UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Statistical Analysis of the Data from PERSSILAA - Personalized ICT Supported Service for Independent Living and Active Ageing

Joana Horta Fernandes

Mestrado em Gestão de Informação
Especialização em Gestão e Análise de Dados

Trabalho de projeto orientado por:
Prof[a] Doutora Marília Cristina de Sousa Antunes

2016

Acknowledgments

During my academic journey and, in particular, in the last months of conclusion of my Master's project, the support of some really important people was crucial. To those I want to give a special thanks.

First of all, to my advisor, Professor Marília Antunes, who helped me in a very difficult time of my life, giving me the opportunity of developing the project. I already recognized her as an exceptional teacher, but in the course of this Master's project I had the opportunity and the pleasure to work directly with her, sharing knowledge and also a good personal relationship. I have to thank for all the seconds, minutes and hours dispensed to support and guide me in this final work, but also for the affection and the thoughtfulness with which the professor always treated me. Her friendship was very important in all the ups and downs of this journey.

To Professor Maria Antónia Turkman, who always behaved as a real advisor towards me, ready to help and guide me through this work, even not baring the official title. All her commitment to this project and the knowledge which provided me are not possible to measure in words. I have also to thank her for giving me the opportunity, by participating in this project, to know better the work done in a scientific research and to go through other experiences that greatly enriched me. My gratitude and affection for her are immense.

To Filipe, for all the love and understanding with which he dealt with me during this long and difficult process. For believing in me even when I doubted myself, minimizing all the obstacles on the way and giving me always something positive to think about. For always being by my side and never letting me give up.

To my parents, without whom this achievement would not have been possible. For all the values that they transmitted to me, which allowed me to get here, but also for giving me the freedom to choose my path and always supporting me unconditionally.

To Zé and Manuela, for all the supporting and encouragement words and the sound advice. For being my second parents and accompanying me all the way.

To all my friends, who always recognize in me more capabilities that I really know and make me believe that I can do anything. A special thanks: to Catarina, my old and future colleague, for motivating me and always putting a smile on my face; to Jessica for helping me with all of my doubts, sharing all the concerns and celebrating together all the successes; to Gru, João, Bia, Raquel and Nicole for all the friendship and support during this important phase, but also during all the years that they have been by my side; to Carolina, Ana and Raquel for being there from the beginning to the end of this academic adventure; to Joana and Gonçalo for all the affection and concern with me.

Thank you all for these and many other reasons. With you, all of this was much easier.

Abstract

The report of this project concerns the results of a statistical analysis carried out to assess the performance of the PERSSILAA's screening process. An initial framework on the PERSSILAA project is presented with emphasis regarding the screening process. A brief description of the statistical methodologies used to perform this analysis is also included in this report.

PERSSILAA is a FP7 funded European project with the objective of developing a new health care service supported by a technological platform access for older people. The project has three main phases: the screening module where the individuals who will participate in the next phases are selected; the monitoring module to assess the daily health status of the participants; the training module which contains physical exercises, nutritional advices and cognitive tasks to be performed or used by the participants. The screening process, which is the object of this work, is composed by a first step of self-assessment and a second step of face to face assessment. This work starts by explaining in detail how and with which tools the screening process is performed.

In the theoretical part of this report some well-known statistical methods used to analyse the data, such as regression models, cluster analysis, discriminant analysis and others, are summarized and a recent technique to principal components analysis of mixed data (numerical and categorical) is presented. It is assumed a basic knowledge in Statistics.

A validation of the database from PERSSILAA was also necessary to be carried out, before making any analysis. The errors spotted in the database and all the procedure to find them are also reported in this document.

The most important questions which the statistical study developed in this work attempts to answer are: "can the questionnaire of the 1^{st} screening assign a classification close to the one of the 2^{nd} screening (assuming that a face to face evaluation is more reliable than a self-assessment)?"; "what are the most relevant questions of the 1^{st} screening to the classification of the 2^{nd} screening?";"how the individual questions, regarding the specific domains (physical, cognitive, nutritional) from the 1^{st} screening relate to the respective scores and classifications of the 2^{nd} screening?"; "can a classification rule based on the results of the individual questions of the 1^{st} screening questionnaire better classify the individuals than the one defined in PERSSILAA protocol?".

Preliminary to answer these questions, the report also presents a descriptive statistical analysis on the characteristics of the participants in this study, a characterization of the different classification profiles of the 1^{st} screening based on the results from the questionnaire, a comparative study on the classifications, obtained according to the tools, between the populations of the four municipalities which participated in the project and an exploratory analysis to check the validity of the 1^{st} screening.

Keywords: PERSSILAA, Frailty, Screening Tools, Older Adults, Multivariate Data Analysis.

Resumo

Neste projeto final de mestrado foi desenvolvida uma análise estatística sobre a base de dados do projeto europeu PERSSILAA, com o principal objetivo de avaliar o desempenho do seu processo de triagem. Para além da descrição da metodologia aplicada, dos resultados obtidos, e da discussão dos mesmos, é também realizado um enquadramento no projeto PERSSILAA, com mais pormenor no processo de triagem avaliado neste trabalho, e uma sumariação dos principais conceitos teóricos envolvidos na aplicação dos métodos estatísticos utilizados nesta análise.

O projeto PERSSILAA é um projeto financiado pelo programa europeu FP7 que pretende desenvolver um novo serviço de cuidados de saúde para os mais idosos, com o apoio de uma plataforma tecnológica de fácil utilização para os participantes. Este projeto foi aplicado apenas em algumas regiões de Itália e Holanda, mas pretende ser alargado a toda a Europa. O programa oferecido pelo PERSSILAA organiza-se em três módulos: módulo de *screening* - onde é feita a triagem dos indivíduos que devem participar nos dois outros módulos do PERSSILAA, através de avaliação do estado de saúde dos mesmos, principalmente nos domínios físico, cognitivo e nutricional; módulo de *monitoring* - que pretende fazer uma monitorização diária do estado de saúde do idoso ao longo do programa, através de métodos simples e não intrusivos; módulo de *training* - programa de treino, disponibilizado aos idosos considerados em risco no módulo de *screening*, desenvolvido para prevenir o seu progresso para um estado de saúde frágil através de exercício físico, dicas nutricionais e tarefas de estimulação cognitiva. No processo de triagem, alvo da análise realizada neste trabalho, existem duas fases distintas. No primeiro momento de triagem é realizado um questionário de auto-avaliação por cada participante, enquanto no segundo a avaliação é feita por elementos da equipa do PERSSILAA e com testes que testam mais as capacidades reais do indivíduo nos domínios avaliados. Toda a explicação sobre o projeto PERSSILAA e o modo como se desenvolve o processo de triagem é mais detalhada neste relatório.

Vários métodos estatísticos de análise de dados multivariados foram aplicados sobre os dados do projeto, para a recolha de informações que pudessem gerar conclusões importantes sobre o processo de triagem. Neste documento é fornecido um suporte teórico que resume os principais conceitos necessários à compreensão das metodologias utilizadas, sem entrar em demasiado detalhe e admitindo alguns conhecimentos básicos de estatística. A análise de regressão é um dos temas abordados, com especial atenção nos modelos de regressão linear e logística, incluindo uma explicação sobre a estimação e interpretação dos modelos. Também a análise de *clusters* foi utilizada nesta análise e, portanto, são incluídas no resumo teórico definições das medidas de proximidade mais utilizadas e algumas explicações sobre os dois processos de classificação: hierárquica e não hierárquica. Na análise discriminante linear foram apresentadas soluções de discriminação para o caso com apenas dois grupos e para o caso com mais do que dois grupos. Neste resumo teórico também se inclui uma descrição dos três testes de hipóteses usados na comparação da classificação entre municípios: o teste do Qui-quadrado para homogeneidade, o teste Mann-Whitney-Wilcoxon e o teste de Kruskal-Wallis. É realizada também uma pequena exposição sobre a curva ROC, sendo mencionadas as medidas principais para a sua compreensão. Por fim, uma descrição pormenorizada de uma nova abordagem de análise por componentes principais para dados mistos (categóricos e numéricos) é incluída neste suporte teórico.

Antes de ser feita a análise estatística dos dados provenientes do processo de triagem do PERSSILAA, foi necessário fazer uma validação da base de dados, de forma a garantir que os valores analisados tinham sido calculados corretamente de acordo com o protocolo do projeto. Neste processo de validação foram procurados valores estranhos que fugissem claramente à definição das variáveis e verificados os cálculos dos *scores* e classificações dos vários testes. Os erros encontrados foram corrigidos posteriormente, criando uma nova base de dados que foi utilizada durante a análise estatística. Toda a descrição do processo de validação da base de dados e dos erros encontrados está também incluída neste documento.

Relativamente à parte central do projeto, a análise dos dados do PERSSILAA, esta foi dividida em três partes (1º *screening*, 2º *screening* e comparação entre *screenings*) nas quais a análise realizada e os conjuntos de dados utilizados foram diferentes, já que pretendiam responder a diferentes questões sobre o processo de triagem.

A análise realizada sobre os dados do 1º *screening* focou-se em descrever os indivíduos tanto quanto às suas características físicas e comportamentais, como quanto aos resultados dos testes e sua classificação e perceber se esta se mantém quando realizada por outros métodos, que não o descrito no protocolo do PERSSILAA. É portanto apresentada uma análise preliminar sobre as características dos participantes, suportada por representações gráficas que ajudam a visualizar os resultados. Para além da apresentação dos resultados da classificação (final e em cada domínio) são descritos, neste relatório, os resultados de um estudo comparativo das classificações entre as populações dos vários municípios participantes neste projeto. Uma caracterização dos vários perfis de classificação é mostrada através de *faces de Chernoff*, tentando identificar quais as principais diferenças nos questionários dos participantes com classificações distintas. Por fim, com base em métodos como análise de *clusters*, análise discriminante linear e regressão logística multinominal e assumindo a classificação do PERSSILAA como um padrão de excelência, foram atribuídas novas classificações aos indivíduos e comparadas com as classificações originais para perceber se ocorrem grandes alterações.

Os dados dos participantes do 2º *screening* produzem uma amostra muito menos diversificada da população, porque apenas os indivíduos classificados como *pre-frail* no 1º *screening* participam no 2º, com poucas exceções de alguns indivíduos classificados como *robust* e *frail* que também participaram. Por este motivo este conjunto de dados é muito desequilibrado e não permite produzir análises tão interessantes como as realizadas para o 1º *screening*. Consequentemente são apenas apresentados para os dados do 2º *screening* os resultados de uma análise preliminar sobre as características dos indivíduos, os resultados das classificações do 2º *screening* e um estudo comparativo das classificações entre os municípios, tal como foi realizado numa fase inicial da análise do 1º *screening*.

Finalmente, foi realizada uma comparação entre os resultados da classificação dos indivíduos que participaram nos dois *screenings*, procurando responder a algumas questões consideradas de interesse. Antes de mais, são mostradas as classificações dos indivíduos nos dois *screenings* e assim observadas as possíveis alterações de classificação do indivíduo, não devido à alteração real da sua condição, mas sim a diferenças na avaliação realizada no 1º e 2º *screenings*. Através de curvas ROC, incluídas e analisadas neste documento, foi possível avaliar o desempenho dos *scores* do 1º *screening*, correspondentes aos três domínios avaliados, como testes de diagnóstico para as respetivas classificações do 2º *screening*. A mesma análise foi realizada com o *score* do domínio físico do 1º *screening* e os resultados das classificações associadas a cada teste físico realizado no 2º *screening*. Foi também importante perceber quais as variáveis do 1º *screening*

que se mostravam mais relevantes para a classificação do 2º *screening*, tanto quanto às questões específicas de cada domínio, como a todas as questões do 1º *screening*. Para avaliar quais as questões mais relevantes de cada domínio do 1º *screening* para a classificação ou *score* do respetivo domínio do 2º *screening* foram utilizadas a regressão logística e a regressão linear múltipla. Quanto à classificação final do 2º *screening*, o objetivo foi identificar as variáveis mais informativas de todo o questionário do 1º *screening*, mas também perceber se seria razoável construir uma regra de classificação que tivesse como base estas perguntas e como resultado probabilidades de classificação como *pre-frail* para cada indivíduo. Esta última análise foi realizada também através da estimação de um modelo de regressão logística.

Este trabalho é importante para avaliar a qualidade do processo de triagem do PERSSILAA e por consequência aferir se a seleção dos participantes no programa é feita de forma correta. É crucial uma boa seleção dos indivíduos para que o programa utilizado pelos participantes, que foi construído unicamente para idosos com um certo estado de saúde, possa ter resultados positivos e ajudar a melhorar ou manter as condições de vida dos mesmos.

Palavras-chave: PERSSILAA, Fragilidade, Ferramentas de Triagem, Idosos, Análise de Dados Multivariados.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

ACF: *Autocorrelation Function*
AD8: *The AD8: The Washington University Dementia Screening Test*
AUC: *Area Under the Curve*
BMI: *Body Mass Index*
CSRT: *Chair Sit and Reach Test*
CST: *Chair-Stand Test*
FDR: *False Discovery Rate*
FP7: 7^{th} *Framework Programme*
FPF: *False Positive Fraction*
FWER: *Family-Wise Error Rate*
GFI: *The Groningen Frailty Indicator*
GSVD: *Generalized Singular Value Decompostition*
ICT: *Information and Communications Technology*
KATZ ADL: *The KATZ Index of Independence in Activities of Daily Living*
LDA: *Linear Discriminant Analysis*
MCA: *Multiple Correspondence Analysis*
MLR: *Multinomial Logistic Regression*
MNA: *Mini Nutritional Assessment*
MNA-SF: *Mini Nutritional Assessment Short Form*
MST: *Two-Minute Step Test*
NA: *Not Available*
PCA: *Principal Component Analysis*
PERSSILAA: *Personalised ICT Supported Service for Independent Living and Active Ageing*
QMCI: *The Quick Mild Cognitive Impairment*
ROC: *Receiver Operating Characteristic*
SF36: *36-Item Short Form Survey Instrument*
TNF: *True Negative Fraction*
TPF: *True Positive Fraction*
TUGT: *Timed Up and Go Test*

Chapter 1

Introduction

The increasing number of older people in developed societies is an issue of growing concern for the governments of several countries and some international organizations. Europe is one of the world's regions where this phenomenon occurs, and it can bring some serious problems, such as the unsustainability of the health care system. Consequently it is important to promote the welfare of older adults and ensure an healthy and active aging.

In order to achieve these objectives, first it is necessary to identify the individuals at risk of developing adverse health situations. Frailty is a common term to describe this health status, which may cause disability, morbidity, institutionalization or death. Although frailty is recognized as an important issue which needs to be treated, there is a lack of clinical assessment tools to help making this identification, and those that already exist have many limitations. Some are part of a time consuming process, others represent a very expensive solution and it is not possible to achieve a consensus on many of them. Furthermore, there is no known tool which can be used to evaluate the overall status (physical, cognitive and nutritional domains) of older adults which can be implemented in the European Union.

The existence of programmes which can identify older adults at risk of a frailty condition and prevent degradation allows avoiding some serious problems for health systems. All institutionalization and hospitalization represent a heavy burden on the existing health systems, so the increasing number of older adults may lead to an increase of costs in these services. This is another reason why a programme to prevent frailty and maintain a healthy status in older adults is needed.

The PERSSILAA project, developed by a group of research teams of European universities and institutions, intended to fill this gap building a new solution of frailty prevention, with the help of ICT tools. In addition to contributing to solve a critical problem of our society, PERSSILAA also proposes to do it through a reliable, efficient, easy to use and unexpensive program.

Concerning the need to identify older adults who should participate in the PERSSILAA programme, a screening procedure was developed and implemented. Screening focuses on the use of accessible and user friendly screening instruments to get an overall picture of an individual's nutritional, physical and cognitive state. However, there is a need to study the performance of these instruments so that a recommendation of its use can be given. The work presented in this report aimed precisely to evaluate the performance of these tools by means of adequate statistical methodologies and also to understand which are the main factors that affect functional decline.

Some well-known statistical methodologies were used to assess the performance of the screening process: regression models were carried out to select the most relevant questions of the 1^{st} screening tools to the classification of the 2^{nd} screening; some hypotheses tests were performed to find out if there were significant differences in the classification (final and in each domain) of the 1^{st} and 2^{nd} screenings between the populations of the different municipalities involved in the study; linear discriminant analysis, multinomial logistic regression and cluster analysis built three alternative ways of classifying the participants of the 1^{st} screening, based on the individ-

ual questions or scores of the questionnaire; receiver operating characteristic analysis helped to evaluate the performance of the tests, related to each domain of the 1st screening, as diagnostic tests for the corresponding classification in the 2nd screening. Besides these four main issues examined, other statistical studies have been conducted in this work.

This report describes the statistical analysis carried out, the results obtained and it also summarizes some important statistical concepts necessary to understand the results. The document is organized as follows: on Chapter 2 an overview of the PERSSILAA project and a detailed description of the screening process and the tools used in it are given; Chapter 3 is a theoretical summary of the statistical methods which were used in the practical part of this project (Chapters 5, 6 and 7); the database validation process is described in Chapter 4; Chapters 5 and 6 present the results and some comments of the analyses performed on the data of the 1st and 2nd screenings, respectively; a comparative study between the two screening results is also done and discussed in Chapter 7; finally, in Chapter 8 the main conclusions of this statistical study are summarized and some other statistical analysis that could be important for the PERSSILAA project are mentioned.

Chapter 2

PERSSILAA

For a good understanding of this Master's project it is necessary to first understand the main project in which it is inserted. This chapter intends to explain the objectives of the PERSSILAA project, how this programme is developed, giving a more detailed description of the screening process which is the study object of this work.

2.1 Overview

PERSSILAA (Personalised ICT Supported Service for Independent Living and Active Ageing) is a FP7 funded European project, which intends to develop and validate a new service to screen and prevent frailty in older adults. This new service model comprises the following three modules:

Screening: screening methods, easy to perform and understand, were developed giving an overall picture of the subject's health;

Monitoring: unobtrusive processes of monitoring were used to assess the everyday functioning of older adults;

Training: some health promotion programs were available for those elderly people who were selected in the screening module.

These modules are carried out based on three different domains of health: physical, cognitive and nutritional.

The service developed by PERSSILAA intends to create a new approach for the way care services are organized. The main objective is to fragment disease management into preventive personalized services provided by a proactive team of caregivers and health professionals in local community service, integrated into existing healthcare services. In the project, a technological platform was developed to provide these three types of service in an efficient, reliable and easy to use way, specially designed for its end-users, older adults. As it was referred to in the deliverables of the project, this infrastructure included work on gamification, interoperability and clinical decision support tools.

Frailty prevention is the main point of this project. Since there is not a unique definition of frailty, several definitions were discussed as well as the risk factors commonly associated to it, in order to build an universal definition in PERSSILAA's context. Hence, PERSSILAA's definition for frailty is

"Frail older adults are those at increased risk for future poor clinical outcomes, such as the development of disability, dementia, falls, hospitalisation, institutionalisation or increased mortality".

The PERSSILAA's project was implemented in community centers of some municipalities of the Netherlands and of the Campania region in Italy. However, only the data from the Netherlands' implementation was analyzed in this Master's project, since the data from Italy was not made available during its course.

2.2 Screening process

The screening process was initiated in 2014 and it was from then on a continuous procedure to portray the actual health status of the older adults in some municipalities of Netherlands. The assessment of the people's condition was done mainly based on physical, cognitive and nutritional domains, as mentioned above. The screening module is divided in two connected phases: 1st screening and 2nd screening. The main objective of the screening process is to establish an easy and reliable procedure to select pre-frail older people who can benefit from the PERSSILAA programme.

In the 1st screening, a self assessment questionnaire is filled in by the subjects, eventually with help of a family member. The questionnaire contains questions from different tests, set outside PERSSILAA's project. A 1st round of the 1st screening was performed at the beginning of the project with a questionnaire which is slightly different from the one applied on the 2nd and further rounds (which can be performed in the future) of the 1st screening and which is the basis of the statistical analysis developed in this work. While MNA-SF, SF36 and GFI tests were maintained in the 2nd round of the 1st screening, it became evident that KATZ ADL instrument was not bringing extra benefit in the physical classification of the individuals and hence was taken out from the questionnaire. The three first mentioned tests are used to assess the individuals' condition in nutritional and physical domains and general status, respectively.

In the questionnaire from the 1st round, it was also detected that evaluation of cognitive domain was insufficient (just one question from MNA-SF and another from GFI were used to assign a cognitive classification). For this reason the AD8 test was added to the questionnaire in the 2nd round of the 1st screening.

Besides the questions regarding those screening tools, the 1st screening's questionnaire contains also other general questions which are useful to characterize the individuals but were not used to classify the subjects. Each test included in the questionnaire allows the computation of a score which is used to classify the individuals in the three domains and also in a more general evaluation (GFI). With these scores PERSSILAA project proposes a classification rule to classify the individuals in the 1st screening in one of three classes: "FRAIL", "PRE-FRAIL" and "ROBUST". The process of classifying based on the scores is addressed later in this report.

The 2nd screening is done in person (face to face) in principle only with individuals who were classified as pre-frail in the 1st screening. All the tests of the 2nd screening are conducted by health professionals or members of PERSSILAA's team. Four practice tests and scores are used to classify the subjects with respect to physical domain. In the cognitive domain, some cognitive exercises are performed in the QMCI test. Regarding the nutritional domain, the complete MNA test is used to classify the individuals, being the only questionnaire applied in the 2 nd screening. No changes were made during the project concerning the 2nd screening procedures.

In Figure 2.1, all the steps in the screening process can be observed.

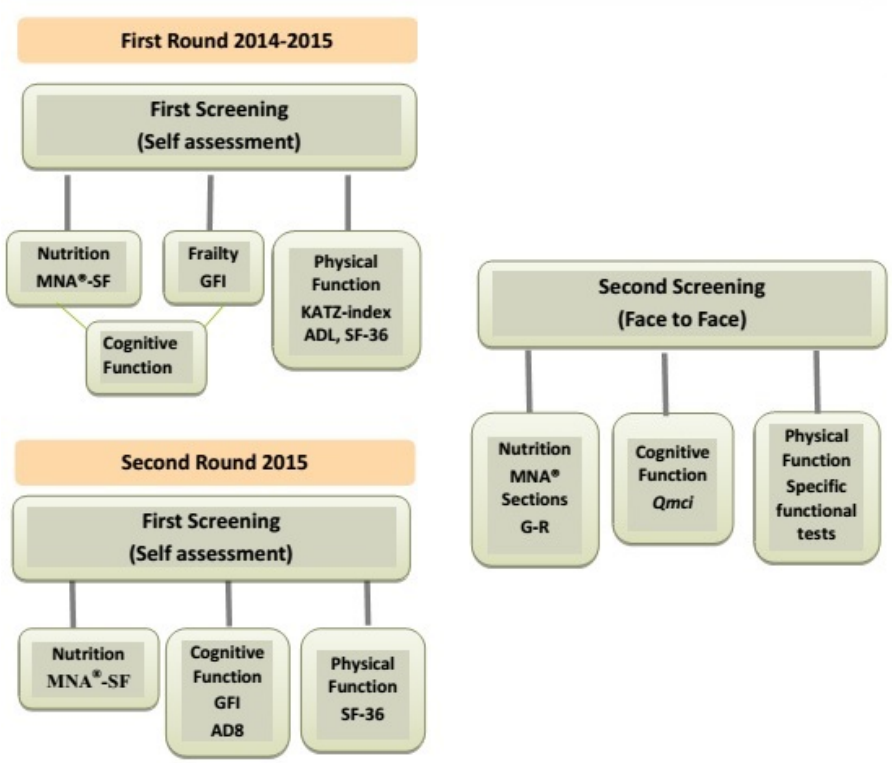


Figure 2.1: Screening process

2.3 Screening tools

The screening process uses several tools to classify the individuals with respect to their health status, including physical, cognitive and nutritional domains, as mentioned. In this section, the tests are described for each screening step and the correspondent classification rules are explained, as well as the final classification of each screening.

2.3.1 1st screening

Physical Domain

As mentioned above, the physical domain is evaluated in the 1st screening using the SF36 test. This test is performed in the 1st screening using 10 questions from the complete 36-item Health Survey (RAND-36). The focus of this questionnaire is to measure the limitations in daily activities as a result of health (physical) problems. The questions are scaled in a categorical variable with 3 classes (1-Yes, limited a lot; 2-Yes, limited a little; 3-No, not limited at all) and are set as:

1. to do vigorous activities;
2. to do moderate activities;
3. lifting or carrying groceries;
4. climbing several flights of stairs;
5. climbing one flight of stairs;
6. bending, kneeling or stooping;

7. walking more than one mile;
8. walking several blocks;
9. walking one block;
10. bathing or dressing yourself.

The SF36 score is the sum of a transformation of the questions results with values in a scale from 0 to 100. This transformation is given by the following expression:

$$T(\text{question result}) = (\text{question result} - 1) * 10/2 \tag{2.1}$$

In physical domain, the classification rule is:

- if SF36 score ≤ 60, classify the individual in "FUNCTIONAL DECLINE" class;
- if SF36 score > 60, classify the individual in "NORMAL" class.

Cognitive Domain

The test currently used to classify the subjects in relation to cognitive domain in the 1st screening is AD8 (The AD8: The Washington University Dementia Screening Test). In this tool the participants answer to 8 questions related to possible changes in their memory in a variety of tasks, during the past years. The answers for AD8 questions can be "yes", "no" or "don't know", but the results are codified as a binary variable (1 for "yes" and 0 for the other two), as it is specified in Galvin, et al., 2005 [7]. A descriptive list of the questions from AD8 test is shown below,

1. experiences any problems with judgement;
2. experiences less interest in hobbies and activities;
3. repeats the same things over and over;
4. has trouble learning how to use a tool, appliance, or gadget;
5. forgets correct month or year;
6. has trouble handling complicated financial affairs;
7. has trouble remembering appointments;
8. has daily problems with thinking and/or memory.

In cognitive domain, the individuals get a score between 0 and 8, which is just the sum of all the results of AD8.

The classification rule associated with AD8 and the cognitive domain is then,

- if AD8 score < 2, classify the individual in "NORMAL" class;
- if AD8 score ≥ 2, classify the individual in "FUNCTIONAL DECLINE" class.

Nutritional Domain

Regarding the nutritional domain, the tool used for classification in the 1st screening is a short form of the Mini Nutritional Assessment (MNA). This is a validated tool which includes anthropometric measurements, global assessment, dietary assessment and subjective evaluation on their questions. The MNA test was developed by Nestlé and is commonly used to identify older adults who are malnourished or in risk of it. The results from question to question vary in MNA, since they are quite different. In the list below the questions A-F from the short form of the MNA (MNA-SF) are described and the possible answers are in brackets for those which have multiple choice:

1. started to eating less as a result of loss of appetite, digestive problems, difficulty in chewing and / or swallowing in the last 3 months (0-Significant (greatly reduced appetite), 1-A little (moderate loss of appetite), 2-No (no loss of appetite));
2. loss of weight during the last 3 months (0-weight loss greater than 3 kg (6.6 lbs), 1-does not know, 2-weight loss between 1 and 3 kg (2.2 and 6.6 lbs), 3-no weight loss);
3. the extent that the subject is able to move (0-bed or chair bound, 1-able to get out of bed / chair but does not go out, 2-goes out);
4. has suffered psychological stress or acute disease in the past 3 months (0-Yes, 2-No);
5. experiences neuropsychological (0-severe dementia or depression, 1-mild dementia, 2-no psychological problems);
6. height (in m);
7. weight (in kg).

The original last question is about the Body Mass Index (BMI), but in the 1st screening questionnaire of PERSSILAA, just the weight and height of each individual are required to the participants for subsequent calculation and codification of the BMI on the PERSSILAA's platform. The calculation of BMI and its codification is as follows:

Calculation: $\text{BMI} = \frac{\text{weight (in kg)}}{\text{height (in m)}^2}$.

Codification:

- if BMI < 19, result=0;
- if $19 \leq$ BMI < 21, result=1;
- if $21 \leq$ BMI < 23, result=2;
- if BMI ≥ 23, result=3.

The score in the short form of MNA is the sum of all the results in a scale from 0 to 14. With this tool, the individuals are classified as undernourished, at risk of malnutrition or normal, based on the following rule,

- if MNA-SF score < 8, classify the individual in "UNDERNOURISHED" class;
- if $8 \leq$ MNA-SF score ≤ 12, classify the individual in "RISK OF MALNUTRITION" class;

- if MNA-SF score > 12, classify the individual in "NORMAL" class.

In the PERSSILAA's screening process this classification was converted into the same categories of the other domains of the 1^{st} screening. Therefore, the individuals considered as undernourished or in risk of malnutrition were classified in "FUNCTIONAL DECLINE" class, while the other class is the same.

