	–	–	–	–	–	–	0.0082	–	–	–	–
alcoholLightWine	–	–	–	–	–	0.0031	–	–	-0.0318	–	–	–	–	–	–
sleepingTime	–	-0.1532	–	-0.0650	–	–	–	–	-0.0107	–	-0.0627	–	-0.1664	–	–
whatAgeWillLive	–	-0.0230	–	-0.0806	–	–	–	-0.0659	0.0083	–	-0.1427	–	–	0.1476	–
bestTimeToLive	–	0.5290	–	–	–	–	–	-0.0291	–	–	–	–	–	–	–
cholesterolTotal	–	0.1683	–	–	–	–	–	0.0606	0.0371	–	–	–	–	-0.0688	–
cholesterolHDL	–	–	–	-0.1200	–	–	–	–	–	–	-0.0095	–	–	–	–
bloodPressureSystole	–	0.2461	–	0.2198	–	–	–	0.0879	–	–	–	–	–	–	–
bloodPressureDiastole	–	0.1626	–	0.2576	–	–	–	–	–	–	–	–	–	–	0.1072
numberOfCigarettes	–	–	0.0004	0.0542	–	–	–	0.0415	–	–	0.8024	–	–	–	–

Table E.5: Variables and their coefficients selected while using the Forward stepwise selection algorithm for regression. Intercept for all variables is zero due to the normalization of data.

Variables Y	MSE training	MSE testing	Improvement training	Improvement testing
age	0.8248	0.8294	0.18	0.17
height	0.3931	0.3871	0.61	0.61
weight	0.6445	0.6534	0.36	0.35
education	0.9665	0.9645	0.03	0.04
income	0.8087	0.8211	0.19	0.18
alcoholBeer	0.8494	0.8892	0.15	0.11
alcoholDrink	0.9967	1.0273	0.00	-0.03
alcoholConcentrated	0.9336	0.9196	0.07	0.08
alcoholWine	0.9640	0.9613	0.04	0.04
alcoholLightWine	0.9989	0.9624	0.00	0.04
sleepingTime	0.9415	0.9467	0.06	0.05
whatAgeWillLive	0.9321	0.9321	0.07	0.07
bestTimeToLive	0.7170	0.7187	0.28	0.28
cholesterolTotal	0.9533	0.9555	0.05	0.04
cholesterolHDL	0.9853	0.9657	0.01	0.03
bloodPressureSystole	0.8456	0.8488	0.15	0.15
bloodPressureDiastole	0.8750	0.8775	0.13	0.12
numberOfCigarettes	0.3292	0.3290	0.67	0.67

Table E.6: Mean squared errors and improvements for the training and testing data sets while using Forward stepwise selection algorithm. In this case the baseline values to compare the result against is not the Mean predictor but value 1.0, this is due to the normalization of the data.

