POLITECNICO DI TORINO

Master degree course in Computer Engineering

Master Degree Thesis

Health Records Analysis to build Final Patient dataset for dementia disease

Supervisor:

Prof. Paolo Garza

Prof. Ernestina Menasalvas Ruiz

Candidate:

Gaetano Ferrara

matricola: 254904

July 2020

Abstract

Big Data and data mining technologies are becoming an essential technology in everyday life. Banking use this tools to monitor the financial market through network activity monitors to minimize fraudulent transition. Automotive to reduce the implementation of opt having an underperforming market. The Healthcare sector, however, still has not benefitted from the wide-scale use of Big Data technologies. The reason for it is Healthcare sector has been traditionally slow in adopting new ICT techniques because most clinical data was stored in paper form. The distribution of Electronic health record (EHR) in healthcare facilities, it was an important step towards digitalization of health.

Using the current hospital IT system is impossible to use data to make the clinical decisions because more of that has been collected in a wrong way or rather to achieve operational purposes. Dementia disease is one of the most widespread disease in the world among older people. Someone in the world develops dementia every 3 seconds. Around 50 million people have dementia, and there are nearly 10 million new cases every year. the most common form of dementia is Alzheimer disease that is associated with 60-70 % of cases. The focus of this master thesis is dementia disease using anonymized data collecting various information during the disease progression. We present a data-driven approach to build the main dataset called Final Patient Data and with it define a set of three key performance indicators that it a preliminary step of a more complex machine learning analysis: 1) Multiple Filter Analysis 2) identification of strong correlation 3) rate of progression after the first diagnosis.

In order to achieve the goal this thesis presents a method to be able, first of all, to extract processes of the patient from unstructured data sets. To structure the project and make it replicable, CRISP-DM has been adopted as a methodology to fulfill the goals. The thesis also presents methods to analyze, clean and prepare data to obtain structured datasets from which the mentioned KPIs can be measured.

Acknowledgements

I would like to acknowledge Ernestina Menasalvas Ruiz and the entire CTB team who supported the development of the research performed in order to reach the conclusion of this thesis. It was a pleasure working with you.

Contents

List of Figures

List of Tables

1 Introduction and Motivation

1.1 Introduction

Data are one of the most useful resources for a huge variety of productive sectors. Data into computer science meaning is an elementary description of an information, an entity, a phenomenon or an event[4].

With the advent of technology, there has been an impressive increase of data and new branches of science were born with the goal to exploit it. An IDC report[5] said that global data traffic will reach 175 zettabytes by 2025. This uncontrollable data traffic increase is due to the spread of smart devices. In 2020 connected smart devices are 30 billion. they are estimated to be 75 billion by 2025 thanks to a trend in constant and exponential growth.

Another IDC report said that now all production sectors make full use of data to carry out analysis, statistical studies and so on. This report said that health care is growing the usage of data science tools, infact it is assumed that there will be 36% growth in generated healthcare data by 2025. The growth in accumulated data is due to the integration of IT systems within hospitals and the use of the Electronic Health Record (EHR).

This increase in the healthcare sector has spread the belief of using it to extract new knowledge from the data. The idea is to exploit all the tools provided by big data, machine learning and deep learning techniques in order to provide a powerful mean to improve the diagnoses, the treatments provided and their effectiveness for identifying the risk factors associated to a specific type of illness or to translate the aquired knowledge into new hospital policies to reduce costs and waste of resources.

One of the many diseases that we want to study using the methods provided by data science is dementia that is an extremely common disease spread through the human population. In Italy in 2019 there are 1,271,000 people suffering from dementia which in 2050 it is expected to double and in the world it is estimated to have 48.8 million dementia sufferers, where more than 60% are affected by Alzheimer's disease.

The promoter of this study is the Oxford Project to investigate Memory and Aging (OPTIMA)[6] that has the aim of identifying which are the still unknown causes of dementia and in particular Alzheimer.

The guidelines published by the World Health Organization (WHO)[7] on May 14 2019 emphasize how a change in lifestyles (such as smoking habits, excessive alcohol consumption, unbalanced nutrition) or the control of certain diseases (such as hypertension, diabetes, obesity, depression, hypercholesterolemia) and factors that are not strictly sanitary (such as social isolation and cognitive stimuli) may be implicated in the onset of dementia and, in general, of cognitive impairment.