General

The Groningen Frailty Indicator (GFI) is a tool which identifies the degree of frailty in older adults and is used in the screening process to immediately exclude the participants classified as frail. This test contains 15 questions, some about the general status of the individual and others from each domain involved in the classification. All the questions have a binary result, but sometimes with different meanings and values. Hence, the questions and their possible results, in brackets, are listed above,

1. can perform grocery shopping without assistance from another person (0-Yes, 1-No);
2. can walk outside house without assistance from another person (0-Yes, 1-No);
3. can getting (un)dressed without assistance from another person (0-Yes, 1-No);
4. can visiting the restroom without assistance from another person (0-Yes, 1-No);
5. how would the subject rate their own physical fitness (0-6 = 1, 7-10 = 0);
6. encounters problems in daily life because of impaired vision (0-Yes, 1-No);
7. encounters problems in daily life because of impaired hearing (0-Yes, 1-No);
8. unintentionally lost a lot of weight in the past 6 months (0-Yes, 1-No);
9. takes 4 or more different types of medication (0-Yes, 1-No);
10. has any complaints on your memory (0-No, 0-Sometimes, 1-Yes);
11. ever experience emptiness around yourself (0-No, 1-Sometimes, 1-Yes);
12. ever misses the presence of other people around their or miss anyone they love (0-No, 1-Sometimes, 1-Yes);
13. ever feels left alone (0-No, 1-Sometimes, 1-Yes);
14. is feeling down or depressed lately (0-No, 1-Sometimes, 1-Yes);
15. is feeling nervous or anxious lately (0-No, 1-Sometimes, 1-Yes).

The sum of all the binary results gives the score of GFI test. This score varies between 0 and 15 and based on this the following rule is built:

- if GFI score < 4, classify the individual in "ROBUST" class;
- if GFI score $= 4$, classify the individual in "PRE-FRAIL" class;
- if GFI score ≥ 5, classify the individual in "FRAIL" class.

Final Classification

Based on the scores from the last four mentioned tests of the 1st screening, it is possible to classify the individuals in relation to their total health status. As it is proposed by PERSSILAA the "FRAIL" status is assigned to a subject taking into account only the GFI score, while for the other two classes ("PRE-FRAIL" and "ROBUST") it is necessary a combination of all the scores. The classification rule for the final status of the 1st screening is then,

- if GFI score ≥ 5, classify the individual in "FRAIL" class;
- if GFI score < 4 & MNA-SF score > 11 & SF36 score > 60 & AD8 score < 2, classify the individual in "ROBUST" class;
- otherwise, classify the individual in "PRE-FRAIL" class.

2.3.2 2nd screening

Physical Domain

In the physical part of the 2nd screening, which is supposed to be done only with the participants classified as pre-frail in the 1st screening, four physical exercises are performed: the "timed up and go test" (TUGT), the "chair sit and reach test" (CSRT), the "chair-stand test" (CST) and the "two-minute step test" (MST). These tests are described below.

TUGT: The timed up and go test focuses in the assessment of the agility and balance of the subject, essentially to identify people in risk of falling. In this test the individuals should rise from a chair without the support of their arms, walk 10 feet as quickly as possible, turn and return to the chair. This test is timed in seconds and the value is inserted in the PERSSILAA's platform. TUGT is represented in PERSSILAA's database with three variables corresponding to each phase of the test, namely T2_PHY_TUGT_01, T2_PHY_TUGT_02 and T2_PHY_TUGT_03.

CSRT: The chair-stand test assesses the flexibility of the individuals, which is an important competence for good posture and some mobility tasks. In this test, the participant should sit at the front edge of a chair, bent one leg with the foot flat on the floor and extend straight the other with heel on floor and foot flexed. The goal is, with the hands of each other and arms outstretched, to reach as far forward as possible in the direction of the toes. The score of the test is the distance between the tip of the middle fingers and the toes, in centimetres. If the fingertips touch the toes then the score is zero. If they do not touch, the score is a negative value, while if they overlap, the score is positive.

CST: The chair-stand test has the objective of measuring the lower body strength of each individual. An individual should start the test seated in a chair with the arms crossed. For 30 seconds, the subject has to get up and sit as many times as possible. The result of this test is the number of times the individual completes this exercise.

MST: The two-minute step test is used to measure the aerobic endurance of the individuals. This test is simply a walking exercise in place during 2 minutes, in which the participant must raise the knee at a certain height. The number of correct steps with the right leg are recorded and are used as a measure.

All results of these tests are converted into a classification for each test, that combined result in a final classification for the physical domain in the 2nd screening. This classification procedure can be done, for each test, gender and age, based on Table 2.1, in which are recorded the values to classify an individual as normal. Otherwise the classification is always "functional decline".

Table 2.1: Range of values to classify an individual as normal in physical tests of 2nd screening

Test	Gender / Age range	60 - 64	65 - 69	70 - 74	75 - 79	80 - 84	85 - 89
TUGT	Men	≤ 5.6	≤ 5.9	≤ 6.2	≤ 7.2	≤ 7.6	≤ 8.9
	Women	≤ 6.0	≤ 6.4	≤ 7.1	≤ 7.4	≤ 8.7	≤ 9.6
CSRT	Men	≥ -6.35	≥ -7.62	≥ -7.62	≥ -10.16	≥ -13.97	≥ -13.97
	Women	≥ -1.27	≥ -1.27	≥ -2.54	≥ -3.81	≥ -5.08	≥ -6.35
CST	Men	≥ 14	≥ 12	≥ 12	≥ 11	≥ 10	≥ 8
	Women	≥ 12	≥ 11	≥ 10	≥ 10	≥ 9	≥ 8
MST	Men	≥ 87	≥ 86	≥ 80	≥ 73	≥ 71	≥ 59
	Women	≥ 75	≥ 73	≥ 68	≥ 68	≥ 60	≥ 55

Taking this into account, the physical classification in the 2nd screening is:

- if the individual is in functional decline for all 4 tests, classify the individual in "FRAIL" class;
- if the individual is normal for all 4 tests, classify the individual in "ROBUST" class;
- otherwise, classify the individual in "FUNCTIONAL DECLINE" class.

Cognitive Domain

The Quick Mild Cognitive Impairment (QMCI) screen is used to evaluate the participants of the 2nd screening according to their cognitive domain. This test is composed by 6 questions, each one assessing a different aspect of the cognitive domain and have different types of results. A summary of the questions and their results are shown below,

1. **Orientation** - to answer 5 questions which test spatial and temporal ideas of the individual in 1 minute (2-correct answer, 1-attempted but incorrect, 0-no attempt) maximum score=10;
2. **Word Registration** - to repeat 5 heard words in 30 seconds (1-per word repeated, in any order), maximum score=5;
3. **Clock Drawing** - to draw a clock, showing a specific time, in 1 minute (Give 1 mark for each number, 1 for each hand and 1 for the pivot correctly placed. Loose 1 mark for each number duplicated or greater than 12), maximum score=15;
4. **Delayed Recall** - to repeat the same words of question 2 in 1 minute (4-per word repeated, in any order), maximum score=20;
5. **Verbal Fluency** - to name as many animals as she/he can in 1 minute (0.5-per animal named until a maximum of 40), maximum score=20;
6. **Logical Memory** - to listen to a story and after that to tell as much of the story as she/he can in 30 seconds (2-per highlighted word in a table recalled), maximum score=30.

The score of QMCI, which varies between 0 and 100, is the sum of the 6 questions results and, based on this, the following classification rule for the cognitive domain in the 2nd screening can be constructed,

- if QMCI score < 50, classify the individual in "FRAIL" class;
- if $50 \leq$ QMCI score ≤ 70, classify the individual in "PRE-FRAIL" class;
- if QMCI score > 70, classify the individual in "ROBUST" class.

Nutritional Domain

In the nutritional domain of the 2nd screening the questions of the MNA test not used in the 1st screening are to be answered. With the aim of providing more detailed assessment and stratification of the individuals classified as pre-frail, the questions G-R are used, completing the original test. The results of the questions have again different scales. A list of the description of the questions and their possible results or codification (in brackets) is provided below,

g lives independently (1-Yes,0-No);

h takes more than 3 prescription drugs per day (0-Yes, 1-No);

i have pressure sores or skin ulcers (0-Yes, 1-No);

j number of subject's full meals daily (0-1 meal, 1-2 meals, 2-3 meals);

k1 eats at least one serving of dairy products (1-Yes, 0-No);

k2 eats two or more servings of legumes and eggs per week (1-Yes, 0-No);

k3 eats meat, fish or poultry every day (1-Yes, 0-No);

l consumes two or more servings of fruit or vegetables per day (1-Yes, 0-No);

m number of fluid consumed by the subject per day (0.0-less than 3 cups, 0.5-3 to 5 cups, 1.0-more than 5 cups);

n subject's mode of feeding (0-unable to eat without assistance, 1-self-fed with some difficulty, 2-self-fed without any problem);

o self view of nutritional status (0-views self as being malnourished, 1-is uncertain of nutritional state, 2-views self as having no nutritional problem);

p subject's perception of their health status compared with other people of the same age (0.0-not as good, 0.5-does not know, 1-as good, 2-better);

q mid-arm circumference of the subject in cm (0.0-MAC less than 21, 0.5-MAC 21 to 22, 1-MAC greater than 22);

r calf circumference of the subject in cm (0-CC less than 31, 1-CC 31 or greater).

The question K of the original MNA test is codified using its three subquestions (k1, k2, k3), as it is shown below

- if there are 0 or 1 ”yes”, K score=0.0;
- if there are 2 ”yes”, K score=0.5;
- if there are 3 ”yes”, K score=1.0.

As a result of this, the total score of MNA test is obtained through the sum of the MNA-SF score with all the results of the questions G-R. This score varies between 0 and 30 and it assigns to each individual a nutritional classification in the 2nd screening, based on the following classification rule

- if MNA score < 17, classify the individual in “FRAIL” class;
- if $17 \leq$ MNA score < 24, classify the individual in “PRE-FRAIL” class;
- if MNA score ≥ 24, classify the individual in “ROBUST” class.

Final Classification

Finally, to classify the participants of this 2nd screening as frail, pre-frail or robust, another classification rule was created, this time with the results of the classifications in the three evaluated domains. This rule is presented below,

- if the individual is classified as robust in the three domains, the final classification of the 2nd screening is “ROBUST”;
- if the individual is classified as frail in the three domains, the final classification of the 2nd screening is “FRAIL”;
- otherwise, the final classification of the 2nd screening is “PRE-FRAIL”.

In this chapter, all the points of the PERSSILAA project, needed to understand the assessment of the screening process developed in this project, were mentioned. More details about the PERSSILAA project and, in particular, about the screening process can be found in the site of the project (https://perssilaa.com/) and in the questionnaires attached to this report.

Chapter 3

Statistical Methods

During the development of this Master's project, some classic statistical methods for the treatment of multivariate data were used. This chapter is intended to summarize these methods, based on some basic statistical concepts and focusing in the main topics that help to understand the work done.

3.1 Regression models

When working with multivariate data it is often of interest to explore the relationships between the different variables involved. Some act as response variables and others as explanatory variables, that is, variables which are expected to be responsible for the variability observed in the response variables. In some cases and areas this study is very crucial, because it allows to describe and control a variable through other more accessible variables. The regression models have an important role in carrying out this type of data analysis and they currently are one of the most popular statistical methods to deal with this kind of questions.

In addition to describe the relation between a response (dependent) variable and one or more covariates (predictors, explanatory, or independent variables), the regression models can be used to predict response values when the values of the variables which are acting as predictors are known [12]. In the context of the present work, prediction does not play an interesting role; instead the regression was used to identify the most important variables in the description of a certain response variable in the case of the multiple regression; and in the case of multinomial and logistic regression was used, as well, for cross validation. For this reason, goodness of fit measurements were not of concern in this project and hence they are not addressed in this chapter.

Before entering into a more detailed description of specific regression models applied in this project, notation common to all models should be introduced. Let Y denote the response variable to be studied, $\mathbf{X} = (X_0, X_1, ..., X_p)$ denote a vector with a constant $X_0 = 1$ plus p covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ denote the regression coefficients (the parameters of the model), where β_0 is the intercept and the other parameters are the weights corresponding to the p covariates [10].

During this work only multiple linear regression and logistic regression (binomial and multinomial) were performed. These two types of regression are part of the generalized linear regression models group which are models that can be reduced to a weighted sum of the covariates after a transformation. Therefore, denoting $C(Y|\mathbf{x})$ as a property of the distribution of Y given the values of $\mathbf{X}$ are know, say $\mathbf{X} = \mathbf{x}$, these models include the relation [10]

$$C(Y|\mathbf{x}) = g(\mathbf{x}\boldsymbol{\beta}), \tag{3.1}$$

as well as the distribution of $Y|\mathbf{x}$.

Categorical covariates:

The introduction of categorical covariates in the model is performed differently from the quantitative covariates. Some of them are represented by numbers, but they have no numerical significance, so it is incorrect to treat them as quantitative variables. In this case, for each categorical covariate, a set of $k-1$ dummy variables have to be introduced, where k is the number of levels of the covariate, and a level of reference has to be chosen. The dummy variables are binary and produce a specific code for each level of the corresponding categorical variable.

For a better understanding of this method, an example of a variable that represents the color of the eyes (brown, blue and green) is shown. The resulting code from the dummy variables, with brown as the reference color, can be observed in the Table 3.1.

Table 3.1: Dummy coding for color of the eyes example

Color	Dblue	Dgreen
brown	0	0
blue	1	0
green	0	1

If the color of the eyes is a covariate in a regression model, then two variables will be introduced in the model, corresponding to its dummy variables.

Interpretation of the parameters:

The interpretation of the model parameters is normally a difficult task, because the property $C(Y|\mathbf{x})$, which describes the way $\mathbf{x}$ affects Y, is not necessarily linear in the parameters. But it is known that for the generalized linear models, which the logistic and multinomial models are examples, there is a function $h(u)$ which transforms the property $C(Y|\mathbf{x})$ into a linear function in the parameters, $h(C(Y|\mathbf{x})) = \mathbf{x}\boldsymbol{\beta}$, hence it is possible to do the interpretation of parameters on the transformed property instead of on the non linear property.

With this transformation, the regression coefficient β_j can be interpreted as the change in $h(C(Y|\mathbf{x}))$ per unit change in X_j, when the other covariates are constant [10],

$$\beta_j = h(C(Y|X_1, ..., X_j + 1, ..., X_p)) - h(C(Y|X_1, ..., X_j + 1, ..., X_p)). \tag{3.2}$$

For categorical covariates this interpretation is different, because they are represented by their dummy variables in the model. Considering the previous example of the color of the eyes, it is known that for this variable there must be two covariates in the model representing the dummy variables. Denote those covariates by X_j and X_{j+1}, corresponding to Dblue and Dgreen, respectively. Supposing that each observation corresponds to an individual and the model have more $p-2$ covariates, the model representations for the 3 possible cases are shown below.

Brown eyes: $h(C(Y|X_j = 0, X_{j+1} = 0)) = \beta_0 + ... + X_{j-1}\beta_{j-1} + X_{j+2}\beta_{j+2} + ... + X_p\beta_p$

Blue eyes: $h(C(Y|X_j = 1, X_{j+1} = 0)) = \beta_0 + ... + X_{j-1}\beta_{j-1} + \beta_j + X_{j+2}\beta_{j+2} + ... + X_p\beta_p$

Green eyes: $h(C(Y|X_j = 0, X_{j+1} = 1)) = \beta_0 + ... + X_{j-1}\beta_{j-1} + \beta_{j+1} + X_{j+2}\beta_{j+2} + ... + X_p\beta_p$

As it can be seen in the equations, the β_j coefficient indicates the difference on $h(C(Y|X))$ between an individual with blue eyes and other with brown eyes, when the values for all the other covariates are equal, while β_{j+1} represents the same difference but between an individual with green eyes and other with brown eyes [10].

3.1.1 Linear Regression

The most commonly applied regression model is the linear regression model, where the response variable must be continuous. When there is only one covariate related to Y the model is called simple linear regression, while if there is more than one covariate, it is given the name of multiple linear regression model.

In the linear regression the property of interest is the conditional mean, $E(Y|\mathbf{x})$, which is just equal to the weighted sum of the covariates,

$$C(Y|\mathbf{x}) = E(Y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}. \tag{3.3}$$

In this case it is assumed that the response variable can be described by the sum of the conditional mean with the amount of deviation from this,

$$Y = E(Y|\mathbf{x}) + \varepsilon = \mathbf{x}\boldsymbol{\beta} + \varepsilon. \tag{3.4}$$

The ε value represents the deviation referred before, often called error, and it is assumed to follow a normal distribution with mean zero and a constant variance σ^2. For this reason, in the linear regression, it is assumed that the response, Y, follows a normal distribution with mean $E(Y|\mathbf{x})$ and variance σ^2 [12]. Considering the linear regression applied to a set of n individuals, the model is represented, in matrix notation, by [16]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.5}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Regarding the interpretation of the parameters, this is much easier in the linear regression, because the property of interest is linear in $\boldsymbol{\beta}$. Therefore, for non categorical covariates, the coefficient β_j is simply the difference in the expected value of Y per unit change in X_j [10].

Estimation of the parameters:

To fit the model to the training set, the estimation of the unknown parameters has to be done, in this case only the estimation of $\boldsymbol{\beta}$. In linear regression is more frequent to do this with the least squares method, in which the values for $\boldsymbol{\beta}$ are chosen such that the sum of squared deviations between the observed values and the predicted values, $\sum_{i=1}^{n} (y_i - \mathbf{x_i}'\boldsymbol{\beta})^2$, where $\mathbf{x_i}' = \begin{bmatrix} 1 & x_{i,1} & \cdots & x_{i,p} \end{bmatrix}$, is minimized.

In order to find the $\boldsymbol{\beta}$ which minimizes this sum, it is necessary to obtain the derivatives of it with respect to β_j, for $j = 0, 1, ..., p$, and set the $p+1$ expressions equal to zero. The resulting estimator from this process is [16]

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.6}$$

Significance of the covariates:

A covariate is considered statistically significant for the model when the regression coefficient associated with it is significantly different from zero. In a model with categorical covariates, these only are not significant, if all the parameters corresponding to their dummy variables are not significantly different from zero.

Therefore, to test if a variable is statistically significant, it is necessary to test if its regression coefficients are different from zero, i.e., $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. It is known that under the null hypothesis and if the errors are normal,

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{VAR}(\hat{\beta}_j)}} \frown t_{n-k-1}, \tag{3.7}$$

where n is the number of observations in the training set and k is the number of covariates.

It has been already defined in equation (3.6) a possible estimator to $\boldsymbol{\beta}$, but it is also necessary, for the calculation of this statistic, to estimate the errors variance, since $VAR(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is the covariance matrix of $\boldsymbol{\beta}$. An unbiased estimator presented in the literature is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}})^2}{n-k-1}. \tag{3.8}$$

With all the estimates known, it is possible to use the statistic T_j in the hypothesis test to evaluate the importance of a covariate in the model [16].

3.1.2 Logistic Regression

In many medical, epidemiologic and social studies it is increasingly interesting to analyse dichotomous or polychotomous variables, and for these the linear regression is not an option. The use of the logistic regression in the dichotomous case is recurrent, because it uses the logistic distribution in the analysis, which is a flexible and easily function to use and to interpret.

As in the linear regression, also the logistic regression can be applied in case the model has one or more covariates (simple and multiple logistic regression). Since the case with multiple covariates is the more complex, but also the more general, it is the one which is explained in this work.

Assuming that the response Y_i for the individual i is binary with possible values 0 and 1 and these can be taken with probabilities π_i and $1 - \pi_i$, the variable Y_i is said to follow a Bernoulli distribution with parameter π_i, described as

$$P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \tag{3.9}$$

where $y_i = 0, 1$. In this case the expected value for Y_i is π_i and it is this that is intended to relate with the covariates in the logistic regression model. Since the probability of occurring $Y_i = y_i$ has to be in the interval [0,1], the relation between the expected value and the covariates is defined as

$$\pi_i(\mathbf{x}) = \frac{e^{\mathbf{x_i}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x_i}'\boldsymbol{\beta}}}. \tag{3.10}$$

The transformation necessary to make the equation (3.10) linear in $\mathbf{x}$ is called logit and is

defined below as a function of $\pi(\mathbf{x})$,

$$logit(\pi_i(\mathbf{x})) = \log\left(\frac{\pi_i(\mathbf{x})}{1-\pi_i(\mathbf{x})}\right). \tag{3.11}$$

This analysis can also be done for grouped data, in which there are some groups of individuals that have identical values for covariates and can be studied together. For that the model is slightly different, since the variable Y_i will represent a count for the group i and it will have a Binomial distribution (generalization of Bernoulli) [19]. Little attention will be given to this case here, because in the practical part of this work it was done only an individual approach of the logistic regression.

The interpretation of the regression coefficients in the logistic regression is a bit more complex than in linear regression, because the function of interest is not linear in $\mathbf{x}$. But, as it was referred to before, the interpretation can be done based on $logit(\pi(\mathbf{x}))$. In the case of non categorical covariates, the parameter β_j is interpreted as the change in log odds of $Y = 1|\mathbf{x}$ per unit change in X_j. This interpretation can also be done for the odds of $Y = 1|\mathbf{x}$, in this case a change of one unit in X_j causes an increment of e^{β_j} in the odds [9].

Estimation of the parameters:

The method used for the estimation of the parameters of the logistic regression model is normally the maximum likelihood. As the name implies, the method aims to find the values for $\boldsymbol{\beta}$ which maximize the likelihood function. In the logistic regression, the likelihood function is represented by the expression

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x_i})^{y_i}\left[1-\pi(\mathbf{x_i})\right]^{1-y_i} = \prod_{i=1}^{n}\left(\frac{e^{\mathbf{x_i}'\boldsymbol{\beta}}}{1+e^{\mathbf{x_i}'\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+e^{\mathbf{x_i}'\boldsymbol{\beta}}}\right)^{1-y_i}. \tag{3.12}$$

Similar to what was done in linear regression for the least squares method, it is necessary to obtain the derivatives of the log likelihood function with respect to β_j, for $j = 1, ..., p$, and set the resulting equations equal to zero. These set of equations do not have analytical solution and numerical methods for nonlinear system of equations have to be used [12].

Significance of the covariates:

As it was referred to for linear regression, to check the significance of a covariate X_j it is necessary to test if the parameter β_j is different from zero, i.e., to perform the hypothesis test $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. In this case the Wald test statistics can be used,

$$W_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{VAR}(\hat{\beta}_j)}}, \tag{3.13}$$

which follows approximately a standard normal distribution under the null hypothesis.

After estimating the parameters, the respective variances must be estimated to perform the tests. In the logistic regression the covariances matrix, which contains the variances and the covariances for all pairs of parameters, can be estimated by the inverse of the observed information matrix with dimensions $(p+1)\times(p+1)$, $VAR(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. This matrix includes the negative of the second partial derivatives terms of log likelihood function and is represented

bellow

$$\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{V}\mathbf{X}, \tag{3.14}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}, \mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}.$$

Multinomial Logistic Regression:

The multinomial logistic regression is a generalization of the model previously seen for a polychotomous response, i.e., a categorical variable with more than two categories. In this approach, to an individual i, the variable Y_i can take one value from a discrete set $1, 2, ..., J$ with probabilities π_{ij} and it is assumed that the J categories of Y_i are mutually exclusive and exhaustive, which means that $\sum_{j=1}^{J} \pi_{ij} = 1$.

Giving again attention only to the individual case, underlying this model there is a dummy variable Y_{ij} which takes the value 1 if the individual i belongs to the j_{th} category and 0 otherwise and due to this it has a multinomial distribution that can be described as

$$P(Y_i = y_i) = P[Y_{i1} = y_{i1}, ..., Y_{iJ} = y_{iJ}] = \binom{1}{y_{i1}, ..., y_{iJ}} \pi_{i1}^{y_{i1}} \cdots \pi_{iJ}^{y_{iJ}}. \tag{3.15}$$

All the analysis with the multinomial logistic regression can be done similarly to the one explained for the dichotomous case, but in what concerns the definition of the odds there is a significant difference. Since in this case there are more than two categories for the response variable, a reference category has to be chosen to define the logit function as

$$logit(\pi_{i1}) = \log\left(\frac{\pi_{i1}}{\pi_{iR}}\right), \tag{3.16}$$

where R represents the reference category [19].

3.2 Cluster analysis

The cluster analysis can be seen in two different perspectives: some authors describe it as the set of techniques that reduces the number of rows in a data matrix, which has one observation per row, combining them in groups of similar observations [13]; but most of the literature refers to it simply as an analysis to identify natural groups of objects in a data set [20] [14].

The resulting groups from clustering, called clusters, can also be interpreted as a classification assigned to each observation based on their similarities and dissimilarities with the others, without imposing pre-defined classes or other assumptions. In this procedure, it is intended that the clusters are formed so that objects in the same group are very similar and objects in distinct groups are not similar.

Because this analysis focuses on similarities and dissimilarities between observations and groups of them, it is necessary to properly define these measures, which will be done in the following subsections. Additionally, to perform a cluster analysis it is essential to choose one of two ways of doing it: hierarchically or not hierarchically. This topic will also be discussed later

in this chapter.

3.2.1 Proximity between observations

Proximity measures have an important role in creating the groups of observations in clustering. The objects in the same group are similar which implies having a high degree of similarity and a low dissimilarity degree. A proximity measure p_{rs}, between to objects r and s, is defined as a similarity or a dissimilarity measure if it satisfies the following assumptions [13].

Similarity:

1. $0 \leq p_{rs} \leq 1, \forall_{\mathbf{x_r}, \mathbf{x_s}}$;
2. $p_{rs} = 1 \Leftrightarrow \mathbf{x_r}$ and $\mathbf{x_s}$ are identical;
3. $p_{rs} = p_{sr}$.

Dissimilarity:

1. $p_{rs} \geq 0, \forall_{\mathbf{x_r}, \mathbf{x_s}}$;
2. $p_{rs} = 0 \Leftrightarrow \mathbf{x_r}$ and $\mathbf{x_s}$ are identical;
3. $p_{rs} = p_{sr}$.

For different types of variables and contexts there are multiple proximity measures. The data is represented by a $n \times p$ matrix, where in the columns are the variables $X_1, \ldots, X_p$ and the objects r and s are rows of the matrix defined by the observed vectors $\mathbf{x_r} = (x_{r1}, ..., x_{rp})$ and $\mathbf{x_s} = (x_{s1}, ..., x_{sp})$. The majority of similarity measures for binary data, some of them defined in Table 3.2 (a), are based on counts of concordant and discordant values between two objects, which can be observed in Table 3.2 (b).

Table 3.2: Similarity measures for binary

(a) Similarity measures

Coefficient name	Formula
Simple matching	$s_{rs} = \frac{a+d}{p}$
Double matching	$s_{rs} = \frac{2(a+d)}{2(a+d)+b+c}$
Jaccard	$s_{rs} = \frac{a}{a+b+c}$
Rogers-Tanimoto	$s_{rs} = \frac{a+d}{a+d+2(b+c)}$
Dice	$s_{rs} = \frac{2a}{2a+b+c}$

(b) Counts of values for individuals r and s

		Individual r		
	Value	1	0	Total
Individual s	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p = a+b+c+d$

Regarding categorical variables with more than two classes, the approach is similar to binary variables and some of the measures, seen before, can even be used, if it is designated by concordance the case where the two individuals have the same categories for the variable and discordance otherwise. But when the variables are ordinal, these measures must be used in an adapted version. In this case, different scores should be given to different degrees of proximity between the categories. An example of this can be the variable educational level with values 1-elementary school, 2-high school, 3-college, where it can be attributed the highest score (1)

when the individuals have the same education level, the following score (0.5) when the difference is only 1 unit and the lowest score (0) when one individual has an elementary school educational level and the other has a college degree.

The continuous data have also many coefficients defined for calculating the proximity between two objects, the most common are the dissimilarity measures because they are usually based on the differences between the values of variables. Some of the dissimilarity measures most mentioned in the literature are summarized in Table 3.3.

Table 3.3: Dissimilarity measures for continuous data

Distance name	Formula
Euclidean distance	$d_{rs} = \sqrt{\sum_{i=1}^{p} (x_{ri} - x_{si})^2}$
Standardized euclidean distance	$d_{rs} = \sqrt{\sum_{i=1}^{p} \left(\frac{1}{s_i^2}\right)(x_{ri} - x_{si})^2}$
Manhatan distance	$d_{rs} = \sum_{i=1}^{p} \lvert x_{ri} - x_{si} \rvert$
Mahalanobis distance	$d_{rs} = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}(x_{ri} - x_{si})\left(\frac{1}{s_{ij}}\right)(x_{rj} - x_{sj})}$
Minkowski distance	$d_{rs} = \left(\sum_{i=1}^{p} \lvert x_{ri} - x_{si} \rvert^{\lambda}\right)^{1/\lambda}$

Note: s_i^2 denotes the variance of variable with index i and s_{ij} denotes the covariance between the variables with index i and j.