Variables Y	sex	age	height	weight	education	economicSituation	housesSizeAdults	alcoholBeer	alcoholWine	smokerForAYear	smokingNow	sleepingTime	stress	lifeSatisfaction	totalAlcoholConsumption
age	–	–	-0.1868	0.2561	-0.0774	–	–	-0.0575	0.1843	0.1611	-0.2065	-0.1422	-0.1218	–	–
height	-0.5816	-0.1018	–	0.3165	0.1055	0.0150	–	–	0.0177	-0.0118	0.0166	–	–	0.0155	–
weight	-0.0568	0.1957	0.5167	–	-0.0108	-0.0088	–	0.0302	-0.0267	0.0660	-0.0186	-0.0251	0.0388	-0.0271	–
education	0.2028	-0.0912	0.2245	–	–	0.0963	-0.0728	–	0.1331	-0.0838	-0.0701	-0.0228	0.0684	0.0546	–
income	-0.0152	0.1459	0.0664	–	0.1748	0.1921	0.2262	0.0114	0.1214	-0.0169	-0.0220	-0.0250	0.0330	0.1090	–
alcoholBeer	-0.2640	-0.0653	0.0095	0.0416	0.0146	–	-0.0217	–	0.0560	0.0863	0.1664	-0.0163	0.0303	-0.0428	–
alcoholDrink	–	-0.0873	–	0.0487	-0.0398	–	–	–	0.0412	–	0.0548	-0.0239	0.0189	-0.0211	–
alcoholConcentrated	-0.1585	0.0850	–	0.0771	–	–	–	0.1263	–	–	0.0961	–	–	–	–
alcoholWine	0.0528	0.2094	0.0433	-0.0381	0.1380	0.0411	0.0178	0.0618	–	0.0490	–	–	0.0194	0.0272	–
alcoholLightWine	0.0182	0.0291	–	-0.0072	0.0075	0.0047	-0.0050	-0.0240	-0.0382	-0.0200	–	–	0.0067	0.0031	–
sleepingTime	–	-0.1447	–	-0.0686	–	–	–	–	–	–	–	–	-0.1707	–	–
whatAgeWillLive	–	–	–	-0.0814	0.0546	0.0224	–	-0.0604	–	-0.0345	-0.1122	0.0260	-0.0628	0.1159	–
bestTimeToLive	0.0477	0.5137	-0.0331	0.0515	-0.0181	0.0223	0.0022	-0.0164	-0.0045	0.0187	–	–	0.0015	0.0396	–
cholesterolTotal	–	0.1641	-0.1582	0.2031	-0.0267	–	–	–	–	-0.0118	–	-0.0261	0.0502	-0.0422	0.0927
cholesterolHDL	0.1031	0.0380	-0.0322	–	–	–	–	–	0.0171	–	-0.0191	–	0.0219	–	0.0673
bloodPressureSystole	-0.1624	0.2202	–	–	–	–	–	–	–	–	–	–	–	-0.0554	0.0773
bloodPressureDiastole	-0.1017	0.1568	–	0.0565	–	–	-0.0136	–	0.0076	-0.0362	0.0136	-0.0260	0.0521	-0.0407	0.0931
numberOfCigarettes	-0.0436	0.0584	-0.0448	0.0938	-0.0626	-0.0177	–	–	-0.0626	–	0.7879	-0.0237	0.0178	-0.0114	0.0656

Table E.7: Variables and their coefficients selected while using the SISAL algorithm for regression. Intercept for all variables is zero due to the normalization of data.

Variables Y	MSE training	MSE testing	Improvement training	Improvement testing
age	0.8248	0.8212	0.18	0.18
height	0.3914	0.3881	0.61	0.61
weight	0.6369	0.6267	0.36	0.37
education	0.9053	0.8908	0.09	0.11
income	0.7780	0.7752	0.22	0.22
alcoholBeer	0.8360	0.8144	0.16	0.19
alcoholDrink	0.9823	0.9477	0.02	0.05
alcoholConcentrated	0.9032	0.8624	0.10	0.14
alcoholWine	0.9320	0.9090	0.07	0.09
alcoholLightWine	0.9954	0.9281	0.00	0.07
sleepingTime	0.9467	0.9240	0.05	0.08
whatAgeWillLive	0.9244	0.8987	0.08	0.10
bestTimeToLive	0.7120	0.7058	0.29	0.29
cholesterolTotal	0.9447	0.9315	0.06	0.07
cholesterolHDL	0.9722	0.9567	0.03	0.04
bloodPressureSystole	0.8303	0.8114	0.17	0.19
bloodPressureDiastole	0.8606	0.8543	0.14	0.15
numberOfCigarettes	0.3126	0.3087	0.69	0.69

Table E.8: Mean squared errors and improvements for the training and testing data sets while using SISAL algorithm. In this case the baseline values to compare the result against is not the Mean predictor but value 1.0, this is due to the normalization of the data.

Variables Y	NN size	MSE training	MSE testing	Improvement training	Improvement testing
Forward stepwise selection ALL					
cholesterolTotal	4-9-1	0.9454	0.9462	0.0546	0.0538
cholesterolHDL	2-4-1	0.9823	0.9825	0.0177	0.0175
bloodPressureSystole	3-7-1	0.8355	0.8357	0.1645	0.1643
bloodPressureDiastole	3-7-1	0.8586	0.8589	0.1414	0.1411
numberOfCigarettes	4-8-1	0.2748	0.2751	0.7252	0.7249
Forward stepwise selection TOP					
cholesterolTotal	3-6-1	0.9470	0.9475	0.0530	0.0525
cholesterolHDL	1-3-1	0.9829	0.9830	0.0171	0.0170
bloodPressureSystole	2-5-1	0.8437	0.8438	0.1563	0.1562
bloodPressureDiastole	2-5-1	0.8668	0.8670	0.1332	0.1330
numberOfCigarettes	1-2-1	0.2971	0.2972	0.7029	0.7028

Table E.9: Neural network models using parameters from the Forward selection algorithm. The first table shows the results when using all the selected parameters. The second table shows the results when using only the most important parameters from the selected ones (higher weight). Enough parameters are selected to have at least 80% of the total weight represented.