In the context of this work, none of these measures is really useful, because the database which is the objective of this study has variables of different types. The most appropriate measure in this case is the Gower coefficient, which contains in a single value the global degree of similarity, from variables of different types, between two individuals. The Gower coefficient S_{rs} is defined by the following equation,

$$S_{rs} = \frac{\sum_{k=1}^{p} s_{rsk} w_{rsk}}{\sum_{k=1}^{p} w_{rsk}}, \tag{3.17}$$

where:

- $w_{rsk} = 0$ if the value of kth variable is missing for one or both of individuals r and s, $w_{rsk} = 1$ otherwise;
- for binary variables:
  - $w_{rsk} = 0$ if there is a negative match (kth variable are 0 for individual r and s);
  - $s_{rsk} = 1$ if the value of kth variable is 1 for both of individuals r and s and $s_{rsk} = 0$ otherwise;

- for categorical variables (more than two classes): $s_{rsk} = 1$ if the individuals r and s have concordance in the kth variable value and $s_{rsk} = 0$ if they have discordance;
- for quantitative variables: $s_{rsk} = 1 - |x_{rk} - x_{sk}|/R_k$, where R_k is the range of the values in the kth variable [8].

When the data set contains ordinal variables, it is possible to adapt this coefficient to take this into account. In this project it was necessary to define s_{rsk} for ordinal variables as

$$s_{rsk} = (N_k - |x_{rk} - x_{sk}|)/N_k, \tag{3.18}$$

where N_k is the number of categories of the kth variable.

These proximity measures were defined for a pair of observations, but they are normally used in cluster analyses of databases with n observations. When they are used for this purpose, in specific clustering methods, proximity matrices have to be constructed. Proximity matrices contain in the (i, j) entry the proximity measure between individuals i and j [13]. Due to the symmetry property and because the elements in the diagonal are either 1 (for similarity measures) or 0 (for dissimilarity measures), only an upper or lower triangular matrix needs to be computed.

3.2.2 Proximity between groups

In the previous subsection methods for measuring the proximity between observations were defined, but in some cases to perform a cluster analysis is also necessary to calculate proximity measures for two groups of objects (clusters). This subsection intends to do a summary of the most frequently used proximity measures between groups. The methods shown here are commonly used as a criterion to join clusters. For this purpose, it will be assumed that two clusters denoted by C_a and C_b have, respectively, n_a and n_b objects. The objects of C_a are denoted by $\mathbf{a_i} = (a_{i1}, ..., a_{ip}), i = 1, ..., n_a$, while the elements in C_b are defined as $\mathbf{b_i} = (b_{i1}, ..., b_{ip}), i = 1, ..., n_b$ [20]. For some measures it is also useful to consider the centroid for each cluster, which is defined as the mean vector of all the cluster elements. Thus, in this case, the centroids of C_a and C_b are defined as $\bar{\mathbf{a}} = (\bar{a_1}, ..., \bar{a_p})$ and $\bar{\mathbf{b}} = (\bar{b_1}, ..., \bar{b_p})$

Single Linkage: The single linkage method, also called nearest neighbor, defines the proximity between two clusters as the minimum dissimilarity or the maximum similarity between two objects, each of a different cluster. This can be described by the following equation,

$$d_{C_aC_b} = min\{d_{ab} \mid a \in C_a \text{ and } b \in C_b\}. \tag{3.19}$$

Complete Linkage: The complete linkage or furthest neighbor assigns the opposite definition of single linkage to the proximity between clusters, i.e., as the maximum dissimilarity between two objects of the two clusters. Similar to equation (3.19), complete linkage is defined by the formula below,

$$d_{C_aC_b} = max\{d_{ab} \mid a \in C_a \text{ and } b \in C_b\}. \tag{3.20}$$

Average Linkage: Contrary to single and complete linkage which only use one proxim-

ity measure between a pair of observations from two different clusters to define the proximity between the groups, the average linkage intends to take into account the similarities or dissimilarities of all pair of objects of the two clusters. It corresponds to the average of all the proximity measures between the pairs, as it is described in the following equation (in this case dissimilarities),

$$d_{C_aC_b} = \frac{\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} d_{a_ib_j}}{n_a n_b}. \tag{3.21}$$

Ward's Method: The Ward's method also involves all the observations, but it is based on the sum of squares criterion. Given a cluster C_a, the sum of squares of the deviations from the centroid can be calculated as

$$SSE_a = \sum_{i=1}^{n_a} (\mathbf{a_i} - \bar{\mathbf{a}})'(\mathbf{a_i} - \bar{\mathbf{a}}) = \sum_{i=1}^{n_a} \|\mathbf{a_i} - \bar{\mathbf{a}}\|^2. \tag{3.22}$$

It is possible to define the incremental increase in the sum of squares of the deviations from the new centroid resulting from joining two clusters C_a and C_b, creating a new cluster C_t, as

$$ISSE_{ab} = n_a\|\mathbf{a_i} - \bar{\mathbf{a}}\|^2 + n_b\|\mathbf{b_i} - \bar{\mathbf{b}}\|^2 = \left(\frac{n_a n_b}{n_a + n_b}\right)\|\bar{\mathbf{a}} - \bar{\mathbf{b}}\|^2 \tag{3.23}$$

In this case should be joined the clusters which have the lowest $ISSE$.

3.2.3 Hierarchical and Non-hierarchical clustering

The construction of a cluster analysis can be done in one of two ways: a hierarchical approach in which clusters will be joined or divided in each iteration and a non-hierarchical approach in which a set of k clusters are initially randomly formed and, in each iteration, the observations will be reallocated according to a criterion.

To initiate a hierarchical clustering a proximity matrix is calculated, based on the measures mentioned in subsection 3.2.1, for the n individuals of the study. In an agglomerative hierarchical method the process starts with n objects which can be as n single clusters and, based on the proximity matrix, the two objects with lowest dissimilarity or highest similarity will be joined. From the moment when clusters are created, the proximity measures have to be recalculated based on the techniques introduced in subsection 3.2.2. These steps are repeated until there is only one cluster formed by all the objects of the data set. On the other hand, the divisive hierarchical method starts with a single group, which contains all n objects, and this is divided in two clusters with the least similar objects. The process is repeated until n clusters are formed, one per observation.

In this approach there is not a preferential number of clusters, but the best number can be chosen based on the dissimilarities between the groups. In order to choose the number of clusters and to represent the results of these methods, a two-dimensional diagram called dendrogram can be drawn (example in Figure 3.1). This graphical representation contains in one axis the objects of the data set and in the other the proximity measures responsible for each merging or division [6].

In non-hierarchical clustering the number of clusters to be formed is predefined (k) and is

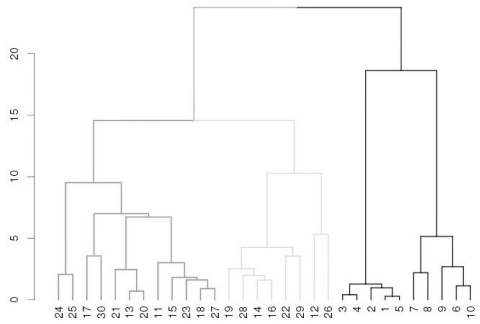


Figure 3.1: Dendrogram

precisely this the starting point of the method. A proximity matrix of the data is not necessary in this case, instead the data matrix with dimensions $n \times p$ is used. The algorithm can be defined by the following steps:

1. select k centroids of the n objects in data;
2. associate each object to the nearest centroid based on some dissimilarity measure;
3. recalculate the centroids of each cluster based on the objects which have been associated;
4. repeat step 2 and 3 until there are no more changes in the clusters or some stop criterion is reached.

The most popular non-hierarchical procedure is the K-means algorithm which calculates the centroids by averaging and usually uses the Euclidean distance as dissimilarity measure.

For the practical part of this work it was decided to use a hierarchical approach of the cluster analysis, since there are important categorical variables in the database which would be more conveniently treated with Gower coefficient than with Euclidean distance associated to K-means.

3.3 ROC curve

The receiver operating characteristic (ROC) curve is a statistical tool which is used to evaluate the performance of a test, whose result is measured in a continuous scale, as a binary classification test for a binary known status, while the threshold value for the classification varies. The ROC analysis is frequently used in medical studies to evaluate the accuracy (sensitivity and specificity) of diagnostic tests.

Generally it is assumed that higher values of the test, Y, suggest a positive status, which in medical cases is the disease state. But when the test follows the opposite idea, it is possible to do an inverse transformation changing the test to evaluate for $Z = -Y$ [18].

Before defining the ROC curve, it is important to introduce two measures of the performance of binary classification tests: sensitivity and specificity. The sensitivity or true positive fraction (TPF) is the proportion of positive cases which are correctly classified as such by the Y test. This can be more formally defined as

$$\text{Sensitivity} = \frac{\text{Positive State \& Positive Test}}{\text{Positive State}} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}. \tag{3.24}$$

Along the same lines, but with respect to negative results, the specificity or true negative fraction (TNF) is defined as the proportion of negative cases correctly classified by the test. This is

described by the following formula,

$$\text{Specificity} = \frac{\text{Negative State \& Negative Test}}{\text{Negative State}} = \frac{\text{True Negative}}{\text{True Negative + False Positive}}. \tag{3.25}$$

Assuming that c is a threshold in the continuous test Y, its binary response is defined as positive in the case that $Y \geq c$ and negative if $Y < c$. In this way, it is possible to describe the two measures, which are part of the construction of ROC curve, in function of c as

$$\begin{aligned} TPF(c) &= P[Y \geq c \mid D = 1] \text{ (Sensitivity)}, \\ FPF(c) &= P[Y \geq c \mid D = 0] \text{ (1-Specificity)}, \end{aligned} \tag{3.26}$$

where D is the binary indicator variable that represents disease in a medical case, but can represent the relevant state defined for the test as positive. This variable gives the true state of the individual or observation regardless of the test result.

The ROC curve is finally defined as the set of pairs of true and false positive fractions resulting from the binary response of Y with different possible thresholds (example in Figure 3.2). The next equation represents the formal definition of the ROC curve applied to a Y test with continuous results,

$$ROC(Y) = \{(FPF(c), TPF(c)), c \in]-\infty, +\infty[\}. \tag{3.27}$$

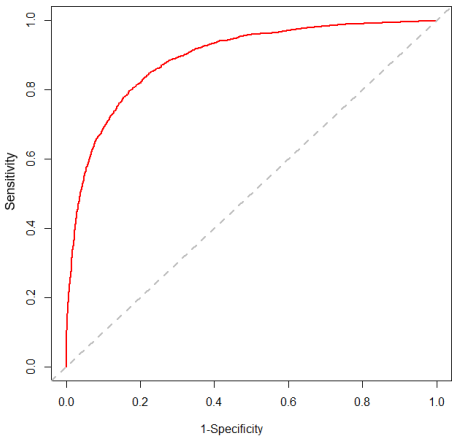


Figure 3.2: ROC curve

A test is totally uninformative if the probability distributions of Y are equal for the two populations of the study, i.e., if for all possible thresholds c it is verified that $TPF(c) = FPF(c)$. In what follows it is denoted by Y_D the diagnostic test variable for the population where the binary true status is positive and by $Y_{\bar{D}}$ for the population where it is negative. The best case for a test is when it can separate completely the two status or populations, that is when for some c of Y, $TPF(c) = 1$ and $FPF(c) = 0$. Taking this into account, a ROC curve is indicative of a better performance of the test the more above the bisecting of odd quarters in the graphic. Therefore, when comparing two tests based on their ROC curves, it is considered that test A is better than B if for any thresholds c_A and c_B which verify $FPF_A(c_A) = FPF_B(c_B)$ occurs $TPF_A(c_A) > TPF_B(c_B)$ [18].

Some measures can be used to characterize the ROC curve and to evaluate the test under study. The area under the ROC curve is the most used measure in this analysis and it is formally described by the following integral,

$$AUC = \int_0^1 ROC(t)\mathrm{d}t. \tag{3.28}$$

The area under the ROC curve (AUC) is a measure of how much above the ROC curve is in relation to the bisecting of odd quarters. A test which separates perfectly the two classes has

AUC=1, while an uninformative test has AUC=0.5. Then, a test is better the closer to 1 the AUC is and a test A is better than a test B if $AUC_A \geq AUC_B$.

For estimating the ROC curve there are three possible procedures:

- applying non-parametric empirical methods to obtain an empirical ROC curve;
- modeling the distributions of Y_D and $Y_{\bar{D}}$ and based on this estimate the ROC curve;
- modeling the ROC curve as a smooth parametric function.

The empirical estimation is very popular when the test results are continuous, which is the case in this work, because it does not require strong assumptions and it is very easy to apply. Several values of c are used, usually the ordered values in increasing order of the observed values of Y, and, for each c, values of the specificity and sensitivity are obtained. Then the definition of the ROC curve in equation (3.27) is applied to these data and the graphic with the resulting points is drawn. There are also several options regarding the choice of the optimal cut-off point, which is the threshold value of Y for future classification of the test as positive or negative. The most popular cut-off point is the one which maximizes the Youden's index, that is, sensitivity+specificity-1 of the resulting test.

3.4 Linear discriminant analysis

When analyzing a dataset with information about observations of several groups, it may be interesting to obtain functions of the associated variables which leads to a maximum separation among the groups. Normally this procedure is also used to create classification rules which can classify new vectors of observations in one of the known classes.

Some authors defend an independent study of these two methodologies, called as discrimination and classification, respectively. While others believe that it makes no sense to treat them separately, because their objetives, as the separation of the classes, tend to overlap.

A particular case in discriminant analysis is when the functions are linear and it is called linear discriminant analysis. In this section this topic is discussed in more detail, having been one of the methodologies used in the practice part of this project.

3.4.1 Case with two groups

Assuming the existence of two groups (populations or categories), G_1 and G_2, and an individual defined by p variables in a vector $\mathbf{x}$, let $f_1(\mathbf{x}|\boldsymbol{\theta_1})$ and $f_2(\mathbf{x}|\boldsymbol{\theta_2})$ denote the distributions of the random vector $\mathbf{X}$ when the individual belongs to G_1 or G_2, respectively. Therefore, it is possible to classify the observation $\mathbf{x}$ as belonging to G_1 if $f_1(\mathbf{x}|\boldsymbol{\theta_1}) > f_2(\mathbf{x}|\boldsymbol{\theta_2})$.

The classification rule is defined as:

- if $\lambda = \frac{f_1(\mathbf{x}|\boldsymbol{\theta_1})}{f_2(\mathbf{x}|\boldsymbol{\theta_2})} > 1$, $\mathbf{x}$ is classified as belonging to G_1,
- otherwise, $\mathbf{x}$ is classified as belonging to G_2.

Normal populations with known parameters:

Consider that $\mathbf{X} \sim N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ when it belongs to G_1, $\mathbf{X} \sim N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$ in case of belonging to G_2 and all the parameters are known. Then

$$\lambda = \exp\left[(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})\right] \tag{3.29}$$

and the discrimination rule is

- if $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) > 0$, $\mathbf{x}$ is classified as belonging to G_1,
- otherwise, $\mathbf{x}$ is classified as belonging to G_2.

The misclassification probabilities can be an important measure to evaluate the quality of a discrimination rule, with the knowledge that the smaller the probabilities are, the better the rule is. As a result of assuming a multivariate normal distribution for $\mathbf{X}$, with different means in distinct classes,

$$\mathbf{Y} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\mathbf{X} \sim N_p(E(\mathbf{Y}), VAR(\mathbf{Y})), \tag{3.30}$$

where

- $E(\mathbf{Y}) = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\boldsymbol{\mu_i} \qquad i = 1, 2,$
- $VAR(\mathbf{Y}) = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) = \delta^2.$

Based on this, it is possible to calculate the misclassification probabilities, which are shown below [17]

$$\begin{aligned} P_{21} &= P[\text{classify } \mathbf{x} \text{ as } G_2, \text{ knowing that } \mathbf{x} \text{ belongs to } G_1] = \Phi\bigg(-\frac{1}{2}\delta\bigg), \\ P_{12} &= P[\text{classify } \mathbf{x} \text{ as } G_1, \text{ knowing that } \mathbf{x} \text{ belongs to } G_2] = \Phi\bigg(-\frac{1}{2}\delta\bigg). \end{aligned} \tag{3.31}$$

Normal populations with unknown parameters:

When the populations are multivariate normal, but the parameters are unknown, other approach is used. Assuming that in the training set there are n_1 observations of G_1 category, n_2 from G_2 and let $(\mathbf{x}_1, ..., \mathbf{x}_{n_1}, \mathbf{x}_{n_1+1}, ..., \mathbf{x}_{n_1+n_2})$ be the vectors which compose the training set, then the parameters estimates are:

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_i \quad \text{and} \quad \bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_2}\mathbf{x}_i \tag{3.32}$$

for the mean vectors corresponding to each group and

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{3.33}$$

for the covariance matrix, where

$$S_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)' \qquad j = 1, 2. \tag{3.34}$$

The classification rule, called Anderson's classification rule, is then defined as

- if $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) > 0$, $\mathbf{x}$ is classified as belonging to G_1,
- otherwise, $\mathbf{x}$ is classified as belonging to G_2.

In this case, it is not possible to calculate exactly the misclassification probabilities, since $\mathbf{Y} = (\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2})' S^{-1}\mathbf{X}$ does not follow a normal distribution. A solution for this problem is to estimate the misclassification errors, which can be done using several methods available in the literature but not referred to in this work [17].

Fisher's linear discriminant function:

Fisher suggested a different approach do discriminate individuals from two different classes based on their multivariate observations $\mathbf{x}$. The concept was to create univariate observations y which would be linear combinations of the elements of $\mathbf{x}$. In this approach is not assumed a normal distribution for the two populations, but a pooled estimate of the covariance matrices is used [15].

The goal is to calculate $Y = \mathbf{a}'\mathbf{x}$ so that the separation between the groups is maximum. For this purpose, Fisher proposed that $\mathbf{a}$ should be choosen so that the ratio of the variance between the groups and the variance within groups is maximum. This ratio is defined in the following equation,

$$\frac{[\mathbf{a}'(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})]^2}{\mathbf{a}'\Sigma\mathbf{a}} = \frac{\left[\sum_{i=1}^{p} a_i(\mu_{1i} - \mu_{2i})\right]^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sigma_{ij}}, \tag{3.35}$$

where $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the mean vectors of the two populations and Σ is the common covariance matrix. It can be shown that the vector $\mathbf{a}$ which maximizes this ratio is $\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ and the function

$$L(\mathbf{x}) = \mathbf{a}'\mathbf{x} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\mathbf{x} \tag{3.36}$$

is called Fisher's linear discriminant function.

Based on the Fisher's function, it was created a new classification rule without assuming any distribution for the two groups,

- if $L(\mathbf{x}) > k$, $\mathbf{x}$ is classified as belonging to G_1,
- otherwise, $\mathbf{x}$ is classified as belonging to G_2,

where $k = \frac{1}{2}\left[(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})\right]$ is the midpoint between $L(\boldsymbol{\mu_1})$ e $L(\boldsymbol{\mu_2})$.

Obviously, to apply this rule when the parameters of the distribution are unknown, the parameters $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ and Σ are substituted by the respective estimates. This results in the previous rule for normal populations with unknown parameters.

3.4.2 Case with multiple groups

Extending this analysis to the case where there are k groups $(G_1, G_2, ..., G_k)$, the objective is to classify a vector $\mathbf{x}$ in one of the k groups. Let $P(G_1), P(G_2), ..., P(G_k)$ denote the prior probabilities of an individual to belong to each of the groups and let $f_i(\mathbf{x})$ represent the distribution of X in each group. In addition, let $c_{i/j}$ denote the cost of classifying $\mathbf{x}$ in G_i group, when it belongs to G_j.

Using Bayes' theorem the posterior probability membership for each group is

$$P(G_i|\mathbf{x}) = \frac{P(G_i)f_i(\mathbf{x})}{\sum_{j=1}^{k} f_j(\mathbf{x})P(G_j)} \tag{3.37}$$

and, from that, to set the expected cost of classifying $\mathbf{x}$ in G_i as

$$\pi(G_i) = \sum_{\substack{j=1 \\ j\neq i}}^{k} c_{i/j}P(G_j|\mathbf{x}). \tag{3.38}$$

The classification rule, which results from this, is then:

- if $\pi(G_i) \leq \pi(G_j), \forall j = 1, ..., k$, $\mathbf{x}$ is classified as belonging to G_i.

This rule can be applied to any distribution of $\mathbf{X}$ according to the class, but these distributions have to be specified.

Normal populations with unknown parameters:
In the case when the responses of the groups are described by multinormal random variables with the same covariance matrices, the classification rule is similar to those defined in the case with only two groups. The linear discriminant scores W_{ij}, used to classify $\mathbf{x}$, are described by

$$W_{ij} = \mathbf{x}'\mathbf{S}(\bar{\mathbf{x}_i} - \bar{\mathbf{x}_j}) - \frac{1}{2}(\bar{\mathbf{x}_i} + \bar{\mathbf{x}_j})'\mathbf{S}(\bar{\mathbf{x}_i} - \bar{\mathbf{x}_j}). \tag{3.39}$$

The discriminant rule is then:

- if $W_{ij} > 0, \forall j \neq i$, $\mathbf{x}$ is classified as belonging to G_i.

3.5 Principal component analysis of mixed data

The principal component analysis (PCA) of mixed data is a relatively new approach developed by researchers from various institutes of Bordeaux, culminating in the creation of a new package for the R software and an article published in December, 2014 [1]. This method intends to perform a PCA for numerical and categorical data simultaneously based on the generalized singular value decompostition (GSVD). In addition, it allows to introduce weights to rows and columns of the studied data matrix. The multiple correspondence analysis is also used in this process.

3.5.1 GSVD

GSVD is a matrix decomposition method which intends to decompose a matrix $\mathbf{Z}$ of dimension $n \times p$ based on two positive square matrices $\mathbf{N}$ and $\mathbf{M}$ of dimensions $n \times n$ and $p \times p$, respectively. The decomposition of $\mathbf{Z}$ provided by GSVD is then

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}', \tag{3.40}$$

where:

- $\mathbf{\Lambda} = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$ is the diagonal matrix of dimension $r \times r$ which contains the singular values of $\mathbf{ZMZ'N}$ and $\mathbf{Z'NZM}$, where $r = \text{rank}(Z)$;
- $\mathbf{U}$ is the matrix of dimension $n \times r$ which contains the first r eigenvectors of $\mathbf{ZMZ'N}$ such that $\mathbf{U'NU} = \mathbb{I}_r$;
- $\mathbf{V}$ is the matrix of dimension $p \times r$ which contains the first r eigenvectors of $\mathbf{Z'NZM}$ such that $\mathbf{V'MV} = \mathbb{I}_r$.

3.5.2 PCA with metrics

As mentioned above, GSVD allows to introduce weights to rows and columns of $\mathbf{Z}$ in PCA using $\mathbf{N}$ and $\mathbf{M}$, which are the diagonal matrices of those weights. This decomposition is done with the aim of calculating the factor scores of the rows and columns from the data matrix.

The $n \times r$ matrix $\mathbf{F}$ contains, by definition, the factor scores of the rows. This scores are the coordinates of the orthogonal projections of the $\mathbf{Z}$ rows weighted by $\mathbf{M}$ onto the axes resulting from the $\mathbf{V}$ columns, i.e., the columns of $\mathbf{F}$ are the principal components of $\mathbf{Z}$. For this reason $\mathbf{F}$ is defined by the following equation,

$$\mathbf{F} = \mathbf{ZMV} \tag{3.41}$$

and by equation (3.40) it can be also defined as

$$\mathbf{F} = \mathbf{U\Lambda}. \tag{3.42}$$

To define the factor scores of the columns the process is similar, using the matrices $\mathbf{N}$ and $\mathbf{U}$ instead of $\mathbf{M}$ and $\mathbf{V}$. If $\mathbf{A}$ denotes the factor scores matrix of the columns of dimension $p \times r$, then it contains the loadings of $\mathbf{Z}$. The loadings are the coordinates of the orthogonal projections of the $\mathbf{Z}$ columns weighted by $\mathbf{N}$ onto the axes resulting from the $\mathbf{U}$ columns. Hence $\mathbf{A}$ can be obtained from the next equation

$$\mathbf{A} = \mathbf{Z'NU} \tag{3.43}$$

which, again by equation (3.40), can be also defined as

$$\mathbf{A} = \mathbf{V\Lambda}. \tag{3.44}$$

3.5.3 Standard PCA and MCA

Now it is necessary to understand how the theory of GSVD is applied in the standard PCA and MCA, defining for each case the matrices $\mathbf{Z}$, $\mathbf{M}$ and $\mathbf{N}$.

Standard PCA:

In PCA the studied data matrix $\mathbf{X}$ of dimension $n \times p$ is the set of n observations described by p numerical variables. Since it is possible that the variables are not all described by the same unit or have very different variances it is important to standardize the data. Thus the first step in a standard PCA procedure is the pre-processing, where the $n \times p$ matrix $\mathbf{Z}$ resulting from a standardization of $\mathbf{X}$ is computed.

Regarding $\mathbf{N}$ and $\mathbf{M}$ matrices they are defined as

$$\mathbf{N} = \frac{1}{n}\mathbb{I}_n \quad \text{and} \quad \mathbf{M} = \mathbb{I}_p, \tag{3.45}$$

since in PCA the observations are weighted by $\frac{1}{n}$ and the variables are weighted simply by 1.

Based on this, the factor scores matrices $\mathbf{F}$ and $\mathbf{A}$ can be calculated using the equations (3.42) and (3.44), respectively.

Some important properties of PCA can be described using these matrices, as the authors Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J. mention in [1]. These are presented below.

- a_{ji} is the linear correlation between $\mathbf{x}_j$ and the principal component $\mathbf{f}_i$:

$$a_{ji} = \mathbf{z}_j'\mathbf{N}\mathbf{u}_i = r(\mathbf{x}_j, \mathbf{f}_i); \tag{3.46}$$

- λ_i is the variance of the principal component $\mathbf{f}_i$:

$$\lambda_i = \|\mathbf{f}_i\|^2_{\mathbf{N}} = VAR(\mathbf{f}_i); \tag{3.47}$$

- λ_i is also the sum of the squared correlations between the numerical variables $\mathbf{x}_j$, where $j = 1, ..., p$, and the principal component $\mathbf{f}_i$:

$$\lambda_i = \|\mathbf{a}_i\|^2_{\mathbf{M}} = \sum_{j=1}^{p} r^2(\mathbf{x}_j, \mathbf{f}_i). \tag{3.48}$$

Standard MCA:

In MCA the $n \times p$ matrix of data contains n observations, as in PCA, described by p categorical variables and it can be denoted by $\mathbf{X}$. Denote the levels of the categorical variables by m_j, where $j = 1, ..., p$, and the total number of levels by m. In this case the pre-processing step is a bit more complicated. First it is necessary to create a $n \times m$ matrix $\mathbf{G}$, where each level of $\mathbf{X}$ is coded as a binary variable. After that the $\mathbf{Z}$ matrix centering the elements in $\mathbf{G}$ is built.

In this case, the weights for the observations are the same as in PCA, but the weights for the levels are $\frac{n}{n_s}$, where n_s denotes the number of observations which belong to level s. Therefore, the matrices $\mathbf{N}$ and $\mathbf{M}$ are defined as

$$\mathbf{N} = \frac{1}{n}\mathbb{I}_n \quad \text{and} \quad \mathbf{M} = diag\left(\frac{n}{n_s}, s = 1, ..., m\right). \tag{3.49}$$

Again using GSVD of $\mathbf{Z}$ with the matrices previously defined, in particular the equation (3.42), it is possible to calculate the matrix $\mathbf{F}$. However, the matrix of the factor scores of the levels can not be calculated as in PCA. In the MCA case, the matrix $\mathbf{A}^*$ is described by

$$\mathbf{A}^* = \mathbf{M}\mathbf{V}\mathbf{\Lambda}. \tag{3.50}$$

MCA has some properties that are mentioned below [1].

- a^*_{si} is the mean value of the (normalized) factor scores of the observations that belong to level s:

$$a^*_{si} = \frac{n}{n_s} a_{ij} = \frac{n}{n_s} \mathbf{z}_s^t \mathbf{N} \mathbf{u}_i; \tag{3.51}$$

- λ_i is the sum of the correlation ratios between the categorical variables $\mathbf{x}_j$ and the principal component $\mathbf{f}_i$:

$$\lambda_i = \|\mathbf{a}_i\|^2_{\mathbf{M}} = \|\mathbf{a}^*_i\|^2_{\mathbf{M}^{-1}} = \sum_{j=1}^{p} \eta^2(\mathbf{f}_i | x_j), \tag{3.52}$$

(The correlation ratio $\eta^2(\mathbf{f}_i|x_j)$ measures the part of the variance of $\mathbf{f}_i$ explained by the variable j).

3.5.4 PCA of mixed data

The method propose by Chavent et al. [1] to perform a PCA for mixed data is based on standard PCA and MCA and uses the GSVD approach. The data set to be analyzed with this method contains n observations described by p_1 numerical variables and p_2 categorical variables. This information is represented in a $n \times p_1 + p_2$ matrix $\mathbf{X}$ which can be divided in two matrices with variables of different types. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the matrices of the numerical and categorical data with dimensions of $n \times p_1$ and $n \times p_2$, respectively. Let also m denote the total number of levels of the categorical variables. This method is done in two main steps which are described below.

1st step:

In the first step the pre-processing phase is carried out and its main objective is to calculate the matrices $\mathbf{Z}$, $\mathbf{N}$ and $\mathbf{M}$. The $n \times (p_1 + m)$ matrix $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ consists of the two matrices built on the pre-processing steps of standard PCA and MCA. Thus, $\mathbf{Z}_1$ results from the standardization of $\mathbf{X}_1$ and $\mathbf{Z}_2$ is built centering the indicator matrix $\mathbf{G}$ of $\mathbf{X}_2$.

The matrix $\mathbf{N}$ contains the weights of the rows of $\mathbf{Z}$, which were equally defined for the two methods (PCA and MCA) as $\frac{1}{n}$. For this reason, in this mixed method $\mathbf{N} = \frac{1}{n}\mathbb{I}$. On the other hand, the matrix $\mathbf{M}$ contains different weights on its diagonal. The first p_1 entries of the diagonal are equal to 1 which is the weight in PCA, while the last m entries are equal to $\frac{n}{n_s}$ as in MCA.

2nd step:

The second step of this approach is called the factor scores processing step. The objective in this phase is to build the matrices $\mathbf{F}$ and $\mathbf{A}^*$ of the factor scores, which is the main result of this principal component analysis. Using matrices $\mathbf{N}$ and $\mathbf{M}$, as they are defined in the first step, the decomposition of $\mathbf{Z}$ by GSVD is done, as defined in subsection 3.5.1: $\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$.

The factor scores matrices are calculated using the results of this decomposition and the following equations, which were already mentioned in this section,

$$\mathbf{F} = \mathbf{Z}\mathbf{M}\mathbf{V}, \quad \mathbf{A} = \mathbf{Z}'\mathbf{N}\mathbf{U}. \tag{3.53}$$

Note that the matrix $\mathbf{A}^*$ can be divided in two different matrices as follows, $\mathbf{A}^* = \begin{bmatrix} \mathbf{A}^*_1 \\ \mathbf{A}^*_2 \end{bmatrix}$, where $\mathbf{A}^*_1$ contains the factor scores of the p_1 numerical variables, while $\mathbf{A}^*_2$ contains the factor scores

corresponding to the m levels.

When the PCA of mixed data is applied to a purely numerical or categorical data, the method behaves as a standard PCA or MCA, respectively.

3.6 Some hypotheses tests

Hypothesis testing is a statistical method commonly used to infer from a sample if a statement about a population is true or false. This statement is the hypothesis of the test. In this report, the basic knowledge about hypothesis testing, which can be seen in [5], is assumed. This section intends to explain and contextualize different tests of hypotheses performed in this work.

3.6.1 Chi-squared test for homogeneity

Assuming that the data under analyses is from r populations and each observation is classified into one of c classes, it is possible to construct a $r \times c$ contingency table, as Table 3.4, with r rows and c columns. This table contains the number of observations from the ith population that are classified as j, denoted by O_{ij}. Let n_i denote the number of observations from each population, thus

$$n_i = O_{i1} + O_{i2} + ... + O_{ic}, \quad \text{for} \quad i = 1, ..., r. \tag{3.54}$$

The total number of observations in the jth class from all populations is denoted by C_j,

$$C_j = O_{1j} + O_{2j} + ... + O_{rj}, \quad \text{for} \quad j = 1, ..., c. \tag{3.55}$$

Finally, let N denote the total number of observations of all populations, thus

$$N = n_1 + n_2 + ... + n_r. \tag{3.56}$$

Table 3.4: $r \times c$ contingency table

	Class 1	Class 2	...	Class c	Totals
Population 1	O_{11}	O_{12}	...	O_{1c}	n_1
Population 2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
Population r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Totals	C_1	C_2	...	C_c	N

In order to apply the Chi-squared test to evaluate the homogeneity of the distributions from the different populations, some assumptions have to be satisfied. The samples from each population have to be independent and identically distributed, the populations have to be stochastically independent and it is also crucial that each observation can be classified into one and only one of the c classes [3].

This test intends to infer if the probabilities of a random observation from each population classified in the jth class are equal to each other, for all the classes. Therefore, denoting by p_{ij} the probability of an observation from the ith population be classified in the jth class, the hypotheses to be tested are:

H_0: All the probabilities corresponding to the same class are equal to each other ($p_{1j} = p_{2j} = ... = p_{rj}$, for all j)
H_1: At least two probabilities corresponding to the same class are not equal to each other ($p_{ij} \neq p_{kj}$, for some j and some pair i and k).

For the calculation of the test statistics, it is necessary to define E_{ij} as the expected number of observations from the population i classified in jth class, if the null hypothesis is true. If H_0 is true, the maximum likelihood estimates for p_{ij} are equal to C_j/N and hence E_{ij} can be estimated by $e_{ij} = n_i C_j/N$. Taking this into account, the test statistic T is defined by,

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}. \tag{3.57}$$

The Chi-squared test for homogeneity is performed based on an approximation of the distribution of T when the null hypothesis is true, since the exact distribution of T is difficult to calculate. Thus, under the null hypothesis, T has approximately a Chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. The quantiles of the Chi-squared distribution are tabulated [3]. Considering this approximation, the critical region of size α for this test is $R = \{t : t > x_{1-\alpha}\}$, where $x_{1-\alpha}$ is the $1-\alpha$ quantile of a Chi-squared distribution with $(r-1)(c-1)$ degrees of freedom and t is any observed value for T. Hence, the null hypothesis is rejected when T is above $x_{1-\alpha}$ or alternatively when

$$\begin{aligned} \text{p-value} &= P[T > t_{obs} | H_0 \text{ true}] \\ &= P[\chi^2_{(r-1)(c-1)} > t_{obs}] \\ &\leq \alpha, \end{aligned} \tag{3.58}$$

where α is the level of significance considered in the test and t_{obs} is the observed value of the test statistic.

3.6.2 Mann-Whitney-Wilcoxon test

The Mann-Whitney-Wilcoxon test can be applied to two random samples from two different populations, with independence within and between the two samples and outcomes measured in an ordinal scale. As defined in [4], let $(X_1, X_2, ..., X_n)$ and $(Y_1, Y_2, ..., Y_m)$ denote the random samples of size n and m from populations 1 and 2, respectively. For the execution of the test, the total sample has to be sorted in ascending order and the ranks 1 to $n+m$ must be assigned to the observations. Let $R(X_i)$ and $R(Y_j)$ denote the rank assigned to the correspondent variable, X_i or Y_j, for all values of i and j and N denote the total number of observations, $N = n + m$. A value which is repeated in the samples is called a tie and the rank assigned to repeated values is the mean of the ranks that would have been attributed to them if there had been no ties.

This test can be applied to 3 different alternative hypotheses, all related to the comparison between the distributions of the two populations, represented by the random variables X and Y for populations 1 and 2, respectively. The null hypothesis is always the equality of the two distributions. Let $F(x)$ and $G(x)$ denote the distribution function of X and Y, respectively, the hypotheses are defined as:

Two-Tailed Test:
$H_0 : F(x) = G(x)$, for all x
$H_1 : F(x) \neq G(x)$, for some x.

Lower-Tailed Test:
$H_0 : F(x) = G(x)$, for all x
$H_1 : F(x) > G(x)$, for some x.

Upper-Tailed Test:
$H_0 : F(x) = G(x)$, for all x
$H_1 : F(x) < G(x)$, for some x.

The test statistics is equal for the three tests, but varies with the number of ties in the samples. If there are a small number of ties or even none, the test statistics is simply the sum of the ranks corresponding to population 1, i.e.,

$$T = \sum_{i=1}^{n} R(X_i). \tag{3.59}$$

If there are many ties, the test statistics used is the standardization of T, defined as

$$T^* = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}, \tag{3.60}$$

where $\sum R_i^2$ is the sum of the squares of all the ranks (or average ranks) in both samples.

Under the null hypothesis and for $n \leq 20$ and $m \leq 20$, the quantiles of T are tabulated [4], while for upper values of n or m the quantiles are obtained recursively from

$$w_p = n(n + m + 1) - w_{1-p}. \tag{3.61}$$

A normal approximation can also be used in the case of no ties and n or m greater than 20, obtaining the approximate quantiles given by

$$w_p \cong \frac{n(N+1)}{2} + z_p\sqrt{\frac{nm(N+1)}{12}}, \tag{3.62}$$

where z_p is the pth quantile of the standard normal distribution, which is also tabulated. In the case of samples with many ties, the test statistics T^* is approximately a standard normal random variable assuming that H_0 is true, hence the quantiles are the same as above.

In the two-tailed test, the null hypothesis is rejected at level of significance α if the test statistic, T or T^*, is lower than the correspondent $\alpha/2$ quantile or greater than the $1 - \alpha/2$ quantile. Regarding the lower-tailed test, H_0 is rejected at the same level of significance, if T or T^* is lower than the respective α quantile. Finally, in the upper-tailed test, H_0 is rejected, if T or T^* is upper than the $1 - \alpha$ quantile [4].

The respective p-value for each of the three possible tests, letting t_{obs} denote the observed value of the test statistic, is

- two-tailed test: p-value $= 2\min(P[T > t_{obs}|H_0 \text{ true}], P[T < t_{obs}|H_0 \text{ true}])$ or p-value $= 2\min(P[T^* > t^*_{obs}|H_0 \text{ true}], P[T^* < t^*_{obs}|H_0 \text{ true}])$;
- lower-tailed test: p-value $= P[T > t_{obs}|H_0 \text{ true}]$ or p-value $= P[T^* > t^*_{obs}|H_0 \text{ true}]$;
- upper-tailed test: p-value $= P[T < t_{obs}|H_0 \text{ true}]$ or p-value $= P[T^* < t^*_{obs}|H_0 \text{ true}]$.

3.6.3 Kruskal-Wallis test

As an extension of the Mann-Whitney-Wilcoxon test, a test for $k > 2$ independent samples was developed by Kruskal and Wallis. Therefore, in this test, the hypotheses are inferred based on k random samples denoted by $X_{i1}, X_{i2}, ..., X_{in_i}$ for $i = 1, ..., k$ and where n_i represents the size of the ith random sample. Again the total number of observations is denoted by N and $R(X_{ij})$ represents the rank assigned to X_{ij} taking into account all observations. The ranks are assigned in the same way, for equal observations, as in the last test. Also the assumptions made for the test are equal to those from Mann-Whitney-Wilcoxon test [4].

This test evaluates again the homogeneity between the different populations' distributions. Thus, the hypotheses of the test are the following,

H_0: All of the k population distribution functions are identical

H_1: At least one of the populations tends to yield larger observations than at least one of the other populations.

In the Kruskal-Wallis test, the test statistics used is defined as

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right), \tag{3.63}$$

where R_i is the sum of all ranks assigned to the ith sample and

$$S^2 = \frac{1}{N-1}\left(\sum_{\text{all ranks}} R(X_{ij})^2 - N\frac{(N+1)^2}{4}\right) \tag{3.64}$$

These calculations are made in the general case, but if there are no ties the test statistic can be simplified as

$$T = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1). \tag{3.65}$$

The exact null distribution of T is tabulated only for $k = 3$ and $n_i \leq 5$ [4], but in this case is better to work with the approximation of chi-squared distribution with $k-1$ degrees of freedom. Considering this, the null hypothesis is rejected in the test at the level α, if T is greater than its $1-\alpha$ quantile under the null hypothesis. Alternativelly, this decision can be made with the p-value, which in this case is defined as p-value $= P[T < t_{obs}|H_0 \text{ true}] \approx P[\chi^2_{(k-1)} < t_{obs}|H_0 \text{ true}]$, where t_{obs} is the observed value of the test statistic.

3.6.4 Multiple comparisons

The multiple comparisons problem occurs when a set of statistical inferences is simultaneously considered or, in the specific case of hypothesis testing, a set of hypotheses tests is applied to a group of observations. This situation can be a problem, if the type I error of the tests are not carefully chosen. This is because, as the number of independent null hypotheses tested increases, the chance that at least one of those true null hypotheses will be incorrectly rejected also increases and, consequently, many false positives will be produced.

Supposing that it is intended to test m null hypotheses $H_{0i}, i = 1, ..., m$, if each of this hypotheses uses the significance level of α, all the null hypotheses are true and all the tests are independent, then

$$\begin{aligned} P[\text{at least one } H_{0i} \text{ is rejected}] &= 1 - P[\text{none } H_{0i} \text{ is rejected}] \\ &= 1 - (1-\alpha)^m. \end{aligned} \tag{3.66}$$

The probability mentioned in equation (3.66) is called the family-wise error rate (FWER).

Other measure important for this problem is the false discovery rate (FDR), which is defined as the expected proportion of false positives among all the rejected hypotheses,

$$FDR = E\left[\frac{\text{Number of rejected true null hypotheses}}{\text{Number of rejected null hypotheses}}\right]. \tag{3.67}$$

In order to avoid infering about certain issues with a high percentage of error, some methods can be used to control the FWER or the FDR [21]. When controlling the FWER, the error rate is fixed and the rejection area have to be estimated, while when controlling the FDR the opposite happens.

The statistical method used, in the practical part of this work to adjust the p-values, was the Holm's method (1979) [11], which belongs to the group of methods that control the FWER. This method is a sequentially rejective version of the simple Bonferroni procedure, which is much less conservative and it is defined by the following steps:

1. Compute the unadjusted p-values for each of the m tests;
2. Sort all p-values in ascending order ($p_1 \leq p_2 \leq ... \leq p_m$);
3. Compare each p_i with $\alpha/(m-i+1)$ or $(m-i+1)p_i$ with α:
   - if $(m-i+1)p_i \leq \alpha$ and $(m-j+1)p_j \leq \alpha, \forall j < i$, H_{0i} is rejected;
   - otherwise, H_{0i} and the hypotheses $H_{0j}, \forall j > i$ are not rejected and the procedure stops.

The adjusted p-value for H_{0i} is the smallest significance level at which it is rejected. This can be the value $r_i = (m-i+1)p_i$ or if the unadjusted p_i and p_{i-1} are the same, it is logical that the adjusted p-value for H_{0i} is r_{i-1}.

All the methodologies used in the development of this Master's project to make an evaluation of the quality of the screening process classification were properly introduced in this chapter. It is assumed that after reading it, all the concepts necessary to understand the scientific contents of this report.

Chapter 4

Database validation

Before doing a deeper analysis of the project's database it was necessary to validate it based on PERSSILAA's protocol. For this reason, the validity of the values for some variables was verified and all scores and status variables were recalculated by the rules previously described.

During this process, some problems were found in the database which were reported to the PERSSILAA's team responsible for the creation and maintenance of the database. In this chapter the problems found on the database downloaded on 09-03-2016 at 15:55:27 are described. The score or status variables that did not present any problem in this study are not mentioned in this chapter.

4.1 ID_USER variable

The ID_USER variable should uniquely identify the subjects that participated in PERSSILAA's screening. It is supposed that a value from this variable may only appear in one line of the database, unless the correspondent subjects have participated in the two rounds of 1st screening (SURVEY=0 and SURVEY=1).

In this database 98 cases were found with two lines identified by the same ID_USER value (corresponding to 196 lines).

Table 4.1: Subjects with repeated ID_USER and different gender

ID_USER	SURVEY	T1_ALG_01
14025	1	1
14025	0	2
46177	1	1
46177	2	2
46241	1	2
46241	2	1
47670	1	1
47670	2	2
48058	1	1
48058	2	2
48387	1	1
48387	2	2
48695	1	2
48695	2	1
49151	1	2
49151	2	1

Of the 98 cases, 8 had different genders in their repetition, 13 had different dates of birth and 1 had the same value for variable SURVEY (in the column Notes_1 they have the comment FALSE REPETITION). In the group with different values for SURVEY variable, there were 88 cases with values 0 and 1 and 9 cases with values 1 and 2.

From the 8 cases which had different genders, only 1 could be a subject with 2 valid entries from 1st and 2nd round of the 1st screening (marked on grey in Table 4.1), because only the values 0 and 1 are valid for SURVEY variable. The other 7 cases had SURVEY values 1 and 2 (invalid value for SURVEY). All cases can be observed in Table 4.1.

The 2 lines found with repeated ID_USER (32861) and the same value for SURVEY variable (1), also had equal values for all other variables. One of these lines might have been improperly inserted in the database.

Of the 13 subjects that had 2 entries with different date of birth, 3 had the month and the day values changed (marked on grey in Table 4.2). As these 6 entries had all of the necessary questions to the 1st screening's classification filled, they can be

considered as valid repetitions with an error in the T1_ALG_02 variable (date of birth).

In the group of subjects that had repeated ID_USER, there were 88 with SURVEY=0 and 99 with SURVEY=1. To verify how many of them completed the 1st or/and 2nd screening, we had to treat these cases separately since the questionnaire applied in the 1st round of the 1st screening (SURVEY=0) did not contain the AD8 dementia screening test questions.

Among the 88 subjects that participated in the 1st round of 1st screening (SURVEY=0), none had the 1st and 2nd screening completely filled in, 12 of them had all the questions of 1st screening filled in (in the column Notes_2 they have the comment "Complete 1st screening") and 3 have a completely filled in 2nd screening (in the column Notes_2 of they have the comment "Complete 2nd screening").

Concerning the 99 subjects that participated in the 2nd round of 1st screening (SURVEY=1), 23 filled in all questions of 1st and 2nd screenings (in the column Notes_2 they have the comment "Complete 1st and 2nd screening"), 98 of them had the 1st screening completely filled in and 24 had done all the questions of 2nd screening.

It is worth mentioning that in the group of subjects that have two entries in this database, apart from the problems named before, there are also many missing values for those who participated in the 1st round of the 1st screening. It was also noticed that only 27 participated in the 2nd screening.

Throughout this study, it was observed that all the subjects with SURVEY=2 had repeated value for ID_USER. This confirms that the 9 lines with SURVEY=2 were wrong entries on the platform with ID_USER values assigned before.

Table 4.2: Subjects with repeated ID_USER and different date of birth

ID_USER	SURVEY	T1_ALG_02
23538	1	1949-02-03
23538	0	1949-03-02
29276	1	1938-10-28
29276	0	1940-04-09
29845	1	1946-05-08
29845	0	1946-08-05
29929	1	1945-02-19
29929	0	1946-07-02
29950	1	1945-05-03
29950	0	1945-03-05
29971	1	1946-05-30
29971	0	1948-06-04
46177	1	1945-11-20
46177	2	1948-09-30
46241	1	1941-07-03
46241	2	1940-08-01
47670	1	1945-10-03
47670	2	1947-09-03
48058	1	1941-05-19
48058	2	1943-05-29
48387	1	1946-08-17
48387	2	1947-11-13
48695	1	1946-04-07
48695	2	1946-07-16
49151	1	1946-04-01
49151	2	1944-12-09

4.2 MUNICIPALITY variable

Table 4.3: Frequency table for MUNICIPALITY

MUNICIPALITY	Frequency
	2
Enschede	1213
Hengelo	1023
Tubbergen	428
Twenterand	639

Regarding the MUNICIPALITY variable, the distribution of the values can be observed in a frequency table, Table 4.3, noting that there are 2 cases for which the municipality was not recorded. These subjects are women who answered the 1st screening questionnaire online and have ID_USER's values 12965 and 12923. The ID_USER=12923 also has missing values for the variables SF_12_PCS and SF_12_MCS.

4.3 AGE variable

With regard to AGE variable some invalid values were detected such as -1, 44 and other values outside of the established age interval [65,75] (marked on grey in Table 4.4).

There were 116 lines with values outside of the interval for AGE variable, but only 12 corresponded to odd values such as -1, 44, 114 and 115.

Additionally, 448 different birth dates (T1_ALG_02 variable) were found, that had records with different values to the AGE variable. Since the reference date is not the same for all the individuals, it is possible to have subjects with the same birth date but with different values for age. For example, an individual with the birth date 1.Dec.1940 that participated in SURVEY=0 (September 2014) would have an age equal to 73. Another individual, born in the same year, but in January (1.jan.1940, e.g.), who answered the 1st screening in February 2016 would have an age equal to 76. Ages equal to 74 and 75 are, obviously, also possible for participants born in 1940, depending on their month of birth and date of participation.

According to these principles the cases which fall outside the above mentioned possibilities were identified.

Table 4.4: Frequency table for AGE

AGE	Frequency
-1	2
44	1
62	1
63	6
64	29
65	227
66	345
67	338
68	324
69	370
70	300
71	268
72	297
73	256
74	245
75	219
76	58
77	9
78	1
114	1
115	8

4.4 SF_12_PCS and SF_12_MCS variables

For SF_12_PCS and SF_12_MCS variables (scores related to the SF12 test) the existing values were well calculated. Still, 209 individuals, who had missing values on both scores, were detected and yet had valid values in questions SF12 (see excel file *Appendix B.xlsx*, sheet "Missing SF12 scores"). Hence, the score variables SF_12_PCS and SF_12_MCS should not be missing.

4.5 MNA_short_SCORE variable

For the computation of the MNA_short_SCORE, it is necessary to take into account the subject's BMI, in addition to the answers A to E of MNA test. So, in order to verify if the MNA_short_SCORE was well computed, the values of BMI variable had to be verified first. After some calculations it was confirmed that these values were well calculated.

The values for the MNA_short_SCORE variable were calculated by taking the sum of all values from the questions of 1st MNA questionnaire (questions A to E from MNA questionnaire) with the score resulting from the codification of BMI variable. The frequency table of the results (Calculated MNA_short_SCORE) and the original values for MNA_short_SCORE variable are shown in Table 4.5.

The table shows that the calculated values were different from the originals. There were only 33 records with the same values, that corresponded precisely to the subjects who had value 0 for codified BMI.

Table 4.5: Frequency table for MNA_short_SCORE

Validation study	Database	Frequency
3	0	1
4	1	3
5	2	7
6	3	5
7	4	14
8	5	30
9	6	42
10	7	66
11	8	151
12	9	273
13	10	321
14	11	2389

Adding the original values of MNA_short_SCORE to the codified BMI of each subject, the sum was found to be equal to the calculated score in this validation study. Hence, the values shown in the database for MNA_short_SCORE were not indeed the correct values, since it missed the addition of the codified BMI. However, BMI was taking into account in the nutritional classification process, producing correct values for FIRST NUTRITIONAL STATUS. For example, an individual with MNA_short_SCORE in the database equal to 10 and a codified BMI of 3 is classified as "normal", since in fact the score value is 13.

4.6 FIRST_FINAL_STATUS variable

In FIRST_FINAL_STATUS variable the values in the database were also compared with the recomputed values through frequency tables that are condensed in Table 4.6.

Table 4.6: Frequency table for FIRST_FINAL_STATUS

FIRST_FINAL_STATUS	Validation study	Database
FRAIL	558	592
NULL	37	0
PRE-FRAIL	829	768
ROBUST	1881	1945

The bigger difference shown in this table was for "NULL" value which was not present in original FIRST_FINAL_STATUS variable. The new class "NULL" was attributed, in the validation study, to the individuals who could not be classified due to the existence of missing values in their scores. In addition to these, other differences were found between the values calculated in this study for FIRST_FINAL_STUDY and the database values, in a total of 100 subjects.

From the 100 cases, 37 individuals are part of those who should have been classified as "NULL". If missing values are present, the classification should be "NULL", similarly to what is done for the 2nd screening, unless the GFI_SCORE$>$ 4 in which case the classification should be "FRAIL". These subjects were misclassified as "FRAIL", "PRE-FRAIL" or "ROBUST" in the database (in the column Notes_FFS they have the comment "True NULL").

There were also subjects who had their final classification changed because of differences that occured in MNA_short_SCORE variable (in the column Notes_FFS they have the comment

"Changes in MNA_short_SCORE"), other who were well classified because of the introduction of the cognition class for the individuals who did the 1st round of 1st screening (in the column Notes_FFS they have the comment "Introduction of cognition class") and some of them had different calculated values due to these two reasons (in the column Notes_FFS they have the comment "Changes in MNA_short_SCORE + Introduction of cognition class").

4.7 QMCI_SCORE variable

Regarding QMCI_SCORE variable, another problem was detected. The variable's frequency table had only integer numbers while the calculation results included some decimal values with 5 tenths.

An explanation for the differences could be that values of QMCI_SCORE had been rounded up. Taking this into account the calculations were repeated but the differences were still observed. Therefore, 148 ID_USERS who had the calculated value for QMCI_SCORE different from the original value were identified. However only 8 of these errors changed the final classification of 2nd screening.

4.8 MNA_SCORE variable

Once again the original values of MNA_SCORE variable were compared with the recalculated values and there were 64 differences. The reasons for these differences are not clear, but there are some errors that may have been made in other variables from the questionnaire: wrong BMI's codification; wrong values for T2_MNA_r variable (calf circumference of the subject); wrong codification of MNA's question k (the amount of food eaten from different groups). These possible mistakes can not be properly verified since the codifications are not present in this database.

4.9 SECOND_PHYSICAL_STATUS variable

The classifications according to age and gender of the 3 tests from the physical part of 2nd screening could not be validated because they are not included in the PERSSILAA's database. Instead the score variables of this part are copies or sums of the test values.

Hereupon it was decided to check the SECOND_PHYSICAL_STATUS variable and 5 subjects were detected with original values different from the calculated values, that are shown below.

Except for the subject with ID_USER=25750, that was wrongly classified as "ROBUST" without having valid gender and age values, there was no evidence in the variables associated to the SECOND_PHYSICAL_STATUS that could explain these errors. For all the other subjects listed above no odd values in the variables of 2nd screening physical part were registered. This suggests that some mistakes might have been made in the scores that were not available.

4.10 SECOND_NUTRITIONAL_STATUS variable

Although 64 divergences in MNA_SCORE variable were found, only in 10 situations the difference in the scores implied different SECOND_NUTRITIONAL_STATUS. The 10 subjects

had different values for the original and the recalculated values of MNA_SCORE, implying different nutritional status.

4.11 SECOND_COGNITIVE_STATUS variable

In the cognitive test of 2nd screening (QMCI) 148 original values different from the calculated values for this variable were found. For 79 cases, there were no differences in the SECOND_-COGNITIVE_STATUS variable. This means that 69 individuals had values of QMCI_SCORE and SECOND_COGNITIVE_STATUS variables different from the calculated ones.

These differences may have happened given that SECOND_COGNITIVE_STATUS had not yet been updated after the corrections for T2_QMCI_04 and T2_QMCI_05 variables.

Adding to these subjects there were more 64 cases with divergences on SECOND_COGNITIVE_STATUS that had the original values of QMCI_SCORE equal to the calculated values for this variable. Therefore it can be concluded that there were problems in the computation of the SECOND_COGNITIVE_STATUS variable too.

4.12 SECOND_FINAL_STATUS variable

Finally SECOND_FINAL_STATUS was analysed once again comparing the original values with the recalculated values through frequency tables (Table 4.7). The calculations for SECOND_FINAL_STATUS were based on the recalculated values from the other status variables of the 2nd screening.

Table 4.7: Frequency table for SECOND_FINAL_STATUS

SECOND_FINAL_STATUS	Validation study	Database
null	1	0
PRE-FRAIL	324	345
ROBUST	29	9

Since only 21 subjects were misclassified according to the calculated classification in this validation study, this means that only a few of the errors found in the scores had impact on the final classification.

After the detection of these computation or typing errors, they were corrected and a new database was created, using a program developed in R software. This new database was used to perform the analysis of the subsequent chapters. Thus, the database validation, described in this chapter, was essential for a correct analysis of the data presented in the next chapters.

Chapter 5

1st screening

This chapter has the objective of describing the statistical analysis of the data coming from PERSILAA's 1st screening database.

The study was divided into 7 parts: a brief analysis on the individuals' main characteristics; the presentation of classification results in the 1st screening; a comparison of these results between municipalities; the characterization of each type of *persona* by Chernoff faces; a cluster analysis on the individuals who have participated in the 1st screening of the program; the classification results using LDA; the classification results using MLR.

All the results shown in this chapter were obtained from the database downloaded on 15-04-2016 at 15:56:25 and they relate only to individuals who participated in the 2nd round of the 1st screening (value 1 for SURVEY variable) and completed enough questions to allow classification.

5.1 Preliminary analysis

A total of 3173 individuals completed the 1st screening in the 2nd round, corresponding to 1520 men (48%) and 1653 women (52%). Respecting age, the individuals belong mainly to the group between 65 and 74 years (90%). The distribution of individuals by age groups do not seem to differ much according to their gender as it is visible in Table 5.1.

Table 5.1: Distribution by age and gender

Gender\Age group	≤64	65-69	70-74	≥75	Total
Male	12	724	645	139	1520
Female	8	820	676	149	1653
Total	20	1544	1321	288	3173

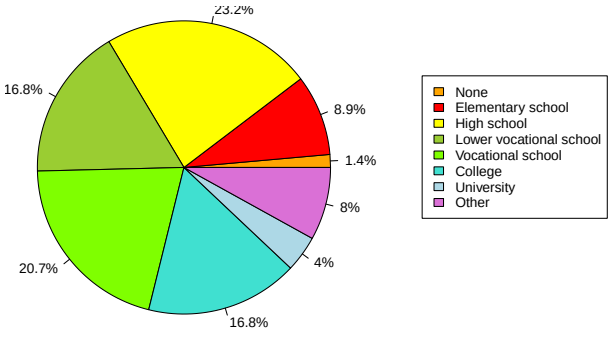


Figure 5.1: Education level

The subjects who participated in this study also gave information about their education level, other feature that may be relevant to their final classification. This population seems to have a medium-high education level since the majority of the subjects has concluded high school (or equivalent) or a higher level.