Variables Y	NN size	MSE training	MSE testing	Improvement training	Improvement testing
SISAL ALL					
cholesterolTotal	11-19-1	0.9212	0.9253	0.0788	0.0747
cholesterolHDL	7-6-1	0.9767	0.9779	0.0233	0.0221
bloodPressureSystole	4-8-1	0.8614	0.8618	0.1386	0.1382
bloodPressureDiastole	11-15-1	0.8475	0.8497	0.1525	0.1503
numberOfCigarettes	12-16-1	0.2440	0.2461	0.7560	0.7539
SISAL TOP					
cholesterolTotal	5-11-1	0.9278	0.9288	0.0722	0.0712
cholesterolHDL	4-8-1	0.9775	0.9785	0.0225	0.0215
bloodPressureSystole	3-7-1	0.8661	0.8663	0.1339	0.1337
bloodPressureDiastole	7-12-1	0.8493	0.8504	0.1507	0.1496
numberOfCigarettes	5-11-1	0.2600	0.2606	0.7400	0.7394

Table E.10: Neural network models using parameters from the SISAL algorithm. The first table shows the results when using all the selected parameters. The second table shows the results when using only the most important parameters from the selected ones (higher weight). Enough parameters are selected to have at least 80% of the total weight represented.

Variables Y	Selected components	MSE training	MSE testing	Improvement training	Improvement testing
Forward stepwise selection ALL					
cholesterolTotal	1,2	0.8020	0.8023	0.0424	0.0421
cholesterolHDL	1	0.4270	0.4271	0.0167	0.0165
bloodPressureSystole	1,2	132.5606	132.6229	0.1529	0.1525
bloodPressureDiastole	1,2,3	61.1725	61.2060	0.1283	0.1278
numberOfCigarettes	1,2,3	58.0981	58.1301	0.1235	0.1230
Forward stepwise selection TOP					
cholesterolTotal	1,2	0.8029	0.8032	0.0414	0.0410
cholesterolHDL	1	0.4283	0.4284	0.0136	0.0134
bloodPressureSystole	1,2	132.9796	133.0368	0.1502	0.1498
bloodPressureDiastole	1,2	62.0634	62.0887	0.1156	0.1152
numberOfCigarettes	1	22.3077	22.3368	0.6634	0.6630

Table E.11: Regression on principal components using the parameters selected by the Forward selection algorithm. The first table shows the results when using all the selected parameters. The second table shows the results when using only the most important components. Enough components were selected to have at least a 95% of the variance represented.

Variables Y	Selected components	MSE training	MSE testing	Improvement training	Improvement testing
SISAL ALL					
cholesterolTotal	1,2,3,4	0.7889	0.7894	0.0581	0.0575
cholesterolHDL	1,2,3	0.4272	0.4274	0.0163	0.0158
bloodPressureSystole	1,2	140.9091	140.9658	0.0995	0.0992
bloodPressureDiastole	1,2,3	61.1734	61.2064	0.1283	0.1278
numberOfCigarettes	1,2,3,4	57.3314	57.3707	0.1350	0.1344
SISAL TOP					
cholesterolTotal	1,2,3,4	0.7928	0.7932	0.0535	0.0529
cholesterolHDL	1,2,3	0.4271	0.4273	0.0165	0.0160
bloodPressureSystole	1,2	141.0715	141.1307	0.0985	0.0981
bloodPressureDiastole	1,2,3	61.0889	61.1229	0.1295	0.1290
numberOfCigarettes	1,2,3	55.5672	55.5992	0.1616	0.1612

Table E.12: Regression on principal components using parameters from the SISAL selection algorithm. The first table shows the results when using all the selected parameters. The second table shows the results when using only the most important components. Enough components were selected to have at least a 95% of the variance represented.