The percentage of individuals with each type of education level can be observed in Figure 5.1. As it was referred to above, 81.5% of the subjects completed high school, an equivalent school level or higher education.

Regarding the subjects' height and weight it was observed that the majority of the individuals

have height between 150 cm and 190 cm (97.7%) and weight between 50 kg and 100 kg (91.2%). However, the interval of values for these attributes is much more extensive. In the group of subjects who participated in the 1st screening there is only a person with the minimum height of 100 cm (could be a mistake) and another with the maximum of 202 cm, while for weight the minimum and maximum are 42 kg and 200 kg, respectively.

Figure 5.2 depicts the classification of the individuals regarding the BMI, according to the scale of the World Health Organization, in 4 main classes: Underweight ($BMI < 18.5$), Normal range ($18.5 \leq BMI < 25$), Overweight ($25 \leq BMI < 30$) or Obese ($BMI \geq 30$). It can be concluded that "Overweight" is the most represented category in the set of subjects who participated in the 1st screening. Besides that, the percentage of people with weight above normal ("Overweight" and "Obese" categories) is equal to 66.27%, which represents the majority of the respondents. This may suggest a greater difficulty in performing certain tasks.

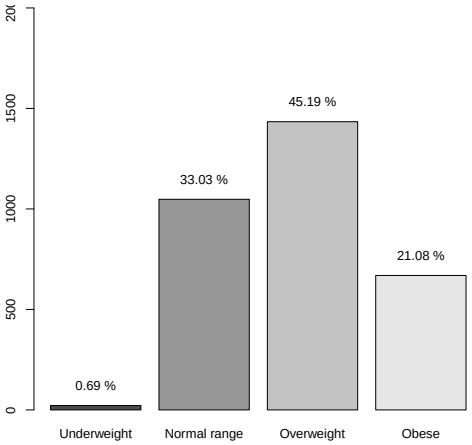


Figure 5.2: Classification according to BMI

The individuals in this study are from 4 different municipalities of Netherlands: Enschede, Hengelo, Tubbergen and Twenterand. The two municipalities which had more residents participating in the 1st screening were Enschede with 1084 individuals (34%) and Hengelo with 1019 (32%). The remaining municipalities are much less represented in the study with 429 individuals from Tubbergen (14%) and 640 from Twenterand (20%).

In order to conclude this preliminary analysis on the participants of 1st screening, 3 behavioral characteristics were studied. In Figure 5.3 are presented 3 bar graphs corresponding to each feature. It is clear that the studied individuals have healthy habits and some independence, since the majority of them do not consume many alcoholic beverages per week, do not smoke and are responsible for their own administration.

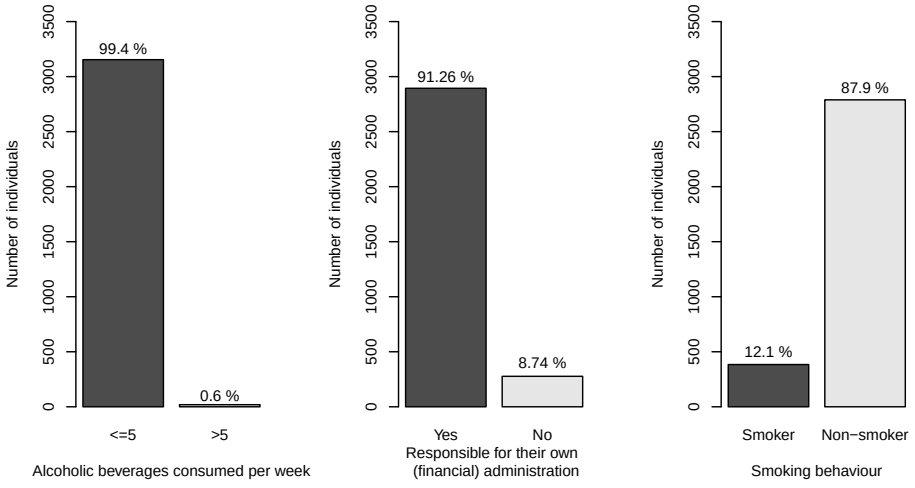


Figure 5.3: Behavioral characteristics

The number of alcoholic beverages consumed per week for each subject who participated in this 1st screening varies between 0 and 12, but, as the graphic shows, 99.4% of the individuals do not drink more than 5 alcoholic beverages per week.

The majority of the individuals (91.26%) are responsible for their own financial administration. Only 277 individuals, from the 3171 who filled in correctly this question, indicated not to be responsible for their administration.

Finally, the difference between the number of smokers and non-smokers is lower than the registered in the previous 2 features, but still large. In this set of older people 12.1% have smoking habits against 87.9% of non-smokers.

5.2 Classification results

The classification results obtained in the 1st phase of PERSSILAA's screening are shown in Table 5.2 and Table 5.3. The subjects listed in the studied database and who completed correctly the 2nd round of this 1st screening were mostly classified in the "ROBUST" class. On the other hand, the "FRAIL" category is the least represented.

Table 5.2: Final classification of 1st screening

FRAIL	PRE-FRAIL	ROBUST
521 (16.4%)	740 (23.3%)	1912 (60.3%)

With respect to the classifications in each domain of 1st screening, it was verified that the domain with more declining individuals is the physical. Nevertheless, the subjects who participated in this 1st screening are in good condition, since the percentage of people classified as "normal" is higher than 80% in all the 3 domains.

Table 5.3: Classification results from 1st screening

Domain\Class	decline	normal
PHYSICAL	619 (19.5%)	2554 (80.5%)
COGNITIVE	475 (15%)	2697 (85%)
NUTRITIONAL	380 (12%)	2793 (88%)

Domain\Class	FRAIL	PRE-FRAIL	ROBUST
GENERAL	521 (16.4%)	264 (8.3%)	2388 (75.3%)

Observing the results in the Table 5.3, it is noted that there are less individuals classified as "ROBUST" in the final status of the 1st screening than in the general classification from the GFI test. This means that the classification is more demanding when, in addition to a general score, the domains classifications are included.

The frail subjects are the same in the two classification methods, because this category is uniquely assigned by the GFI values. Many of the remaining individuals, who were classified as "ROBUST" in the general classification, have been declared as "PRE-FRAIL" when the results of the domains were taken into account.

The final classification of 1st screening is slightly different for each gender and it indicates that, in general, men may be more robust than women. However, this category was the most represented in the 1st screening for both genders, as it was also seen for all individuals as it is

shown in Figure 5.4. For the other two categories the distribution of individuals by gender is also similar to the global one.

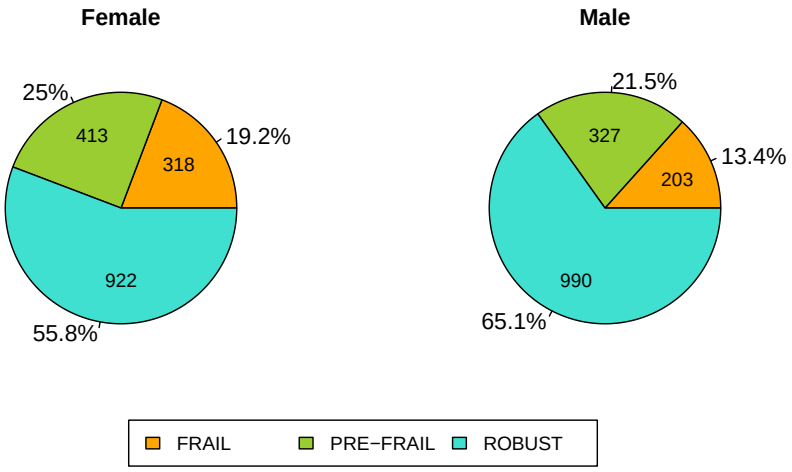


Figure 5.4: Final classification of 1st screening by gender

For the final status by age group of the participants in this 1st screening, it is shown in Table 5.4 how the individuals of each age group are distributed by the different classes. It can be observed that the percentage of robustness seems to decrease with age - in the group of people with age over or equal to 75, it is already much closer to the percentage of frailness.

Regarding the percentage of pre-frail individuals, this is approximately constant across the age groups, being systematically higher than the percentage of frail individuals, except for "≥75" where there is a larger percentage of frail individuals than pre-frail.

Table 5.4: Final classification of 1st screening by age

Age group\Class	FRAIL	PRE-FRAIL	ROBUST
<=64	3 (15%)	5 (25%)	12 (60%)
65-69	221 (14.3%)	327 (21.2%)	996 (64.5%)
70-74	211 (16%)	333 (25.2%)	777 (58.8%)
>=75	86 (29.9%)	75 (26%)	127 (44.1%)

Concerning the 1st screening's final classification according to the municipalities there are some differences, despite the fact that the global pattern is the same, with the "ROBUST" being the predominant category, followed by "PRE-FRAIL" and then "FRAIL". The detailed distribution of the individuals by municipality and final classification of the 1st screening can be observed in Table 5.5. This will be subject to further analysis later in Section 5.3.

Table 5.5: Final classification of the 1st screening by municipality

Municipality\Class	FRAIL	PRE-FRAIL	ROBUST
Enschede	214 (19.74 %)	289 (26.66 %)	581 (53.6 %)
Hengelo	149 (14.62 %)	222 (21.79 %)	648 (63.59 %)
Tubbergen	55 (12.82 %)	87 (20.28 %)	287 (66.9 %)
Twenterand	103 (16.09%)	142 (22.19%)	395 (61.72%)

5.3 Comparison of results by municipality

In order to detect if the differences in the final classification of the 1st screening for the 4 municipalities, observed in Table 5.5, were significant some Chi-squared tests of homogeneity were made resorting to *chisq.test()* function of R. First a global test of homogeneity with a p-value <0.001 rejected the hypothesis that the 4 municipalities were similar regarding the final status.

Hereupon it was necessary to understand which of the municipalities (one or more) are different from the others. For that purpose 6 tests were made, one for each pair of municipalities. The p-values which resulted from the tests were adjusted using the Holm method [11] implemented in the *p.adjust()* function of R.

Through the results presented in Table 5.6 it can be concluded that Enschede's population has a very significant different distribution of the final classification (marked on grey in Table 5.6), although the difference in the Enschede/Twenterand pair is not so strong when compared with the other two. The remaining pairs of municipalities are not significantly different with respect to the distribution of the FIRST_FINAL_STATUS variable (marked on white in Table 5.6).

Table 5.6: Results from the homogeneity Chi-squared tests for the pairs of municipalities

Pair of municipalities	p-value
Enschede/Hengelo	< 0.001
Enschede/Tubbergen	< 0.001
Enschede/Twenterand	0.018
Hengelo/Tubbergen	0.931
Hengelo/Twenterand	0.931
Tubbergen/Twenterand	0.554

Since significant differences between the municipalities have been found in the 1st screening's final status, it would be interesting to understand what happens for each domain. For this reason some tests with the scores and status of all domains were also performed trying to identify in more detail the reason for the differences. Besides tests of homogeneity performed with respect to the classification in each domain, Kruskal-Wallis (*kruskal.test()* in R software) and Mann-Whitney-Wilcoxon (*wilcox.test()* in R software) tests were performed to compare the scores for each domain.

Physical Domain:

In this 1st screening, as it was seen in Table 5.3, 2554 subjects were classified as normal (80.5%) and 619 as decline (19.5%) in the physical domain. To investigate the existence of any differences between municipalities in the physical domain of the 1st screening, two global tests were initialy made (Kruskal-Wallis for SF36_SCORE and Chi-squared of homogeneity for FIRST_PHYSICAL_STATUS). The results of both tests point towards a significant difference in the distributions of these physical variables by municipality, with p-values <0.001 for Kruskal-Wallis test and for Chi-squared test.

As a consequence of these results, it was again necessary to compare the values of SF36_-SCORE and FIRST_PHYSICAL_STATUS in pairs of municipalities. Table 5.7 shows the p-values obtained in the 2 types of tests performed for all possible pairs of municipalities (Wilcoxon and Chi-squared of homogeneity tests). Observing the adjusted p-values for multiple comparisons, the differences on SF36 test score and physical status of the 1st screening are evident and significant for all usual significance levels only for the pairs Enschede/Hengelo and Enschede/Tubbergen.

Table 5.7: Adjusted p-values from the tests in Physical Domain

Pair of municipalities	SF36_SCORE	FIRST_PHYSICAL_STATUS
Enschede/Hengelo	<0.001	<0.001
Enschede/Tubbergen	<0.001	<0.001
Enschede/Twenterand	0.113	0.051
Hengelo/Tubbergen	0.469	0.904
Hengelo/Twenterand	0.113	0.236
Tubbergen/Twenterand	0.086	0.24

Cognitive Domain:

Concerning the cognitive domain, 2697 individuals were declared normal (85%), while 475 had decline classification (15%) in this 1st screening, similar to what happened in the physical domain and as it was observed in Table 5.3. Once again, the two global tests made revealed that there is a significant difference between municipalities in the final cognitive score and status, with p-values <0.001 in the Kruskal-Wallis test for AD8_SCORE and in the Chi-squared homogeneity test for FIRST_COGNITIVE_STATUS.

The results of the tests that compared the cognitive scores and status in all possible pairs of municipalities are presented in Table 5.8. From this, and considering the usual levels of significance, it can be concluded that only the pair Enschede/Hengelo shows significant differences in the AD8_SCORE and besides these two municipalities also the pairs Enschede/Tubbergen and Enschede/Twenterand (this one not significant for all the significance levels) had presented differences in the FIRST_COGNITIVE_STATUS variable.

Table 5.8: Adjusted p-values from the tests in Cognitive Domain

Pair of municipalities	AD8_SCORE	FIRST_COGNITIVE_STATUS
Enschede/Hengelo	<0.001	0.003
Enschede/Tubbergen	0.149	0.006
Enschede/Twenterand	0.117	0.047
Hengelo/Tubbergen	0.292	1
Hengelo/Twenterand	0.288	1
Tubbergen/Twenterand	0.971	≈ 1

Nutritional Domain:

Regarding the nutritional domain, as it was recorded in Table 5.3, 2793 individuals were classified as normal (85%) whereas 380 participants of this 1st screening had a decline classification. In the first two tests done to check the significance of the differences between municipalities in relation to the nutritional classification for 1st screening, the results indicate that there was at least one municipality with a different distribution from the others. The calculated p-values were <0.001, once again, for the Kruskal Wallis test in the MNA_short_SCORE and the Chi-squared test of homogeneity to FINAL_NUTRITIONAL_STATUS.

Once again, due to the fact that significant differences between the municipalities have been found in the nutritional domain of the 1st screening, tests on the several pairs of municipalities were performed to look for any significant differences in the score and status of nutritional domain. In Table 5.9 the results of the tests are presented. It is possible to state that the pairs Enschede/Twenterand and Enschede/Tubbergen have significant differences in MNA_short_SCORE and only Enschede/Tubbergen pair shows differences in FIRST_NUTRITIONAL_STATUS.

Table 5.9: Adjusted p-values from the tests in Nutritional Domain

Pair of municipalities	MNA_short_SCORE	FIRST_NUTRITIONAL_STATUS
Enschede/Hengelo	0.185	0.303
Enschede/Tubbergen	<0.001	<0.001
Enschede/Twenterand	0.003	0.179
Hengelo/Tubbergen	0.047	0.015
Hengelo/Twenterand	0.185	0.56
Tubbergen/Twenterand	0.458	0.091

General Conclusions:

Concerning the distribution of the final classification of the 1st screening, the Enschede's residents are those who presented more significant differences when compared with the other populations. The two pairs of municipalities with stronger differences in the final status (Enschede/Hengelo and Enschede/Tubbergen) also showed significant differences in 2 or 3 domains. On the contrary, the pair Enschede/Twenterand, which presented differences not so significant in the final status of the 1st screening, showed significant differences only in the nutritional domain.

5.4 Characterization for each type of *persona*

From the 1st screening of PERSSILAA, 8 types of individuals can be identified according to their classification in the physical, nutritional and cognitive domains as listed in Table 5.10. Chernoff faces is one graphical representation that makes possible to visualize the differences regarding the main characteristics of the average *persona* for each of the 8 identified types [2].

Table 5.10: Types of *persona*

Persona	Physical	Nutritional	Cognitive
1	normal	normal	normal
2	normal	normal	decline
3	normal	decline	normal
4	decline	normal	normal
5	normal	decline	decline
6	decline	normal	decline
7	decline	decline	normal
8	decline	decline	decline

To obtain the Chernoff faces, one for each type of *persona*, the *faces()* function from *aplpack* package in R software was used. This function produces faces, with each of the face's feature associated to one variable, to a maximum number of 15 features. Hence, it was necessary to reduce the number of variables and calculate averages for each of the 8 types of individuals, for a better understanding of the graphs. For painting the elements of a face, the colors are found by averaging of sets of variables: (7,8)-eyes:iris, (1,2,3)-lips, (14,15)-ears, (12,13)-nose, (9,10,11)-hair, (1,2)-face.

The data matrix was split into 6 groups concerning the questions of ALG, GFI, SF36, SF12, AD8 and MNA_SF. Thereafter the principal components of each subset were calculated with the *PCAmix()* function from *PCAmixdata* package in R software which performs principal components for sets with a mixture of qualitative and quantitative variables. The principal

components were retained according to the following criteria: 3 for the groups corresponding to tests with more general questions such as GFI, SF12 and ALG and 2 for the groups with the domain questions like SF36, AD8 and MNA_SF. These choices were also made to associate the tests to a specific feature of the faces as it can be seen in Table 5.11.

In order to have one single face per type of *persona*, the averages of the principal components by type were calculated across the individuals in the database. The final data set on which the *faces()* function was applied has the averages of the retained principal components per type of *persona*, so the rows correspond to the types and the columns to the principal components.

Table 5.11: Correspondence between features and principal components

Features	Principal components
height of face	ALG dim1
width of face	ALG dim2
structure of face	ALG dim3
height of mouth	GFI dim1
width of mouth	GFI dim2
smiling	GFI dim3
height of eyes	SF36 dim1
width of eyes	SF36 dim2
height of hair	SF12 dim1
width of hair	SF12 dim2
style of hair	SF12 dim3
height of nose	AD8 dim1
width of nose	AD8 dim2
width of ear	MNA_SF dim1
height of ear	MNA_SF dim2

The results of this procedure are presented in Figure 5.5. Observing the faces it is seen that the types 1, 5, 6 and 8 have similar face colors that correspond to some ALG principal components. The types 3, 4 and 7 form another group with similar and darker face colors. Regarding profile 2, this is the one with the lightest and more different face color of the all group.

Regarding to the face feature it is noted that types 1, 2, 3 are very similar, what shows that these types of persona have identical characteristics and behaviors from ALG questions. In the mouth feature, which is related with GFI test, some resemblances can also be detected: 1, 2 and 3 have big lips and an unhappy smile (mean of GFI is 1.57), unlike 4, 6, 7 and 8 that have happy smiles (mean of GFI is 4.47).

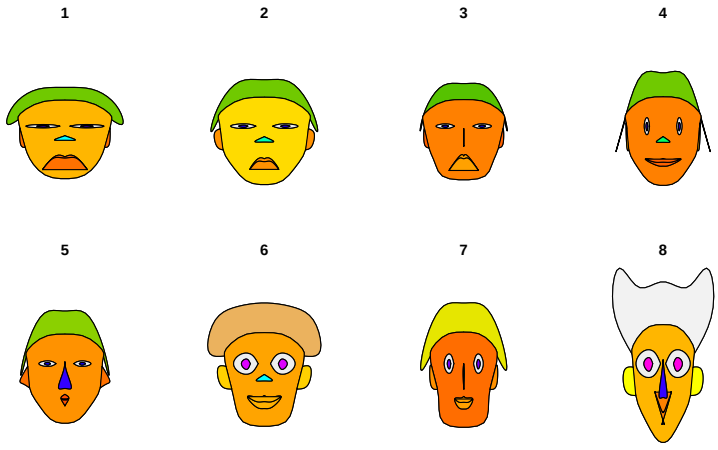


Figure 5.5: Faces for each type of *persona*

The eyes, which are related to SF36 test, are more flat for types 1, 2, 3 and 5 and more open to 4, 6, 7 and 8. This is due to the fact that the first indicated types of *persona* have been classified as normal for physical domain, while the remaining are decline to the same domain.

As for the hair and the hair color, the biggest differences are found between the 6, 7 and 8 types, which shows that they have significant differences from the other and between each other in the characteristics measured by the test SF12.

The types of individuals 3, 5, 7 and 8 have longer noses because they are classified as "decline" in the cognitive domain as measured by AD8 test, while the rest have small noses representing their normality with respect to the cognitive domain.

Finally, it is observed for the ears that types 2, 5, 6 and 8 have wider ears than 1, 3, 4 and 7 precisely because they have different status for nutritional domain.

General conclusions:

The faces 6, 7 and 8 are the most different from each other and also from the remaining faces and these have precisely two or more "decline" classifications, one of which referred to the physical domain.

Another fact to note is that these three types of *persona* along with the type 4 present happier smiles and more open eyes, features that refer to SF36 and GFI respectively, which are the reflection of their "decline" classifications in the physical domain.

The features which in this case undergo more changes are the shape and the color of the hair and the face, which correspond to the principal components of ALG and SF12 questions. These differences are more evident among the most frail persona types (with more "decline" classifications) what suggests that ALG and SF12 questions are useful to distinguish the most fragile individuals.

5.5 Cluster Analysis

With the purpose of understanding if a natural grouping of the individuals, taking into account their answers in the 1st screening, is similar to the division that emerged from its final classification, a cluster analysis was performed through a hierarchical clustering method.

The function used in R was *hclust()* which uses as input a dissimilarity matrix (it contains all the dissimilarity values for each variable between all pairs of individuals) and does an agglomerative hierarchical clustering with the possibility to choose the computing method for the distances between clusters. Due to the different nature of the 1st screening's variables (continuous, binary, ordinal, nominal, etc.) it was used a Gower coefficient [8] adapted to calculate the dissimilarity matrix.

For this analysis it was decided to take into account the variables of the 4 main tests from the 1st screening (GFI, AD8, MNA_SF, SF36) and also the variables from the questions groups ALG and SF12. With regard to the calculation of the distances between clusters in this agglomerative hierarchical clustering, two different methods were applied: complete-linkage method and Ward's method. The results of the cluster analysis were compared with the final classification of 1st screening in two tables which can be observed in Tables 5.12 (a) and (b).

The natural grouping of individuals according to their similarity based on the 1st screening questions is much closer to the final status using the Ward's method than with the complete-linkage method.

In the results of clustering with the Complete method there is a cluster which sticks out from the others (cluster 1), because it has much more individuals than the other two clusters.

Table 5.12: Agglomerative hierarchical clustering results

(a) Complete method

Cluster/FIRST_FINAL_STATUS	FRAIL	PRE-FRAIL	ROBUST
1	373	676	1786
2	89	3	0
3	4	0	0

(b) Ward method

Cluster/FIRST_FINAL_STATUS	FRAIL	PRE-FRAIL	ROBUST
1	218	425	130
2	237	49	0
3	11	205	1656

The cluster 1 holds the majority of subjects from all final classifications, so it is unfeasible to associate each cluster to a particular classification. This method was not capable to make a separation of the individuals from the 1st screening of PERSSILAA project.

Regarding clustering with the Ward's method the situation is different, since for each cluster there is a higher number of individuals of a certain classification. So the most immediate way to associate the resulting clusters with the classes of FIRST_FINAL_STATUS is: cluster 1 - PRE-FRAIL, cluster 2 - FRAIL, cluster 3 - ROBUST. Taking into account this association, it can be concluded that "ROBUST" class is the best grouped with only 7% of the individuals staying out of the cluster 3, while "PRE-FRAIL" and "FRAIL" classes have much higher percentages of wrong grouped people, 37.4% and 49.1% respectively. It is also observed that cluster 1 (associated to class "PRE-FRAIL") is, as expected, the one which contains more individuals from the other classes, because this is the class that is in between the other two, making the separation more difficult.

5.6 Linear Discriminant Analysis

Discriminant analysis can be used to validate a classification process as this screening protocol, comparing its results in a matrix containing the information about actual and predicted classifications, called confusion matrix, and evaluating the percentage of cases well classified. The classification results in discriminant analysis are based on a gold standard, which was not the case here. The purpose of this analysis is just to be able to compute posterior probabilities of membership, although we use as well the results of LDA to obtain the confusion matrix.

Given the present circumstances, it was assumed that the 3 classes are *a priori* equiprobable and the scores of the 3 domains and GFI were used as the predictor variables for the linear discriminant analysis (LDA). The function *lda()* from the MASS package of R was used for this analysis and, due to the fact that it can not deal with NA values, the dataset was reduced in 1 individual who had been classified as "FRAIL" through the GFI_SCORE despite having a NA value for the AD8_SCORE variable.

Table 5.13 is the confusion matrix for this comparison, which records the number of individuals who have been classified in the 9 possible pairs for the two distinct classification processes. The diagonal of this matrix counts the number of individuals who were well classified by the linear discriminant. This represents 90% of the total number of participants from this 1st screening,

Table 5.13: LDA results vs. FIRST_FINAL_STATUS classification

LDA/FIRST_FINAL_STATUS	FRAIL	PRE-FRAIL	ROBUST
FRAIL	520	8	0
PRE-FRAIL	0	602	182
ROBUST	0	130	1730

meaning that the LDA's classifier has produced 10% of wrong classifications for this 1st screening. The "FRAIL" category was the one with the better classification results, with no individuals classified as "PRE-FRAIL" or "ROBUST" by LDA. It is also noted from the table that there is a difficulty in classifying the pre-frail individuals, since 18.6% of these subjects were wrongly classified by LDA as "FRAIL" or "ROBUST". Finally, it can still be observed that the division between "PRE-FRAIL" and "ROBUST" is the hardest, because for both pre-frail and robust individuals there were some wrong classifications by LDA in the other group.

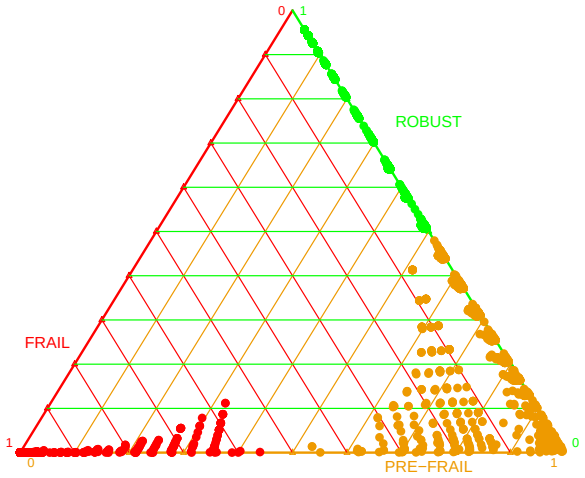


Figure 5.6: Posterior probabilities of LDA for each subject and class

In addition to the classification assigned to each individual, *lda()* function also returns the posterior probabilities for each subject of being classified as "ROBUST", "PRE-FRAIL" and "FRAIL". These probabilities were presented in the graphic of the Figure 5.6, according to the three axes corresponding to each of the three 1st screening's final classes.

Observing the figure it is possible to conclude that the pre-frail individuals have the most dispersed points, i.e., they are not concentrated on the right side of the "PRE-FRAIL" axis. This happens because some subjects classified as pre-frail have probabilities of belonging to another class not so close to zero, despite the higher probability of belonging to "PRE-FRAIL" class. As to the individuals classified as frail and robust, in general, they have probabilities of being classified in other class very close to zero, except some cases who have probabilities of belonging to "PRE-FRAIL" slightly higher.

5.7 Multinomial Logistic Regression

To complete the study of the final classification of 1st screening, an analysis based on multinomial logistic regression models was also carried out. The main objective of this is, as it was done with LDA, to compare the classification resulting from the models with the final status attributed by the 1st screening's questionnaire.

The function used for getting this results was the *multinom()* from *nnet* package in R sotfware, which requires as input the training data set with the values for all the covariates of interest as well as the values for the response variable. Three different approaches were followed: in the first approach only the variables of the SF36, AD8, MNA_SF and GFI tests were used as covariates of the model; the second model considered the covariates of the first model plus the ALG and SF12 variables (except the two repeated in SF36); the last model included the

same covariates as the second, except that variables height and weight were replaced by the BMI_SCORE.

Since the *multinom()* function eliminates the subjects in the data set who contain missing values for any of the variables, the number of individuals classified varies according to the model and the selected covariates. In this specific case, model I was applied to 3171 individuals, while the data set for model II and III recorded 2931.

The results of the prediction, which came from the application of *predict()* function to each model in R, are compared with the final classification for 1st screening in Tables 5.14 (a), (b) and (c). Further the values for residual deviance and AIC of each model are also shown and can be used for comparing the models' performance.

Table 5.14: MLR results vs. FIRST_FINAL_STATUS classification

(a) Model I: Residual deviance=955 ; AIC=1179

MLR/FIRST_FINAL_STATUS	FRAIL	PRE-FRAIL	ROBUST
FRAIL	520	0	0
PRE-FRAIL	0	603	76
ROBUST	0	136	1836

(b) Model II: Residual deviance=846 ; AIC=1250

MLR/FIRST_FINAL_STATUS	FRAIL	PRE-FRAIL	ROBUST
FRAIL	466	0	0
PRE-FRAIL	0	566	73
ROBUST	0	113	1713

(c) Model III: Residual deviance=818 ; AIC=1218

MLR/FIRST_FINAL_STATUS	FRAIL	PRE-FRAIL	ROBUST
FRAIL	466	0	0
PRE-FRAIL	0	577	73
ROBUST	0	102	1713

Doing a quick analysis on the tables it stands out, once again, the separation problem between robust and pre-frail individuals. In all the three fitted models the classification of frail individuals was consistent with the recorded status of the 1st screening. While for "ROBUST" and "PRE-FRAIL" classes there are some individuals misclassified, resulting in around 7% of misclassification in the model I and 6% in the two other models.

The models are similar concerning the misclassification but are different in terms of the goodness of fit. The values of the residual deviance indicate that the best model is the third, because it has the lowest value, 818. Regarding the AIC values, the first model is considered the best (1179). Taking into account all these measures (percentage of misclassification, residual deviance and AIC), the model III is considered the best model of the three.

The *multinom()* function provides also the posterior probabilities of belonging to each class for each individual of the set, based on the fitted models. Due to that and only for Model III, it was created a graphical representation in Figure 5.7 with the subjects' probabilities of being classified in each class, similar to what was done for LDA in Figure 5.6.

It is visible that with this classification approach it is achieved a better separation of the classes than with LDA. The frail individuals have probabilities of membership very close to one in the "FRAIL" class and very close to zero or even zero in the other two classes, while the rest

who were classified in the "PRE-FRAIL" and "ROBUST" classes of 1st screening have also high probabilities in their classes and very low for "FRAIL" class. However, some subjects still have relatively high values for the probability of being classified in the remaining class ("PRE-FRAIL" and "ROBUST" for robust and pre-frail people respectively). This confirms once more that the separation between the "PRE-FRAIL" and "ROBUST" classes is a complicated process and the 1st screening is not enough to classify entirely the individuals who are at the border between this two classes.

The results of MLR, which were performed taking into account the questions from the questionnaire of the 1st screening, were better than those obtained using LDA, which only considered the scores. This suggests that a classification based on the questions depicts better the condition of the individuals.

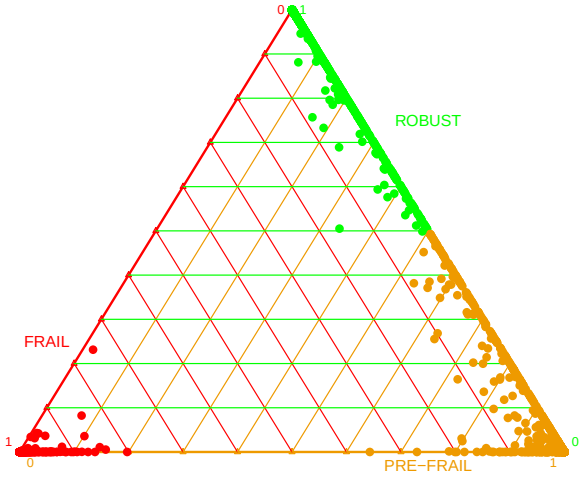


Figure 5.7: Posterior probabilities of MLR for each subject and class

After this analysis, it is possible to have a general idea of the 1st screening's participants and to identify the main questions of the questionnaire of the 1st screening to differentiate the classification profiles (*persona*). Besides that, some very significant differences were found in the 1st screening's classification of the individuals from the Enschede municipality when compared with the other three and other classification methods were performed and compared to the classification obtained according to the PERSSILAA's protocol.

Chapter 6

2nd screening

This chapter refers to the statistical analysis conducted with the data coming from the 2nd screening of the PERSSILAA project. Since only part of the individuals who participated in the 1st screening were called to the 2nd screening, the number of subjects in this analysis is much smaller.

The database used was the same as in the previous chapter (downloaded on 15-04-2016 at 15:56:25). For the present study only the individuals who were classified in the 2nd screening and have participated in the 2nd round of the 1st screening were analyzed. The classification rule of this 2nd screening includes the same three classes of the 1st screening, but in the database there are only records with "PRE-FRAIL" and "ROBUST" classes, since no individual was classified as frail.

This chapter starts with a preliminary analysis of the subjects who are under the above conditions, followed by an analysis of the classification results according to some individuals' characteristics and a comparison of the results by municipality.

6.1 Preliminary analysis

The number of individuals selected for the 2nd screening that actually participated in it was 522. The two genders are nearly equally represented in the sample, as in the 1st screening, with 243 male participants (47%) and 279 women (53%).

In relation to age most people (89%) is between 65 and 74 years old as in the 1st screening, but the age group with more people is 70-74. Comparing with the 1st screening, the group of participants in this 2nd screening is slightly older. As it can be seen in Table 6.1 the distribution of the subjects by age group appears to be equal for both genders.

Table 6.1: Distribution of individuals by age and gender

Gender\Age group	≤64	65-69	70-74	≥75	Total
Male	2	96	119	26	243
Female	2	123	127	27	279
Total	4	219	246	53	522

Concerning the education level of the 2nd screening's participants it is observed through the pie chart in Figure 6.1 that, like in 1st screening, a large majority of people (81.2%) has studied until high school or more. Besides that it is also noticed the big percentage of elderly who have a undergraduate degree (17%).

With respect to weight, the maximum value was 170 kg, while the minimum continued to be 42 kg. About 50% of the participants in this 2nd screening recorded a weight between 69 kg and 90 kg.

The minimum height recorded for a participant was 134 cm, while the maximum value was 196 cm. Half of individuals has recorded values for height between 165 cm and 177 cm, i.e., in a 12 cm range interval.

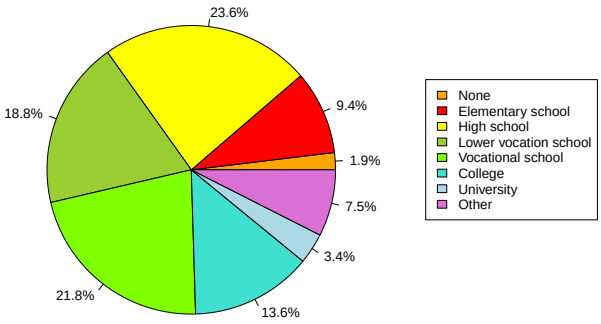


Figure 6.1: Education level of individuals

For classifying the subjects according to their Body Mass Index it was resorted again to the scale from World Health Organization, as it can be seen in the bar graph of Figure 6.2. As it was expected, due to the fact that no large differences in height and weight values of participants have been detected in the two groups (1st and 2nd screening), the results of this classification are also identical to the ones from the 1st screening. The main represented category is "Overweight" with 41% of the 2nd screening's participants and the people with weight well above normal ("Overweight" and "Obese" categories) remain as the majority of the group. Despite the resemblance, the group of participants in this 2nd screening has a higher percentage of underweight people (1.53%) than in the 1st screening.

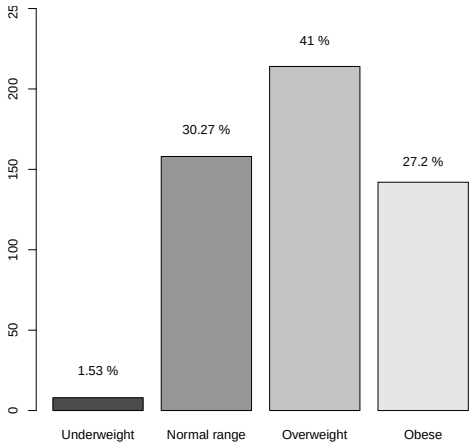


Figure 6.2: Classification of individuals according to BMI

Regarding the distribution by municipalities, the two with the highest participation were also Enschede and Hengelo with 270 (52%) and 131 (25%) individuals, respectively, but it is observed a larger difference between the two, with the residents of Enschede being the majority of the group. Twenterand and Tubbergen had much lower values of participation with only 74 (14%) and 47 (9%) residents in this 2nd screening, respectively. The data from the 2nd screening is much more unbalanced regarding the proportion of participants from each municipality than the one from the 1st screening (Enschede-34%, Hengelo-32%, Twenterand-20%, Tubbergen-14%).

To conclude the preliminary analysis of the characteristics of the participants in the 2nd screening, the questions related to people's behavior concerning alcohol consumption, smoking habits and financial administration were studied. Again, the results (Figure 6.3) reveal that the majority of participants have a healthy lifestyle.

As the first graphic shows, 99.43% of the individuals do not drink more than 5 alcoholic

beverages per week, i.e., a large majority of the participants consume moderate or even very low amounts of alcohol. The minimum consumption per week was 0, while the maximum was 12 beverages.

In relation to responsability and independence, 474 (91.15%) declared to be independent, while only 8.85% assumed not to be.

Observing the rightmost bar chart, the smoking habits of the 2nd screening's participants are rather less healthier than for the other two previously analyzed. This is confirmed by the percentage of smokers, 14.94%.

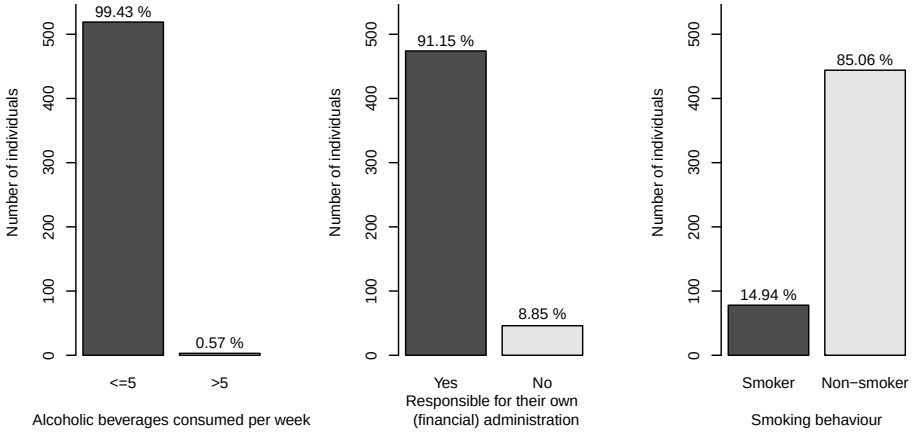


Figure 6.3: Behavioral characteristics of individuals

6.2 Classification results

The 2nd screening's main results are presented and discussed in this section with the support of Tables 6.2 and 6.3, which contain respectively the final classification and the classification by domains. As it was said before, the objective of the 1st screening was to select participants to go to the 2nd screening, who in principle should have been classified as pre-frail in the 1st screening. As a consequence, the data collected in the 2nd screening is not representative of the whole elderly population, thus the results of the new classification are much more unbalanced than in the 1st screening, with only subjects classified as "PRE-FRAIL" and "ROBUST" and pre-frail people being the large majority of the participants with 89.8% of representation.

Table 6.2: Final classification of 2nd screening

PRE-FRAIL	ROBUST
469 (89.8%)	53 (10.2%)

Table 6.3: Classification in the 3 domains of 2nd screening

Domain\Class	FRAIL	FUNCTIONAL DECLINE	null	ROBUST
PHYSICAL	20 (3.8%)	382 (73.2%)	0 (0%)	120 (23%)
COGNITIVE	41 (7.9%)	289 (55.4%)	5 (1%)	187 (35.8%)
NUTRITIONAL	0 (0%)	45 (8.6%)	1 (0.2%)	476 (91.2%)

The classification of the 3 domains which contributes for the 2nd screening's final status is carried out with different classes from those of the 1st screening. In this case the categories are

"FRAIL", "FUNCTIONAL DECLINE" and "ROBUST", but there is also the "null" class which is assigned to some individuals who had "NA" values for some questions of the QMCI and MNA tests. The individuals who were not able to perform one (or more) of the specific physical tests were considered frail in the physical domain. It should be noted that the "FRAIL" class is the least represented in all domains, which can also be a consequence of the fact that the frailest individuals were not chosen to participate in the 2nd screening. The "FUNCTIONAL DECLINE" category is the one with more individuals for physical and cognitive domains, but the percentage of participants thus classified is higher for the first domain. On the other hand, in the nutritional domain is the "ROBUST" class the most represented with 91.2% of the participants.

According to gender there are not signifficant differences in the final classification of the 2nd screening as it can be seen in Figure 6.4. The individuals classified as pre-frail are the majority in both genders, but men have a higher percentage of pre-frail individuals than women. This difference between genders on the distribution of the 2nd screening's final classification is equal to 2%, a much smaller value than the one from the 1st screening (see Figure 6.4).

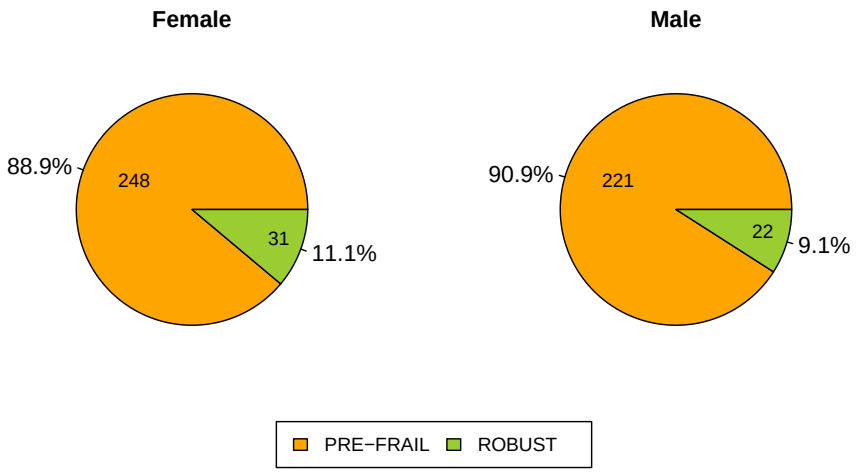


Figure 6.4: Final classification of 2nd screening by gender

The participants in the 2nd screening can be divided in 4 age groups, making possible to study the values of the 2nd final classification for each set of individuals. In Table 6.4 the classification results are displayed by age group. The class "PRE-FRAIL" is the most represented with more than 87% in all the 4 groups, as it should be expected. The participants who were between 70 and 74 years old have had the highest percentage of robustness (12.6%).

Table 6.4: Final classification of 2nd screening by age

Age group\Class	PRE-FRAIL	ROBUST
<=64	4 (100%)	0 (0%)
65-69	200 (91.3%)	19 (8.7%)
70-74	215 (87.4%)	31 (12.6%)
>=75	50 (94.3%)	3 (5.7%)

The same analysis was done for the municipalities where the participants live and similar results were obtained. It can be observed in Table 6.5 the final status of 2nd screening divided by municipalities. Once again in all the 4 municipalities the majority of individuals was classified as "PRE-FRAIL", this time is the Tubbergen municipality which has less percentage of pre-frail

individuals (80.85%), while Twenterand has the highest percentage (94.59%). Similar to what was done for 1st screening, in Section 6.3 a thorough study about the differences between the distributions of the final classification of 2nd screening of each municipality was done.

Table 6.5: Final classification of 2nd screening by municipality

Municipality\Class	PRE-FRAIL	ROBUST
Enschede	240 (88.89 %)	30 (11.11 %)
Hengelo	121 (92.37 %)	10 (7.63 %)
Tubbergen	38 (80.85 %)	9 (19.15 %)
Twenterand	70 (94.59%)	4 (5.41%)

6.3 Comparison of results by municipality

The same methodology as before was used in order to verify if there were significant differences between the second final classification of the participants from various municipalities (Table 6.5). The global test of homogeneity had a p-value equal to 0.066, so there is no substantial statistical evidence to doubt about the homogeneity of the distributions.

Table 6.6: Results from the homogeneity chi-square tests for the pairs of municipalities

Pair of municipalities	adjusted p-value
Enschede/Hengelo	0.764
Enschede/Tubbergen	0.764
Enschede/Twenterand	0.764
Hengelo/Tubbergen	0.276
Hengelo/Twenterand	0.764
Tubbergen/Twenterand	0.226

A test of homogeneity for the 6 pairs of municipalities gave the results presented in Table 6.6 and confirm with great confidence that there are no statistically significant differences between the pairs of municipalities regarding the 2nd screening's final classification of their residents.

The fact that no significant differences have been found for the distributions of 2nd screening's final status by municipality does not imply that there are also no differences in the domain classifications. To complete this study, a few more homogeneity tests on the scores and the status of each 2nd screening's domain were performed. Since these variables are of different types, various homogeneity tests were done: Chi-square tests for categorical variables; Kruskal-Wallis and Mann-Whitney-Wilcoxon tests for quantitative variables. Once again the R tools used for these tests were *chisq.test()*, *kruskal.test()* and *wilcox.test()* functions.

Physical Domain:

As shown in Table 6.3, 382 individuals were classified in the "FUNCTIONAL DECLINE" class (73.2%), 120 in the "ROBUST" class (23%) and finally 20 participants integrated the "FRAIL" class (3.8%). The physical part of this 2nd screening is composed by 4 different practical tests which also have as results the classes "normal" and "functional decline".

Table 6.7: Results from the global tests in Physical Domain

Test	TUGT	CSRT	MST	CST	SECOND_PHYSICAL_STATUS
p-value	<0.001	0.828	<0.001	0.028	0.03

To evaluate any differences in the municipalities on the physical tests or in the final physical

status the Chi-squared homogeneity test (categorical variables) was used. First 5 global tests were performed (physical tests and physical final status) and the results are shown in Table 6.7. The timed up and go test (TUGT) and the two-minute step test (MST) are, according to the presented results, the only tests which show statistically significant differences between municipalities. The p-values corresponding to the other physical tests and the physical final status indicate that the differences in the correspondent classifications are less significant.

The results of the Chi-squared tests of homogeneity for the pairs of municipalities are presented in Table 6.8. It is clear that there is no evidence to doubt on the homogeneity of chair sit and reach test (CSRT) distribution for all pairs of municipalities (p-value=1). Regarding the chair stand test (CST) and the physical classification of the 2nd screening (SECOND_PHYSICAL_STATUS), the results also show that there is no statistical evidence to assume that there is a different behaviour among the participants from the 4 municipalities for these two variables. Concerning the variables which have shown significant differences in the global tests (TUGT and MST), each of them only present differences for some pairs of municipalities. Respecting TUGT test, the pairs Enschede/Tubbergen and Tubbergen/Twenterand showed to have the most significant differences, while in the MST test Enschede/Hengelo and Hengelo/Twenterand were the pairs which showed more significant differences.

Table 6.8: Adjusted p-values from the tests in Physical Domain

Pair of municipalities\Test	TUGT	CSRT	MST	CST	SECOND_PHYSICAL_STATUS
Enschede/Hengelo	0.091	1	<0.001	0.104	0.70
Enschede/Tubbergen	0.002	1	0.685	0.519	0.02
Enschede/Twenterand	0.293	1	0.685	0.421	1.00
Hengelo/Tubbergen	0.119	1	0.285	1	0.31
Hengelo/Twenterand	0.293	1	0.002	0.064	1.00
Tubbergen/Twenterand	0.01	1	0.685	0.421	0.18

Cognitive Domain:

With respect to cognitive domain of the 2nd screening, there were 289 participants with a functional decline classification (55.4%), 187 were declared robust (35.8%) and 41 individuals had "FRAIL" classification, as it can be observed in Table 6.3. In addition 5 subjects failed to perform the QMCI test for the cognitive domain. The homogeneity tests used for evaluating the municipalities differences in this case were the Kruskal-Wallis for the QMCI_SCORE and Chi-squared for SECOND_COGNITIVE_STATUS. The p-values of the global tests were 0.004 and 0.036, respectively for QMCI and the cognitive final status.

To evaluate the homogeneity of cognitive scores and classifications among the various pairs of municipalities present in the study, the Mann-Whitney-Wilcoxon tests were applied to the QMCI_SCORE, while a Chi-squared test was again applied to the SECOND_COGNITIVE_STATUS. The results of this analysis are in the Table 6.9 and it can be seen that only for pair Enschede/Hengelo there is evidence of significant differences in the distributions of the cognitive variables.

Table 6.9: Adjusted p-values from the tests in Cognitive Domain

Pair of municipalities	QMCI_SCORE	SECOND_COGNITIVE_STATUS
Enschede/Hengelo	0.003	0.062
Enschede/Tubbergen	0.394	0.683
Enschede/Twenterand	0.915	0.683
Hengelo/Tubbergen	0.915	1
Hengelo/Twenterand	0.161	1
Tubbergen/Twenterand	0.854	1

Nutritional Domain:

Regarding the 2nd screening's nutritional domain, 476 individuals were classified as robust (91.2%) and 45 participants had a functional decline classification (7.9%), as shown in Table 6.3. Besides there was also one man who has participated in the 2nd screening who did not complete the MNA test.

The tests performed for this domain were the same as for the cognitive domain, because there are also a categorical variable (SECOND_NUTRITIONAL_STATUS) and a quantitative variable (MNA_SCORE). The p-values for the global tests were 0.143 and 0.749, respectively to the Qui-squared test of homogeneity and Kruskal-Wallis test, which shows that there is no significant differences between municipalities with respect to nutritional domain. As a consequence of these findings, it was considered not to be relevant to perform a comparison between the pairs of municipalities.

In this chapter, a general picture of the 2nd screening's participants was given, in addition to be shown that there are no significant differences in the classification between individuals from different municipalities. According to the characterization of the individuals who participated in the 2nd screening, it was possible to conclude that there are no major differences regarding the 1st screening, although this is a specific subset of the participants of the 1st screening.

Chapter 7

Comparison between screenings

The aim of this chapter is to discuss the statistical methods used to compare the results coming from the 1st and 2nd screenings and to study the association between the tools used in each domain in the two screening procedures.

The database studied in this analysis was the same considered in the analysis in chapters (downloaded on 15-04-2016 at 15:56:25), but focusing on the subjects who participated in booth 2nd round of 1st screening and 2nd screening and that have completed enough questions to be classified in the two screenings.

This analysis was divided into 3 sections: the presentation of final and domains classification's results for both screenings; ROC analysis between 1st screening test scores and 2nd domains classification; logistic and multiple linear regression to determine which questions from the 1st screening questionnaire are more relevant for the final classification and for the scores obtained in each domain on the 2nd screening, respectively.

7.1 Results

The classification results of 1st and 2nd screenings for the 521 participants in the conditions mentioned above (all the individuals considered in the study of the previous chapter, except one who did not complete properly the questionnaire of the 1st screening) are recorded in Table 7.1. In addition to the absolute frequency for each pair of classifications, it is also shown how the individuals from the three classes of the 1st screening distribute (in percentage) along the two classes of the 2nd screening.

It is known that PERSSILAA has as an objective of selecting individuals to the 2nd screening who are considered as pre-frail in the 1st screening, because they are the ones who will eventually take full benefit of PERSSILAA program. However, some frail and robust individuals were also chosen to participate in the 2nd screening.

Table 7.1: Final classification of booth screenings

1st / 2nd	PRE-FRAIL	ROBUST
FRAIL	33 (97%)	1 (3%)
PRE-FRAIL	386 (90%)	44 (10%)
ROBUST	49 (86%)	8 (14%)

The results show that the majority of the participants of the 2nd screening were classified in the "PRE-FRAIL" class, including a large percentage of those who had been classified as frail or robust in the 1st screening . This observation was the starting point to undergo a more detailed study with the aim of understanding why the 1st screening may fail in detecting certain types of frailties which can be determinant for a "PRE-FRAIL" classification.

Considering the physical domain, it is possible to observe in Table 7.2 that only 30% of the individuals who had a SF36 score above 60 in the 1st screening, and hence classified as normal, were indeed considered robust in the physical domain using the tools of the 2nd screening. On the other hand 11% of those who had a SF36 score below 60, and hence classified as in decline, were considered physically robust in the 2nd screening.

Table 7.2: Final physical classification of booth screenings

1st / 2nd	FRAIL	FUNCTIONAL DECLINE	ROBUST
decline	16 (8%)	155 (81%)	21 (11%)
normal	4 (1%)	226 (69%)	99 (30%)

Due to these discrepancies it was decided to study whether the SF36 score had enough discriminatory power regarding the classification of the individuals in the physical domain, either regarding their final physical status, or regarding the classification in each of the 4 tests composing the physical test of the 2nd screening. This will be done in the next sections.

The comparison of the cognitive classification of the two screenings, which can be observed in Table 7.3, clearly indicates that probably the tools used in the two screenings evaluate different aspects of the cognitive status of the individuals, since 48% of the total number of participants were differently classified in the 1st and 2nd screenings.

In this case the "null" class refers to some individuals who had "NA" values for some questions of the QMCI test. This may be errors in the records of the values in the PERSSILAA's database.

Table 7.3: Final cognitive classification of booth screenings

1st / 2nd	FRAIL	FUNCTIONAL DECLINE	null	ROBUST
decline	18 (11%)	105 (62%)	3 (2%)	44 (26%)
normal	23 (7%)	183 (52%)	2 (1%)	143 (41%)

Regarding the nutritional domain, the classification results are recorded in Table 7.4. In this domain the people classified as normal in the 1st screening had only 2% of changes, while 66% of the participants in decline on the 1st secreening were classified as robust in the 2nd. The nutritional domain has a very low percentage of subjects who had their class altered from the 1st to the 2nd screening (16%) compared with the other domains and even with the final classification, but this may due to the fact that the MNA test of 2nd screening also covers the questions from the MNASF test of the 1st screening.

Table 7.4: Final nutritional classification of screenings

1st / 2nd	FUNCTIONAL DECLINE	ROBUST
decline	37 (34%)	73 (66%)
normal	8 (2%)	403 (98%)

7.2 ROC analysis

This ROC analysis was used to evaluate the performance of the 1st screening's tools to predict the 2nd screening's classification. In the ROC analysis is necessary to define what means a test being positive and negative. In medical terms this is, in general, clear. Here it was considered that a test would be positive for values corresponding to a decline status and negative when the individual is classified as normal. The study was divided in the 3 domains and the data is slightly different for each, because the cases with "null" values for the variable status of each domain need to be deleted.

Since the data sets are quite unbalanced regarding the 2nd screening's status, two different approaches were taken: 1 - the analysis was performed on the original data; 2 - 100 different

samples with the same number of robust and declining individuals were randomly selected and the same methodology was applied.

During this study some inadequate values were found in the TUGT and CSRT variables. In the TUGT some values very close to 0 (less than 1) were detected and value 124. It is very unlikely that the values of the first case are correct, because this test measures actions as stand, walk and sit in seconds. With regard to the value 124, this is an outlier which does not seem correct, because it is too far from the range of other values [0,27].

Regarding the CSRT variable, the values which were interpreted as incorrect were those which are extremely high, for example greater or equal to 15. Considering that this test measures the reach distance of the hand with the foot as the reference (value 0), it was perceived that values above 14 would not be acceptable for an healthy adult, much less to a elderly. It is believed that these values have been incorrectly inserted into the database as the symmetric of the correct measure.

All the following analysis was done without the values mentioned above and the ROC curves, AUC and confidence intervals for AUC were performed with the *roc()* and *ROC()* functions of pRoc and Epi packages from R, respectively.

The 1st screening's tests of the 3 domains are from two different types: SF36 and MNA (short version) attribute higher scores to people in better physical/nutritional conditions, while AD8 produces lower values to those with more cognitive skills. The R functions used in this analysis have also different forms of acting: *roc()* has a parameter to define the direction to make the comparison and the default option is "auto"; *ROC()* has implemented the rule to classify an individual as positive (in this case is declining) if the score is above the cut-off point. For these reasons it was necessary to reverse the scores for the SF36 and MNA (short version) tests in the use of *ROC()* function.

Physical tests of 2nd screening:

The physical domain was more explored than the other domains, in this ROC analysis, because it is the one which has 4 tests with binary results. In the Figure 7.1 are shown the graphical representation of the ROC curve for all the physical tests done in the 2nd phase of screening.

Observing the graphs it is noted that, for all the 4 physical tests, SF36 test has a different optimal cut-off point and these are not equal to the cut-off point choosen to classify the individuals in the 1st screening's physical domain (marked in red in each chart). The differences between the cut-off points is not surprising, because they refer to distinct tests and may require different physical capacity of the individual.

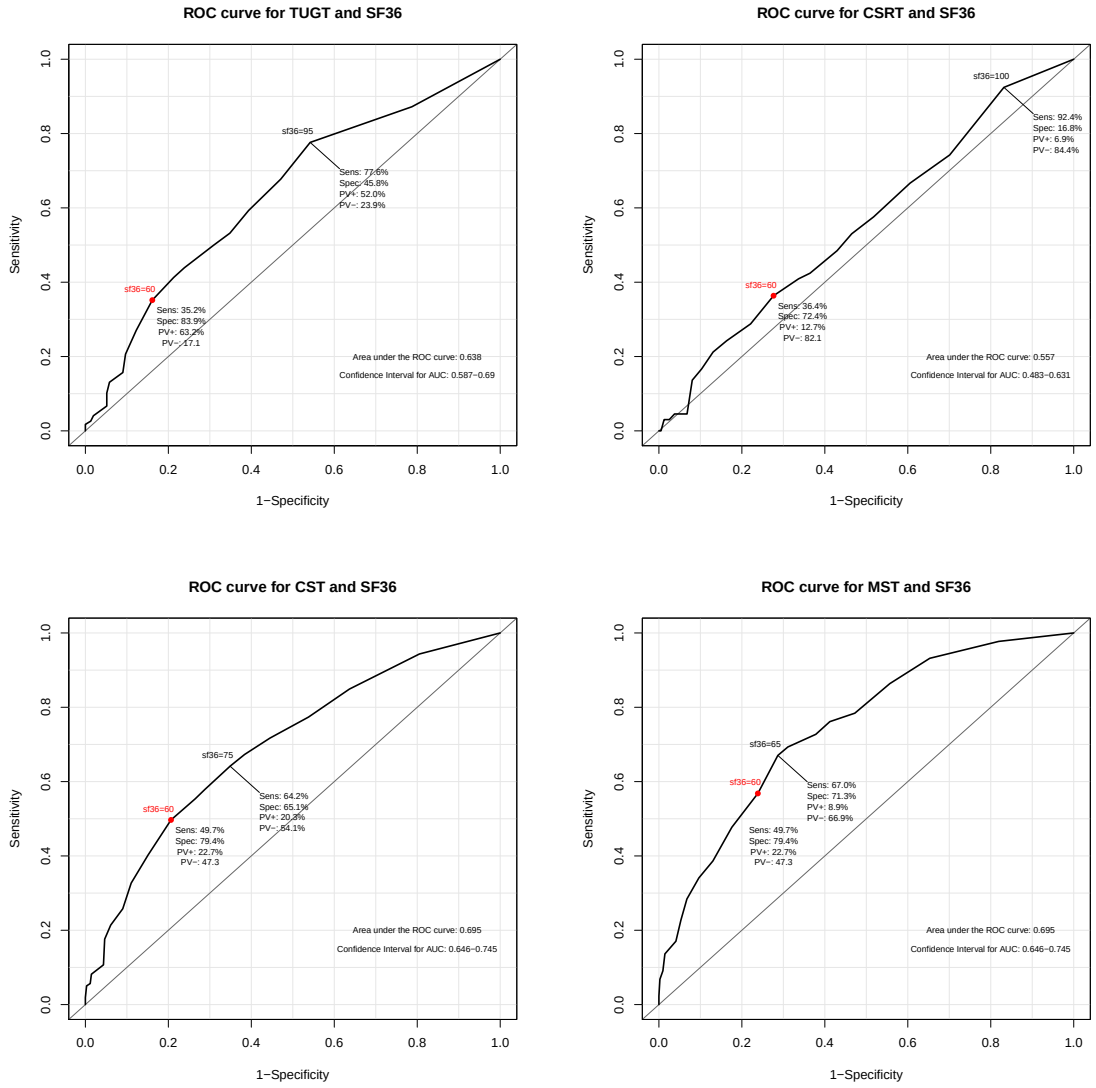
Table 7.5: AUC from ROC curves of samples physical tests

Measure/Test	TUGT	CSRT	CST	MST
Mean	0.657	0.601	0.719	0.773
Standard deviation	0.021	0.05	0.027	0.034
(0.025-0.975) quantiles	(0.616-0.693)	(0.494-0.681)	(0.66-0.762)	(0.693-0.833)

The AUC of these ROC curves does not go much beyond 0.7, meaning that the SF36 test is poor in discriminating the participants between "functional decline" and "normal" classes.

From the application of the mentioned functions to the generated balanced samples resulted some graphs and measures that were saved. In this report are only presented, in the Table 7.5,

Figure 7.1: ROC curves for Physical tests



the means and standard deviations of the AUC obtained in each balanced sample for the 4 tests of physical domain. It is clear that the averages for AUC are all better than the values obtained with the original data and the standard deviations are greatly reduced which shows that should not have had large swings.

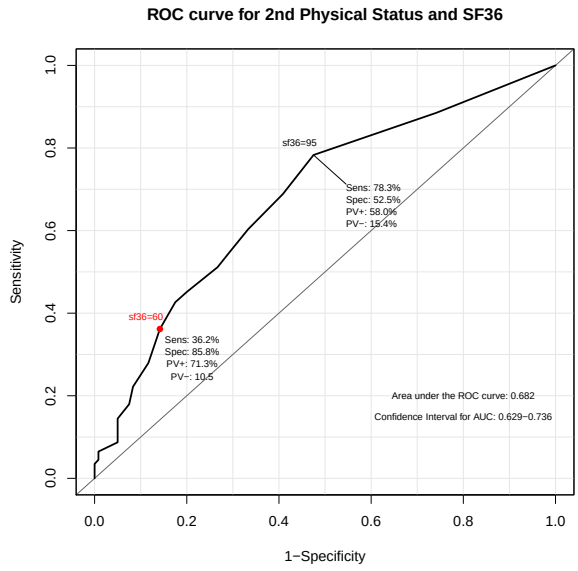
Physical domain:

Regarding the physical domain of the 2nd screening in a more general way, a ROC curve analysis was again performed, with the scores of SF36 test, but this time to evaluate how it can predict the final status of this domain. As the ROC analysis is done only for binary tests and the SECOND_PHYSICAL_STATUS is a categorical variable with 3 classes, it was necessary to alter the data joining the "FRAIL" and "FUNCTIONAL DECLINE" classes into a single class.

The results of this analysis are represented in the Figure 7.2 and they show once again a large difference between the optimal cut-off point and the predefined cut-off point for the SF36 test (marked on red in the chart). The AUC for this curve is 0.682 which is very similar to the values of AUC for the different physical tests and it shows the inefficiency of the test in

discriminating normal and declining individuals, when they are evaluated with the tools of the 2nd screening with respect to the physical domain.

Figure 7.2: ROC curve for Physical Classification



In this ROC curve the very low negative predictive values also stand out, meaning that in the SF36 test many of the negative results (in this case subjects classified as normal) are false negatives, i.e., there are many people wrongly classified as normal.

For the 100 samples produced with the same number of robust and declining individuals, the result of AUC mean was 0.682 and the standard deviation for AUC was 0.021.

Cognitive domain:

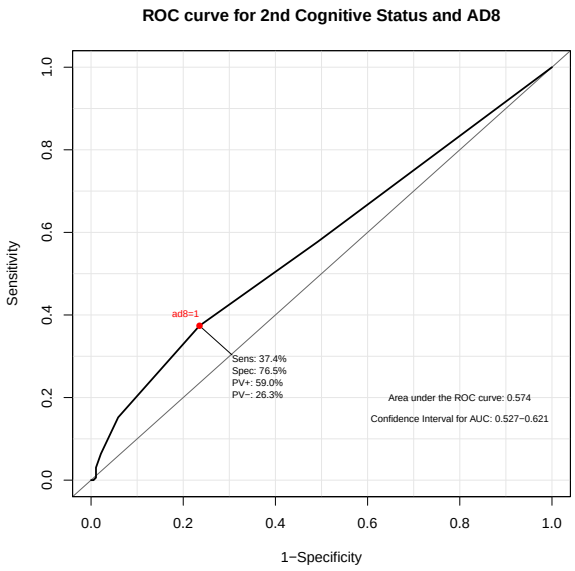
The results of ROC analysis for the cognitive domain are shown in Figure 7.3. The optimal cut-off point calculated coincides with the predefined cut-off point for 1st screening's classification in cognitive domain with AD8 test.

Although the cut-off point for AD8 test is the most appropriate according to this method, this produces a very low sensitivity which means that there are many participants with cognitive functional decline who are not identified as such by the AD8 test.

Furthermore, the AUC from the ROC curve is extremely low (0.574) indicating, once again, the poor performance of the test in discriminating people between the two cognitive classes.

Regarding the results from the balanced samples, the mean and standard deviation of AUC for all the ROC curves performed were recorded and they are, respectively, 0.573 and 0.014.

Figure 7.3: ROC curve for Cognitive Classification



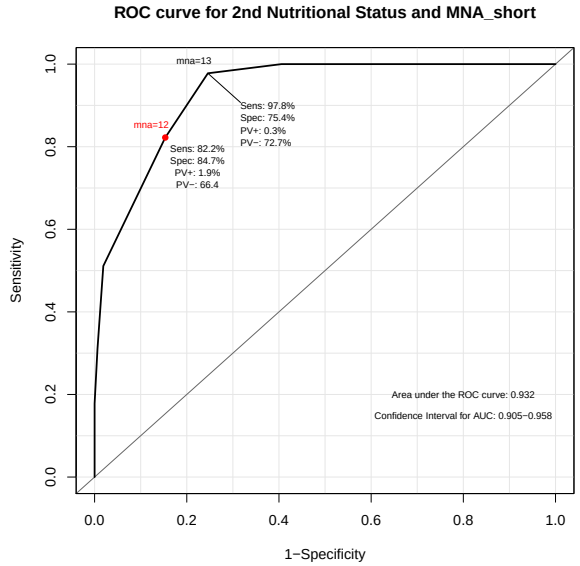
Nutritional domain:

The results of the ROC analysis on the nutritional domain are very different from the other domains. As it can be seen in the Figure 7.4 the optimal cut-off point is not the same as the predefined cut-off point for MNA (short version) in the 1st screening, but they are very close with only 1 unit of difference.

The value for AUC is 0.932 which is relatively high value for this measure and close to the

maximum 1. Usually this means that the test can discriminate with much effectiveness normal and in decline individuals with respect to the nutritional domain, but in this case this idea is questionable because the short version of MNA used in the 1st screening is part of the MNA test that classifies the individuals in the 2nd screening. As the second test includes the first and a few more questions, then the classification based on the two will obviously be very similar.

Figure 7.4: ROC curve for Nutritional Classification



It should also be noted that the positive predictive value for the different cut-off points is very low which means that this test produces many false positives, i.e., the short version of MNA classifies many individuals as declining in nutritional domain, when in true they are not in decline.

Regarding the samples' ROC analysis, the AUC mean of the curves was 0.929, while the standard deviation for the same measure was 0.017.

7.3 Logistic regression and multiple linear regression

After analyzing the results of the last section about how good the scores of the 1st screening can predict the classifications of the 2nd screening in each domain, it was decided to go into the detailed questions on the questionnaire to try to understand these results.

For this reason, two different statistical studies were done. The first one was the logistic regression with the classification for each domain of the 2nd screening as the response variable and the individual questions of the 1st screening for each domain as the explanatory variables.

The objective of this approach is not to predict the classification of the 2nd screening for new individuals, but to understand which questions of the 1st screening have more importance in the 2nd screening's classification process.

For the logistic regression models' estimation the function *glm()* of R was used. The option *family=binomial(link = "logit")* had to be included, since the function is for generalized linear models. Besides that, it was also necessary resort to *stepAIC()* function from the MASS package of R to produce more parsimonious models, that is with only the most significant variables, using the stepwise model selection by AIC.

In this analysis were also performed two distinct approaches for the models construction, like in the ROC analysis (section 7.2), one with the original data and other with 100 randomly selected balanced samples. The results of these two approaches for all domains are shown in the Tables 7.6. The most frequent variables for the samples were defined as those appearing in more than 50 models from the 100 performed. In addition to the variables listed in the tables, also gender and age group variables were taken into account in the models.

To a better understanding of the results, the description of the questions (predictors of the

Table 7.6: Results from the fitted logistic regression model

Domain (test)	Variables in the model	Most frequent variables
Physical (SF36)	4, 6	4, 6
Cognitive (AD8)	3, 4	3, 4
Nutritional (MNASF)	1, 2, 4, 5	1, 2, 4, 6b

model) should be checked in the section 2.3 of chapter 2 and they can also be observed in the completed questionnaire attached to this work.

The most frequent variables obtained in the analysis of the samples may change a little in each resampling, but for logistic regression it is observed that there are no major differences between the obtained variables from the original data and the samples.

The same study was done assuming the test scores of the 2nd screening for each domain as the response variable in a multiple linear regression. The function *lm()* of R was used to estimate these models and it was again necessary to resort to *stepAIC()* function in order to perform a stepwise selection. The results are recorded in Table 7.7.

Table 7.7: Results from the fitted multiple linear regression model

2nd test (1st test)	Variables in the model	Most frequent variables
TUGT (SF36)	2, 3, 5	5
CSRT (SF36)	6	6
CST (SF36)	1, 5, 7, 10	1, 5, 7
MST (SF36)	1, 5, 6, 7, 8	1, 5, 6, 7
QMCI (AD8)	1, 2, 3, 4, 5	4, 5
MNASF	1, 3, 4, 6a, 6b	1, 4

In the multiple linear regression the differences between the relevant variables obtained with the original data and the resampling procedure are bigger than in the logistic regression, since in the samples analysis there were much less variables. However, it can be observed that, in almost all cases, the most frequent variables from the resamples are in the group of variables of the models from original data.

Therefore, both in the logistic regression and in the multiple linear regressions there were no significant differences between the most relevant variables obtained with the complete data and the resampling procedure.

To conclude the comparative study between the 1st and the 2nd screenings, it was made a logistic regression with the 2nd screening final classification as the response variable and all the individual questions from the 1st screening questionnaire as the explanatory variables. We selected the variables which were more informative for classifying the individuals as pre-frail or robust in the 2nd screening.

The selected variables are presented below with the respetive test:

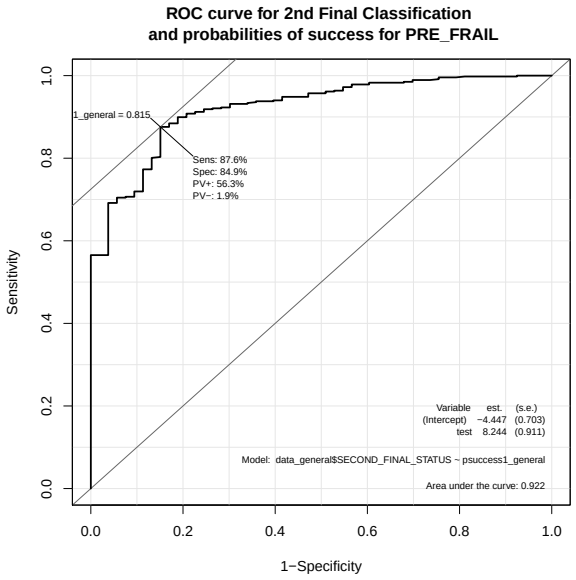
- **General:** 1 - Gender, 3 - Education level, 7 - Consumption of alcoholic beverages per week, 10 - Use of any soft drugs;
- **SF12:** In the past 4 weeks 5 - did you were limited in the kind of work or other activities (physical health)?, 9 - have you felt calm and peaceful?, 12 - how much of the time has your physical health or emotional problems interfered with your social activities?;

- **SF36:** 1 - Vigorous activities, 4 - Climbing several flights of stairs, 6 - Bending, kneeling or stooping, 9 - Walking one block, 10 - Bathing or dressing yourself;
- **GFI:** 1 - Can you perform grocery shopping?, 5 - How would you rate your own physical fitness?, 6 - Do you encounter problems in daily life because of impaired vision?, 7 - Do you encounter problems in daily life because of impaired hearing?;
- **MNASF:** During the last 3 months 1 - did you have loss of appetite, digestive problems, difficulty in chewing and/or swallowing?, 2 - did you have loss of weight?;
- **AD8:** 3 - Repeat the same things over and over, 4 - Have trouble learning how to use a tool, appliance, or gadget.

This procedure would be much more reliable and accurate if the individuals undergoing to the 2nd screening would have not been chosen preferably as pre-frail, but irrespectively of their final classification in the 1st screening.

Another result from a logistic regression are the estimates for the probability of success, which in this case is to be pre-frail. Based on these and defining a cut-off point above which individuals are classified as pre-frail, it is possible to define a new classification rule.

Figure 7.5: ROC curve for 2nd Final Classification



To choose the best cut-off point for this classification rule, a ROC curve was performed for the 2nd final classification, using the probabilities of being pre-frail as the diagnostic test. As it can be seen in Figure 7.5 the optimal cut-off point obtained was 0.815. With this cut-off point, an individual is classified as pre-frail if their probability of being pre-frail is above 0.815. Taking into account the AUC obtained, it is possible to conclude that this diagnostic test has an excellent performance.

Using the previous classification rule, all the individuals in the study were reclassified. In Table 7.8 are presented the comparison between the results of this classification and the 2nd final classification defined by PERSSILAA's protocol. In the creation of this new classification rule, the final classification from the 2nd screening was used as a gold standard. However, this study would have been more interesting and reliable if there was a random selection of individuals undergoing the 2nd screening, rather then giving preference to the pre-frail individuals.

Table 7.8: Results from the fitted logistic regression model for 2^{nd} final classification

2nd FINAL STATUS\LOGISTIC REGRESSION	PRE-FRAIL	ROBUST
PRE-FRAIL	366	101
ROBUST	7	46

To conclude this chapter, it is important to remember the key ideas that came from the analysis. Through the analysis of the ROC curve, it is noticed that the tests AD8 and SF36 of the 1^{st} screening are not good predictors of the condition verified in the same domains of the 2^{nd} screening. Regarding the MNA-SF, the analysis is not entirely conclusive, since the classification of the 2^{nd} screening is also done using questions of MNA-SF. In addition, the most important variables of each domain of the 1^{st} screening to the respective classification in the 2^{nd} screening were identified. Finally a new classification rule was constructed with the most relevant questions of the full questionnaire of the 1^{st} screening to the classification of the 2^{nd}, which seems to have a good performance.

Chapter 8

Conclusion

This report aimed to contextualize and describe the work developed and the results achieved during the statistical analysis of the data resulting from the PERSSILAA's screening process executed in Netherlands. The report starts with a short overview of the PERSSILAA project with special emphasis on the screening protocol. All the main concepts of PERSSILAA required for the understanding of the work, from the description of the variables in the study to the explanation of the classification rules, were summarized in Chapter 2.

In order to give a theoretical support to the statistical work carried out in this project, a description of the methods used in the processing and analysis of data was also done. Without going into too much detail and assuming the reader has some statistical background, some theory of regression models, cluster analysis, discriminant analysis and hypothesis tests used in the work were mentioned. Besides that the use of the ROC curve to evaluate the performance of a diagnostic test and a new approach for principal component analysis in mixed data were studied, explored and applied in this Master's project.

Before performing a statistical analysis on the data, it was necessary to validate the database. In the validation procedure, the computation of the scores and classifications in the platform was verified and some typical errors were spotted. Also, some inconsistencies were found in the database, including some mistakes in the insertion of the data in the platform and in the computation of some variables. These errors were corrected and a new database was created and used in the subsequent analysis.

Preliminary analyses of the data corresponding to the participants of the 1^{st} and 2^{nd} screenings suggested that the general characteristics, such as age and gender, were similar. In the 1^{st} screening the majority of the individuals were classified as robust, while in the 2^{nd} screening they were classified as pre-frail. Since the participants in the 2^{nd} screening are a subset of those participating in the 1^{st} screening and essentially individuals who were classified as pre-frail in the 1^{st} screening, this shows a certain agreement between the classification of the 1^{st} and 2^{nd} screenings. This issue was analyzed in more detail in Chapter 7.

The distribution of the final classification of the two screenings was compared between the populations of the different municipalities from the Netherlands who participated in the study. Regarding the final status of the 1^{st} screening, some significant differences were found between the municipalities, being more evident between Enschede and the other three municipalities. As these differences were found, there was interest in understanding what happened in each domain. It was then concluded that the pair Enschede/Hengelo showed significant differences in physical and cognitive domains, while Enschede/Tubbergen presented differences in physical and nutritional domains. With respect to the pair Enschede/Twenterand, only differences in the nutritional domain were found. Although significant differences between the municipalities were not found in the final classification of the 2^{nd} screening, some slight differences were found when restricting the analysis to specific domains. TUGT and MST were the tests from the physical domain which presented more significant differences in the 2^{nd} screening.

A large part of this work was focused on the analysis of the classification performed by the

1st screening. A characterization for each type of *persona* in the 1st screening process, based on the Chernoff faces, evidenced differences in the questionnaire of the individual classification profiles. The profiles with major differences in the questionnaire of the 1st screening were those who had physical decline and functional decline in at least one of the other domains. The characterization of the individual profiles also suggested that the ALG and SF12 questions (general questions of the 1st questionnaire) provide an important contribution to distinguish the individuals, although these characteristics were not taken into consideration while giving a final 1st screening classification as frail, pre-frail and robust.

The cluster analysis performed in this Master's project emphasized the difficulty of classifying the participants of the 1st screening in the "PRE-FRAIL" class, since this is the intermediate class. This difficulty was confirmed with the linear discriminant analysis and multinomial logistic regression, particularly in the specific discrimination between pre-frail and robust individuals. The classification based on the MLR showed to be closer to the classification of the 1st screening based on PERSSILAA's protocol than the classification resulting from the LDA, which suggests that a classification built using the individual results of the questionnaire and not only the scores is a better tool to determine the health status of the participants.

To complete the objective of this Master's project, which was to do an evaluation on the screening process of PERSSILAA project, and answer some questions raised during the study, a comparison between the two screenings was made. This analysis had two main targets of assessment: the performance of the 1st screening tools for each domain and the performance of the full questionnaire of the 1st screening. The ROC curves obtained revealed that the scores of SF36 and AD8 tests do not have good accuracy as diagnostic tests for the classification in each correspondent domain of the 2nd screening, while the score of MNA-SF had an excellent performance as diagnostic test for the nutritional classification of the 2nd screening. This conclusion was to be expected, since the test responsible for the nutritional classification of the 2nd screening MNA which contains all MNA-SF questions. Using Multiple Linear Regression and Logistic Regression, the variables from the tests of the 1st screening relating to a single domain more relevant to the respective classification of the 2nd screening were selected. In this analysis it was also shown that the variables which are relevant using the original unbalanced data (more pre-frail individuals) are similar to those with balanced samples. Finally, a logistic regression model using all the questions from the full questionnaire of the 1st screening as predictor variables to the final classification of the 2nd screening was fitted. In this last analysis, it was concluded that there is relevant information in the questionnaire of the 1st screening, besides the scores used, to predict the classification of the 2nd screening.

All the work described in this report was done essentially to assess the quality of the classification resulting from the screening process of PERSSILAA, but more statistical work could be done in this Master's project. A longitudinal study to assess the health status of the older adults during their participation in the training module was initially planned. However, this analysis was not performed, since the data was not available until the end of this work. In order to enhance the PERSSILAA project and demonstrate its benefits, further statistical studies could be done, such as:

- including the data from Italy, and other countries if possible, in the analysis to evaluate the applicability of the PERSSILAA's screening process to different countries in Europe, since the programme intends to be implemented in all European countries;

- taking more individuals classified as frail and robust in the 1st screening to the 2nd screening, for more reliable and less biased results;
- a longitudinal study to understand if the participants of the training module improve or maintain their health status, not progressing to frail, longer than the individuals who did not participate in the program.

Bibliography

[1] Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2014). Multivariate analysis of mixed data: The PCAmixdata R package. Bordeaux, France.

[2] Chernoff, H. (1973). The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*, 361 - 368.

[3] Conover, W.J. (1999). Contingency Tables. *Practical Nonparametric Statistics* (pp. 179 - 268) New York: John Wiley & Sons, Inc.

[4] Conover, W.J. (1999). Some Methods Based on Ranks. *Practical Nonparametric Statistics* (pp. 269 - 427) New York: John Wiley & Sons, Inc.

[5] Conover, W.J. (1999). Statistical Inference. *Practical Nonparametric Statistics* (pp. 68 - 122) New York: John Wiley & Sons, Inc.

[6] Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical clustering. *Cluster Analysis* (pp. 71 - 110). London: Wiley.

[7] Galvin JE, Roe CM, Powlishta KK, Coats MA, Muich SJ, Grant E, Miller JP, Storandt M, Morris JC, (2005). The AD8: a brief informant interview to detect dementia. *Neurology*, 65(4), 559-64.

[8] Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4), 857-871.

[9] Harrel, F. E. (2001). Binary Logistic Regression. *Regression Modeling Strategies* (pp. 215 - 268). Virginia: Springer.

[10] Harrel, F. E. (2001). General Aspects of Fitting Regression Models . *Regression Modeling Strategies* (pp. 11- 40). Virginia: Springer.

[11] Holm, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. textitScandinavian Journal of Statistics, 6(2), 65-70.

[12] Hosmer, D. W., & Lemeshow, S. (2000). Introduction to the Logistic Regression Model. *Applied Logistic Regression* (pp. 1 - 27). Canada: Wiley.

[13] Jobson, J. D. (1992). Cluster Analysis and Multidimensional Scaling. *Applied Multivariate Data Analysis* (pp. 483 - 602). New York: Springer-Verlag.

[14] Johnson, R. N., & Wichern, D. W. (2007). Clustering, Distance Methods, and Ordination. *Applied Multivariate Statistical Analysis* (pp. 671 - 756). United States of America: Pearson.

[15] Johnson, R. N., & Wichern, D. W. (2007). Discrimination and Classification. *Applied Multivariate Statistical Analysis* (pp. 575 - 670). United States of America: Pearson.

[16] Long, J. S. (1997). Continuous Outcomes: The Linear Regression Model. *Regression Models for Categorical and Limited Dependent Variables* (pp. 11 - 33). United States of America: SAGE Publications.

[17] Morrison, D. F. (1967). Classification by DiscriminantFuntions. *Multivariate Statistical Methods* (pp. 269 - 290). Pennsylvania: McGraw Hill International Editions.

[18] Pepe, M. S. (2003). The Receiver Operating Characteristic Curve. *The Statistical Evaluation of Medical Tests for classification and Prediction* (pp. 66 - 95). Oxford: Oxford University Press.

[19] Rodriguez, G. (2007). Lecture Notes on Generalized Linear Models. In http://data.princeton.edu/wws509/notes/

[20] Timm, N. H. (2002). Cluster Analysis and Multidimensional Scaling. *Applied Multivariate Analysis* (pp. 515 - 556). Pittsburgh: Springer.

[21] Wright, S. (1992). Adjusted P-Values for Simultaneous Inference. *Biometrics*, 48(4), 1005-1013.

Appendix A

Name: _____

Date: ____ - ____ - _____

General Practitioner: _____

Instructions for completing questionnaire

- Follow the instruction at the separate questions.
- Take the time to complete the questionnaire.
- Please read first the possible answers, before you answer the question.
- In most cased only one answer is allowed. Choose the answer that best fits your situation.
- Something you may enter more than one answer. This information is provided in the instruction of this question.
- It is possible that certain questions look alike.
- It is important that you complete all questions, even when question look alike or you find it difficult to give an answer.
- There are no right and wrong answers. Is concerns your opinion and experience.
- Have you completed the questionnaire? Please check if you filled out all questions.

General characteristics

1. **Are you male of female? [T1_ALG_01]**
   - 1☐ Male
   - 2☐ Female

2. **What is your date of birth? [T1_ALG_02]** ____ - ____ - _____

3. **What is your height? [T1_MNASF_06a]** _____

4. **What is your weight? [T1_MNASF_06b]** _____

5. **What is the highest level of education you have completed? [T1_ALG_03]**
   - 0☐ None
   - 1☐ Elementary School
   - 2☐ High school
   - 3☐ Lower vocation school
   - 4☐ Vocational school
   - 5☐ College
   - 6☐ University
   - 7☐ Other, namely _____

6. **What is your living situation? [T1_ALG_04]**
   - 1☐ Alone
   - 2☐ With partner and/or children

7. **Do you have at home a PC or laptop? [T1_ALG_05]**
   - 1☐ Yes
   - 2☐ No

8. **Do you have at home access to the Internet? [T1_ALG_06]**
   - 1☐ Yes
   - 2☐ No

9. How many alcoholic beverages do you consume on average per week? [T1_ALG_07]

_____

10. Are you responsible for you own (financial) administration? [T1_ALG_08]

1☐ Yes
2☐ No

11. Do you currently smoke? [T1_ALG_09]

1☐ Yes, How many cigarettes do you smoke per day on average? _____
2☐ No

12. Do you (occasionally) use any soft drugs (e.g. cannabis)? [T1_ALG_10]

1☐ Yes
2☐ No

Physical health

The following questions are about your physical health.

13. In general, would you say your health is? [T1_SF12_01]

5☐ Excellent
4☐ Very good
3☐ good
2☐ Fair
1☐ Poor

14. During the PAST 4 WEEKS, how much did PAIN interfere with your normal work (including both work outside the home and housework)? [T1_SF12_08]

5☐ Not at all
4☐ A little bit
3☐ Quite a bit
2☐ A lot
1☐ Extremely

15. To what extent are you able to move (mobility)? [T1_MNASF_03]

0☐ bed or chair bound
1☐ able to get out of bed / chair but does not go out
2☐ goes out

16. The following questions are about you daily activities. Does your health now limit you in these activities? If so, to what extent?

A. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports. [T1_SF36PF_01] / [T1_SF12_02]

1☐ Yes, limited a lot
2☐ Yes, limited a little
3☐ No, not limited at all

B. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf [T1_SF36PF_02]

1☐ Yes, limited a lot
2☐ Yes, limited a little
3☐ No, not limited at all

C. Lifting or carrying groceries. [T1_SF36PF_03]

1☐ Yes, limited a lot
2☐ Yes, limited a little
3☐ No, not limited at all

D. Climbing several flights of stairs. [T1_SF36PF_04]/ [T1_SF12_03]

1☐ Yes, limited a lot
2☐ Yes, limited a little
3☐ No, not limited at all

E. Climbing one flights of stairs [T1_SF36PF_05]

1☐ Yes, limited a lot
2☐ Yes, limited a little
3☐ No, not limited at all

F. Bending, kneeling or stooping [T1_SF36PF_06]

1☐ Yes, limited a lot
2☐ Yes, limited a little
3☐ No, not limited at all

G. Walking more than one mile [T1_SF36PF_07]

1☐ Yes, limited a lot

2☐ Yes, limited a little

3☐ No, not limited at all

H. Walking several blocks [T1_SF36PF_08]

1☐ Yes, limited a lot

2☐ Yes, limited a little

3☐ No, not limited at all

I. Walking one block [T1_SF36PF_09]

1☐ Yes, limited a lot

2☐ Yes, limited a little

3☐ No, not limited at all

J. Bathing or dressing yourself [T1_SF36PF_10]

1☐ Yes, limited a lot

2☐ Yes, limited a little

3☐ No, not limited at all

17. During the PAST 4 WEEKS have you had any of the following problems with your work or other regular activities AS A RESULT OF YOUR PHYSICAL HEALTH?

A. ACCOMPLISHED LESS than you would like: [T1_SF12_04]

1☐ Yes

2☐ No

B. Were limited in the KIND of work or other activities: [T1_SF12_05]

1☐ Yes

2☐ No

18. Do you encounter problems in daily life because of impaired vision? [T1_GFI_06]

1☐ Yes

0☐ No

19. Do you patient encounter problems in daily life because of impaired hearing? [T1_GFI_07]

1☐ Yes

0☐ No

20. How would the you rate your own physical fitness? (0-10 ; 0 is very bad, 10 is very good) [T1_GFI_05]

_____ (0-6 = 1 / 7-10 = 0)

21. Have you unintentionally lost a lot of weight in the past 6 months (6kg in 6 months or 3kg in 3 months)?[T1_GFI_08]

1☐ Yes

0☐ No

22. Did you the past 3 months starting with eating less as a result of loss of appetite, digestive problems, difficulty in chewing and / or swallowing? [T1_MNASF_01]

0☐ Significant (greatly reduced appetite)

1☐ A little (moderate loss of appetite)

2☐ No (no loss of appetite)

23. What is your loss of weight during the last 3 months? [T1_MNASF_02]

0☐ weight loss greater than 3 kg (6.6 lbs)

1☐ does not know

2☐ weight loss between 1 and 3 kg (2.2 and 6.6 lbs)

3☐ no weight loss

Mental health

The following questions are about you mental health.

24. During the PAST 4 WEEKS, were you limited in the kind of work you do or other regular activities AS A RESULT OF ANY EMOTIONAL PROBLEMS (such as feeling depressed or anxious)?

A. ACCOMPLISHED LESS than you would like: [T1_SF12_06]

1☐ Yes

2☐ No

B. Didn't do work or other activities as CAREFULLY as usual: [T1_SF12_07]

1☐ Yes

2☐ No

25. Has suffered psychological stress or acute disease in the past 3 months? [T1_MNASF_04]

0☐ Yes
2☐ No

26. Do you experience neuropsychological? [T1_MNASF_05]

0☐ severe dementia or depression
1☐ mild dementia
2☐ no psychological problems

27. Do you have any complaints on your memory? [T1_GFI_10]

0☐ No
0☐ Sometimes
1☐ Yes

28. Are you feeling down or depressed lately?? [T1_GFI_14]

0☐ No
1☐ Sometimes
1☐ Yes

29. Are you feeling nervous or anxious lately? [T1_GFI_15]

0☐ No
1☐ Sometimes
1☐ Yes

30. The next questions are about how you feel and how things have been DURING THE PAST 4 WEEKS. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the PAST 4 WEEKS…

A. Have you felt calm and peaceful? [T1_SF12_09]

6☐ all the time
5☐ most of the time
4☐ a good bit of the time
3☐ some of the time
2☐ a little of the time
1☐ never

B. Did you have a lot of energy?? [T1_SF12_10]

6☐ all the time
5☐ most of the time
4☐ a good bit of the time
3☐ some of the time
2☐ a little of the time
1☐ never

C. Have you felt downhearted and blue? [T1_SF12_11]

1☐ all the time
2☐ most of the time
3☐ a good bit of the time
4☐ some of the time
5☐ a little of the time
6☐ never

31. Do you experience any problems with judgement (e.g. problems making decisions, bad financial decisions, problems with thinking)? [T1_AD8_01]

1☐ Yes
0☐ No
0☐ Don't know

32. Do you experience less interest in hobbies and activities? [T1_AD8_02]

1☐ Yes
0☐ No
0☐ Don't know

33. Do you repeats the same things over and over (questions, stories, or statements)? [T1_AD8_03]

1☐ Yes
0☐ No
0☐ Don't know

34. Do you have trouble learning how to use a tool, appliance, or gadget (e.g., VCR, computer, microwave, remote control)? [T1_AD8_04]

1☐ Yes
0☐ No
0☐ Don't know

35. Do you forgets correct month or year? [T1_AD8_05]

1☐ Yes
0☐ No
0☐ Don't know

36. Do your have trouble handling complicated financial affairs (e.g., balancing checkbook, income taxes, paying bills)? [T1_AD8_06]

1☐ Yes
0☐ No
0☐ Don't know

37. Do you have trouble remembering appointments? [T1_AD8_07]

1☐ Yes
0☐ No
0☐ Don't know

38. Do you have <u>daily</u> problems with thinking and/or memory? [T1_AD8_08]

1☐ Yes
0☐ No
0☐ Don't know

<u>Relationships with others</u>

The following questions are about your relationships with others.

39. Do you ever experience emptiness around yourself? e.g. You feel so sad that you have no interest in your surroundings. [T1_GFI_11]

0☐ No
1☐ Sometimes
1☐ Yes

40. Do you ever miss the presence of other people around you? Or do you miss anyone you love? [T1_GFI_12]

0☐ No
1☐ Sometimes
1☐ Yes

41. Do you ever feel left alone? e.g. You wish there is someone to go with you for something important? [T1_GFI_13]

0☐ No
1☐ Sometimes
1☐ Yes

42. During the PAST 4 WEEKS, how much of the time has your PHYSICAL HEALTH OR EMOTIONAL PROBLEMS interfered with your social activities (like visiting with friends, relatives, etc.)? [T1_SF12_12]

1☐ all of the time
2☐ most of the time
3☐ a good bit of the time
4☐ some of the time
5☐ a little of the time
6☐ never

Independency

The following questions are about your independency

43. Can you perform the following tasks without assistance from another person (walking aids such as a can or a wheelchair are allowed)?

A. Grocery shopping [T1_GFI_01]

0☐ Yes
1☐ No

B. Walk outside house (around house or to neighbour) [T1_GFI_02]

0☐ Yes
1☐ No

C. Getting (un)dressed [T1_GFI_03]

0☐ Yes

1☐ No

D. Visiting the restroom [T1_GFI_04]

0☐ Yes

1☐ No

Demand of (health)care

The following questions are about your demand of (health)care

44. Do you take 4 or more different types of medication?? [T1_GFI_09]

1☐ Yes

0☐ No

45. At which healthcare professional are you under treatment or receive care? (multiple answers possible) [T1_IM_19] 0/1☐ GP 0 = no 1 = yes

0/1☐ doctor at a nursing home 0 = no 1 = yes

0/1☐ a specialist (e.g. pulmonologist, cardiologist, surgeon) for physical complaints 0 = no 1 = yes

0/1☐ multiple specialist for physical complaints 0 = no 1 = yes

0/1☐ psychologist 0 = no 1 = yes

0/1☐ dietitian 0 = no 1 = yes

0/1☐ social worker 0 = no 1 = yes

0/1☐ physiotherapist 0 = no 1 = yes

0/1☐ speech therapist 0 = no 1 = yes

0/1☐ nurse home care 0 = no 1 = yes

0/1☐ nurse at GP 0 = no 1 = yes

0/1☐ nurse at hospital 0 = no 1 = yes

0/1☐ carers at nursing home 0 = no 1 = yes

0/1☐ last month I was hospitalized or released from the hospital or nursing home 0 = no 1 = yes

0/1☐ I receive no care 0 = no 1 = yes

0/1☐ Other _____ 0 = no 1 = yes

Finally

1. **Would you like to fill in the date you completed this questionnaire [T1_TS_01]**

   ____ - ____ - _____

2. **Has anyone helped you in completing this questionnaire? [T1_TS_02]**

   1☐ Yes, someone helped me to complete this questionnaire.

   0☐ No, I have completed the questionnaire independently → **You are done!**

3. **If so, what was the help? [T1_TS_03]**

   ☐ 1 Someone else noted the answers; I chose the answers themselves

   ☐ 2 I chose the answers with someone and noted

   ☐ 3 Someone has the answers chosen for me and noted

4. **If you were assisted in completing the questionnaire or the questionnaire was completed by another, who was this person? [T1_TS_04]**

   ☐ 1 Partner

   ☐ 2 Family

   ☐ 3 Caregiver

   ☐ 4 Researcher

   ☐ 5 other, namely _____

5. **Space for additional comments: [T1_TS_05]**

_____

_____

Scroll a bit by the questionnaire. Have you completed all the questions? Then you are done! You can return the questionnaire in the enclosed envelope or handed in at the assistant of your GP.

Thanks for you cooperation.

Disclaimer: This questionnaire is developed within the PERSSILAA project (FP7-ICT-610359). This project is financed by the European Union. The following validated questionnaires are part of this questionnaire:

- Groninger Frailty Indicator (GFI)
  - Steverink, N., Slaets, J.P.J., Schuurmans, H., & Lis, M. van (2001). Measuring frailty: development and testing of the Groningen Frailty Indicator (GFI). The Gerontologist (41, special issue 1), 236-237.
- SF-12
  - Botterweck, A., Frenken, F., Janssen, S., Rozendaal, L., De Vree, M., & Otten, F. (2001). Plausibiliteit nieuwe metingen algemene gezondheid en leefstijlen 2001. Heerlen: Centraal Bureau voor de Statistiek.
- AD8
  - Galvin, J.E., et al., Patient's rating of cognitive ability: using the AD8, a brief informant interview, as a self-rating tool to detect dementia. Archives of neurology, 2007. 64(5): p. 725-30.
  - Galvin, J.E., et al., The AD8: a brief informant interview to detect dementia. Neurology, 2005. 65(4): p. 559-64
- RAND-36 Physical Functioning
  - Zee KI van der, Sanderman R. Het meten van de algemene gezondheidstoestand met de RAND-36, een handleiding. Groningen: Rijksuniversiteit Groningen, Noordelijk Centrum voor Gezondheidsvraagstukken; 1992
- Mini Nutrition Assessmetn Short-Form (MNA-SF)
  - Rubenstein LZ, Harker JO, Salva A, Guigoz Y, Vellas B. Screening for Undernutrition in Geriatric Practice: Developing the Short-Form Mini Nutritional Assessment (MNA-SF) . J. Geront 2001; 56A: M366-377

Appendix B

Description of the variables from the PERSSILAA's database

ID_USER - numeric value that identifies uniquely a subject;
SURVEY - numeric value that identifies the round of the 1st screening did by the subject (0-1st round (2014), 1-2nd round (2015));
SCREENING - categorical variable that shows the way the subject performed the 1st screening questionnaire (on paper, on line);
MUNICIPALITY - categorical variable with the municipality of the subject (Enschede, Hengelo, Tubbergen, Twenterand);
T1_ALG_01 - numeric value that indicates the gender of the subject (1-male, 2-female);
T1_ALG_02 - date value that indicates the birthday of the subject (format: yyyy-mm-dd);
AGE - numeric value that indicates the age of the subject;
T1_ALG_04 - numeric value that identifies the living situation (1-alone, 2-with partner and/or children);
T1_ALG_03 - numeric value that identifies the education level's of the subject (0-none, 1-elementary school, 2-high school, 3-lower vocation school, 4-vocational school, 5-college, 6-university, 7-other);
T1_ALG_06 - numeric value that indicates if the subject have or don't have access to the Internet at home (1-Yes, 2-No);
T1_ALG_07 - numeric value that indicates the average of alcoholic beverages consumed per week;
T1_ALG_08 - numeric value that identifies if the subject is responsible for their own (financial) administration (1-Yes, 2-No);
T1_ALG_09a - numeric value that indicates if the subject smokes (1-Yes, 2-No);
T1_ALG_09b - numeric value that indicates the number of cigarettes that the subject smokes per day on average;
T1_ALG_10 - numeric value that indicates if the subject uses any soft drugs (1-Yes, 2-No);
T1_SF12_01 - numeric value that indicates the subject's opinion about their health (1-poor, 2-fair, 3-good, 4-very good, 5-excellent);
T1_SF36PF_01_T1_SF12_02 - numeric value that identifies the subject's limitation to do vigorous activities (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_04_T1_SF12_03 - numeric value that identifies the subject's limitation at climbing several flights of stairs (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF12_04 - numeric value that indicates if the subject accomplished less in the past 4 weeks as a result of their physical health (1-Yes, 2-No);
T1_SF12_05 - numeric value that indicates if the subject were limited in the KIND of work or other activities in the past 4 weeks as a result of their physical health (1-Yes, 2-No);
T1_SF12_06 - numeric value that indicates if the subject accomplished less in the past 4 weeks as a result of any emotional problem (1-Yes, 2-No);
T1_SF12_07 - numeric value that indicates if the subject didn't do work or other activities as CAREFULLY as usual in the past weeks as a result of any emotional problem (1-Yes, 2-No);
T1_SF12_08 - numeric value that indicates if any pain on the subject interfered with their normal work during the past 4 weeks (1-Yes, 2-No);

T1_SF12_09 - numeric value that indicates how much of the time the subject felt calm and peaceful during the past 4 weeks (1-never, 2-a little of the time, 3-some of the time, 4-a good bit of the time, 5-most of the time, 6-all the time);
T1_SF12_10 - numeric value that indicates if the subject had a lot of energy in the past 4 weeks (1-never, 2-a little of the time, 3-some of the time, 4-a good bit of the time, 5-most of the time, 6-all the time);
T1_SF12_11 - numeric value that indicates if the subject felt downhearted and blue in the past 4 weeks (1-never, 2-a little of the time, 3-some of the time, 4-a good bit of the time, 5-most of the time, 6-all the time);
T1_SF12_12 - numeric value that indicates how much of the time the subject's physical health or emotional problems interfered with their social activities (1-never, 2-a little of the time, 3-some of the time, 4-a good bit of the time, 5-most of the time, 6-all the time);
SF_12_PCS - numeric value that indicates the 1st score of SF12 questionnaire;
SF_12_MCS - numeric value that indicates the 2nd score of SF12 questionnaire;
T1_SF36PF_01_T1_SF12_02.1 - numeric value that identifies the subject's limitation to do vigorous activities (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_02 - numeric value that identifies the subject's limitation to do moderate activities (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_03 - numeric value that identifies the subject's limitation to lifting or carrying groceries (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_04_T1_SF12_03.1 - numeric value that identifies the subject's limitation to climbing several flights of stairs (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_05 - numeric value that identifies the subject's limitation to climbing one flights of stairs (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_06 - numeric value that identifies the subject's limitation to bending, kneeling or stooping (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_07 - numeric value that identifies the subject's limitation to walking more than one mile (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_08 - numeric value that identifies the subject's limitation to walking several blocks (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_09 - numeric value that identifies the subject's limitation to walking one block (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
T1_SF36PF_10 - numeric value that identifies the subject's limitation to bathing or dressing yourself (1-Yes,limited a lot, 2-Yes,limited a little, 3-No,not limited at all);
SF_36_SCORE - numeric value that indicates the total score of SF36 questionnaire;
T1_GFI_01 - numeric value that indicates if the subject can perform grocery shopping without assistance from another person (1-Yes, 2-No);
T1_GFI_02 - numeric value that indicates if the subject can walk outside house without assistance from another person (1-Yes, 2-No);
T1_GFI_03_T1_KATZ_02 - numeric value that indicates if the subject can getting (un)dressed without assistance from another person (1-Yes, 2-No);
T1_GFI_04_T1_KATZ_03 - numeric value that indicates if the subject can visiting the restroom without assistance from another person (1-Yes, 2-No);
T1_GFI_05 - numeric value that indicates how would the subject rate their own physical fitness (0-6 = 1 / 7-10 = 0);

T1_GFI_06 - numeric value that indicates if the subject encounters problems in daily life because of impaired vision (1-Yes, 2-No);
T1_GFI_07 - numeric value that indicates if the subject encounters problems in daily life because of impaired hearing (1-Yes, 2-No);
T1_GFI_08 - numeric value that indicates if the subject unintentionally lost a lot of weight in the past 6 months (1-Yes, 2-No);
T1_GFI_09 - numeric value that indicates if the subject takes 4 or more different types of medication (1-Yes, 2-No);
T1_GFI_10 - numeric value that indicates if the subject has any complaints on your memory (0-No, 0-Sometimes, 1-Yes);
T1_GFI_11 - numeric value that indicates if the subject ever experience emptiness around yourself (0-No, 1-Sometimes, 1-Yes);
T1_GFI_12 - numeric value that indicats if the subject ever misses the presence of other people around their or miss anyone they love (0-No, 1-Sometimes, 1-Yes);
T1_GFI_13 - numeric value that indicates if the subject ever feels left alone (0-No, 1-Sometimes, 1-Yes);
T1_GFI_14 - numeric value that indicates if the subject is feeling down or depressed lately (0-No, 1-Sometimes, 1-Yes);
T1_GFI_15 - numeric value that indicates if the subject is feeling nervous or anxious lately (0-No, 1-Sometimes, 1-Yes);
GFI_SCORE - numeric value that indicates the total score of GFI questionnaire;
T1_MNASF_01 - numeric value that indicates if the subject started to eating less as a result of loss of appetite, digestive problems, difficulty in chewing and / or swallowing in the last 3 months (0-Significant (greatly reduced appetite), 1-A little (moderate loss of appetite), 2-No (no loss of appetite));
T1_MNASF_02 - numeric value that indicates the subject's loss of weight during the last 3 months (0-weight loss greater than 3 kg (6.6 lbs), 1-does not know, 2-weight loss between 1 and 3 kg (2.2 and 6.6 lbs), 3-no weight loss);
T1_MNASF_03 - numeric value that indicates the extent that the subject is able to move (0-bed or chair bound, 1-able to get out of bed / chair but does not go out, 2-goes out);
T1_MNASF_04 - numeric value that indicates if the subject has suffered psychological stress or acute disease in the past 3 months (0-Yes, 2-No);
T1_MNASF_05 - numeric value that indicates if the subject experiences neuropsychological (0-severe dementia or depression, 1-mild dementia, 2-no psychological problems);
T1_MNASF_06a - numeric value that indicates the subject's height (kg);
T1_MNASF_06b - numeric value that indicates the subject's weight (cm);
MNA_short_SCORE - numeric value that indicates the total score of MNA questionnaire;
T1_IM_19_a - numeric value that indicates if the subject are under treatment or receive care with a GP (0-No, 1-Yes);
T1_IM_19_b - numeric value that indicates if the subject are under treatment or receive care with a doctor at a nursing home (0-No, 1-Yes);
T1_IM_19_c - numeric value that indicates if the subject are under treatment or receive care with a specialist for physical complaints (0-No, 1-Yes);
T1_IM_19_d - numeric value that indicates if the subject are under treatment or receive care with multiple specialist for physical complaints (0-No, 1-Yes);

T1_IM_19_e - numeric value that indicates if the subject are under treatment or receive care with a psychologist (0-No, 1-Yes);
T1_IM_19_f - numeric value that indicates if the subject are under treatment or receive care with a dietitian (0-No, 1-Yes);
T1_IM_19_g - numeric value that indicates if the subject are under treatment or receive care with a social worker (0-No, 1-Yes);
T1_IM_19_h - numeric value that indicates if the subject are under treatment or receive care with a physiotherapist (0-No, 1-Yes);
T1_IM_19_i - numeric value that indicates if the subject are under treatment or receive care with a speech therapist (0-No, 1-Yes);
T1_IM_19_j - numeric value that indicates if the subject are under treatment or receive care with a nurse home care (0-No, 1-Yes);
T1_IM_19_k - numeric value that indicates if the subject are under treatment or receive care with a nurse at GP (0-No, 1-Yes);
T1_IM_19_l - numeric value that indicates if the subject are under treatment or receive care with a nurse at hospital (0-No, 1-Yes);
T1_IM_19_m - numeric value that indicates if the subject are under treatment or receive care with carers at nursing home (0-No, 1-Yes);
T1_IM_19_n - numeric value that indicates if the subject are under treatment or receive care because last month they were hospitalized or released from the hospital or nursing home (0-No, 1-Yes);
T1_IM_19_o - numeric value that indicates if the subject aren't under treatment or don't receive any care (0-No, 1-Yes);
T1_IM_19_p - string that indicates other healthcare professional that are treating or giving care to the subject;
IM_SCORE - numeric value that indicates the total score of IM questionnaire;
T1_BMI - numeric value that indicates the body mass index of the subject;
BMI_SCORE - numeric value that indicates body mass index codification of the subject;
T1_AD8_01 - numeric value that indicates if the subject experiences any problems with judgement (1-Yes, 0-No, 0-Don't know);
T1_AD8_02 - numeric value that indicates if the subject experiences less interest in hobbies and activities (1-Yes, 0-No, 0-Don't know);
T1_AD8_03 - numeric value that indicates if the subject repeats the same things over and over (1-Yes, 0-No, 0-Don't know);
T1_AD8_04 - numeric value that indicates if the subject has trouble learning how to use a tool, appliance, or gadget (1-Yes, 0-No, 0-Don't know);
T1_AD8_05 - numeric value that indicates if the subject forgets correct month or year (1-Yes, 0-No, 0-Don't know);
T1_AD8_06 - numeric value that indicates if the subject has trouble handling complicated financial affairs (1-Yes, 0-No, 0-Don't know);
T1_AD8_07 - numeric value that indicates if the subject has trouble remembering appointments (1-Yes, 0-No, 0-Don't know);
T1_AD8_08 - numeric value that indicates if the subject has daily problems with thinking and/or memory (1-Yes, 0-No, 0-Don't know);
AD8_SCORE - numeric value that indicates the total score of AD8 questionnaire;

FIRST_PHYSICAL_STATUS - categorical variable that indicate the physical status of the subject in the 1st screening (decline or normal);
FIRST_COGNITIVE_STATUS - categorical variable that indicate the cognitive status of the subject in the 1st screening (decline or normal);
FIRST_NUTRITIONAL_STATUS - categorical variable that indicate the nutritional status of the subject in the 1st screening (decline or normal);
FIRST_GENERAL_STATUS - categorical variable that indicate the general status of the subject in the 1st screening (frail, pre-frail, robust);
FIRST_FINAL_STATUS - categorical variable that indicate the final status of the subject in the 1st screening (FRAIL, PRE-FRAIL, ROBUST);
T2_QMCI_01a - numeric value that punctuates the subject's answer to the question "What country is this?" in Orientation test (2-correct answer, 1-incorrect answer, 0-no attempt);
T2_QMCI_01b - numeric value that punctuates the subject's answer to the question "What year is this?" in Orientation test (2-correct answer, 1-incorrect answer, 0-no attempt);
T2_QMCI_01c - numeric value that punctuates the subject's answer to the question "What month is this?" in Orientation test (2-correct answer, 1-incorrect answer, 0-no attempt);
T2_QMCI_01d - numeric value that punctuates the subject's answer to the question "What is today's date?" in Orientation test (2-correct answer, 1-incorrect answer, 0-no attempt);
T2_QMCI_01e - numeric value that punctuates the subject's answer to the question "What day of the week is this?" in Orientation test (2-correct answer, 1-incorrect answer, 0-no attempt);
T2_QMCI_02 - numeric value that represents the number of correct words repeated by the subject in the Word Registration test (max=5);
T2_QMCI_03a - numeric value that represents the number of correct numbers drawn by the subject in the Clock Drawing test (max=12);
T2_QMCI_03b - numeric value that represents the number of correct clock hands drawn by the subject in the Clock Drawing test (max=2);
T2_QMCI_03c - numeric value that indicates if the pivot was drawn by the subject in the right position in the Clock Drawing test (Yes-1, No-0);
T2_QMCI_04 - numeric value that represents the number of correct words repeated by the subject in the Delayed Recall test (4 points per word for QMCI_SCORE);
T2_QMCI_05 - numeric value that represents the number of correct animals said by the subject the Verbal Fluency test, until 40 (0.5 points per animal for QMCI_SCORE);
T2_QMCI_06 - numeric value that punctuates the subject's story in the Logical Memory test, giving 2 points per each highlighted word exactly recalled (max= 30);
QMCI_SCORE - numeric value that indicates the total score of QMCI questionnaire;
T2_PHY_TUGT_01 - numeric value that represents the time in seconds that the subject took to get out of the chair in the Balance: timed up and go test;
T2_PHY_TUGT_02 - numeric value that represents the time in seconds that the subject took walking in the Balance: timed up and go test;
T2_PHY_TUGT_03 - numeric value that represents the time in seconds that the subject took to back in the Balance: timed up and go test;
TUGT_SCORE - numeric value that represents the average of the last 3 variables;
T2_PHY_CST - numeric value that represents the number of stands did by the subject in the Strength: chair-stand test;
CST_SCORE - numeric value that represents the number of stands did by the subject in the

Strength: chair-stand test;
T2_PHY_CSRT - numeric value that indicates the cm reached by the subject in the Flexibility: chair sit and reach test;
CSRT_short_SCORE - numeric value that indicates the cm reached by the subject in the Flexibility: chair sit and reach test;
T2_PHY_MST - numeric value that indicates the number of steps did by the subject in the Endurance: two-minute step test;
MST_SCORE - numeric value that indicates the number of steps did by the subject in the Endurance: two-minute step test;
T2_MNA_g - numeric value that identifies if the subject lives independently (1-Yes, 0-No);
T2_MNA_h - numeric value that identifies if the subject takes more than 3 prescription drugs per day (0-Yes, 1-No);
T2_MNA_i - numeric value that identifies if the subject have pressure sores or skin ulcers (0-Yes, 1-No);
T2_MNA_j - numeric value that indicates the number of subject's full meals daily (0-1 meal, 1-2 meals, 2-3 meals);
T2_MNA_k1 - numeric value that identifies if the subject eats at least one serving of dairy products (1-Yes, 0-No);
T2_MNA_k2 - numeric value that identifies if the subject eats two or more servings of legumes and eggs per week (1-Yes, 0-No);
T2_MNA_k3 - numeric value that identifies if the subject eats meat, fish or poultry every day (1-Yes, 0-No);
T2_MNA_l - numeric value that identifies if the subject consumes two or more servings of fruit or vegetables per day (1-Yes, 0-No);
T2_MNA_m - numeric value that indicates the number of fluid consumed by the subject per day (0.0-less than 3 cups, 0.5-3 to 5 cups, 1.0-more than 5 cups);
T2_MNA_n - numeric value that indicates the subject's mode of feeding (0-unable to eat without assistance, 1-self-fed with some dificulty, 2-self-fed without any problem);
T2_MNA_o - numeric value that representes the self view of nutritional status (0-views self as being malnourished, 1-is uncertain of nutritional state, 2-views self as having no nutritional problem);
T2_MNA_p - numeric value that representes the subject's perception of their health status compared with other people of the same age (0.0-not as good, 0.5-does not know, 1-as good, 2-better);
T2_MNA_q - numeric value that indicates the mid-arm circumference of the subject in cm (0.0-MAC less than 21, 0.5-MAC 21 to 22, 1-MAC greater than 22);
T2_MNA_r - numeric value that indicates the calf circumference of the subject in cm (0-CC less than 31, 1-CC 31 or greater);
T2_PHY_BUIK - numeric value that indicates the belly size of the subject in cm;
MNA_SCORE - numeric value that indicates the total score of MNA questionnaire;
SECOND_PHYSICAL_STATUS - categorical variable that indicate the physical status of the subject in the 2nd screening (decline or normal);
SECOND_COGNITIVE_STATUS - categorical variable that indicate the cognitive status of the subject in the 2nd screening (decline or normal);
SECOND_NUTRITIONAL_STATUS - categorical variable that indicate the nutritional status of the subject in the 2nd screening (decline or normal);

SECOND_FINAL_STATUS - categorical variable that indicate the final status of the subject in the 2nd screening (FRAIL, PRE-FRAIL, ROBUST).