This master thesis was developed as part of the OPTIMA project in order to build a reference dataset that have been called Patient Final Data and carry out a set of analysis on it.

1.2 State of The Art

In the past, several studies have been done about dementia and its influencing and risk factors. During this study, a different type of features that are perfectly associated with these past studies have been considered .

In the study from E.S Sharp e.t [8] has been studied the relatioship between Dementia and education where has been identified that lower education is associated with a greater risk for dementia in many but not all patients. Another study from C.C. Liu e.t [9] showed that a specific type of gene, allele E4, provide clinical evidence of the relationship between it and the spread of Alzheimer's disease into the population. Although the presence of APOE E4 does not necessarily entail disease development, this genetic isoform probably accelerates the rate of disease conversion and progression.

Lastly, study from T.B. Chen e.t. [10] analyzed the relation between dementia and comorbidities. Regardless of the cognitive condition, people that have more than 60 years have at least one comorbidity disease. The study shows that people who have more than 60 years and at least three different types of comorbidities, have higher possibility to developing cognitive impairment problems. All the features used in these studies have been included in the Patients Final Data created in this project.

1.3 Goals

Starting from unstructured clinical data it will be used Data Science techniques to obtain the following goals:

- Multiple Filter Analysis
- Correlation Analysis on Patients Final Data
- Rate of progression after first diagnosis

In order to fulfil these goals, the following steps must be followed:

- Exploration OPTIMA Data Source
- Extraction patients information from the relevant files into OPTIMA Data Source
- Doctor's rules application.

1.4 Structure of Contents

Project has been divided into the following chapters:

- Chapter 2 Theoretical Framework: It describes the background of this study and explains the main tools used to developed it. Furthermore it describes the primary tools used to develop a Data Science Project and which is the main reason of the study.
- Chapter 3 Materials and Methods: A general overview of the raw OPTIMA Data Source and the process, called Data Pipeline,to build the Patient Final data is describe.
- Chapter 4 Proposed approach to extract pattern of dementia evolution: This is the main chapter on this project.This Chapter, starting from the business project analysis, will describe the primary Key Performance Indicators(KPI) developed.
  Then there is a complete description about Raw Data and the approach used to extract data from it to create Patients Final Data. Furthermore, the Doctor's rules have been described and how to apply it and finally there is a phase of validation of results.
- Chapter 5 Discussion of Results : Chapter describe how the KPI defined into the previous chapter, are been computed and then a discussion of results obtained.

- Chapter 6 Conclusions and Future Lines of Research: We presented the main conclusions and we have been proposed possible ways to improve the accuracy of the result obtained.

2 Theoretical Framework

In this chapter detailed description of dementia disease and what is Optima project have been presented to understand which are the main goals of this ambitious project. Furthermore, During the second part of the chapter, a theoretical explanation has been shown about the key performance indicators and how they have been defined and finally how a data science project is structured.

2.1 Dementia

Dementia is a descriptive term indicating an observable decline in mental abilities. It is an acquired clinical syndrome characterised by deterioration of mental functioning in its cognitive, emotional and conative aspects[11].
In the collective imagination, dementia is directly associated to Alzheimer's disease but this idea is totally wrong. Infact there are over 100 diseases that may cause dementia each one associated with a set of specific causes. In general, it results from neurons degeneration due to disturbances of body systems
The following list shows which are the common causes of dementia :

- Alzheimer's disease: Type of dementia that is totally irreversible. Alzheimer outlines a progressive brain disorder that destroys memory and thinking skills.

- Parkinson's disease: The disease damages nerve cells producing dopamine into the black substance area of the brain. Dopamine has many different tasks to do, such as impacting thought processes, mental functions and memory management.

- Vascular dementia: It is a cognitive impairment problem caused by brain damage due to impaired blood flow to the brain. The main causes of vascular dementia are diseases like stroke that blocks an artery in the brain or conditions that damage blood vessels or blood circulation.

- Frontotemporal dementia: Type of dementia associated with a progressive nerve cell loss into frontal and temporal lobes. Normally FTD leads to loss of important cognitive functions.

- Cortical Lewy Body dementia: Type of dementia normally associated with a Lewy bodies disease.

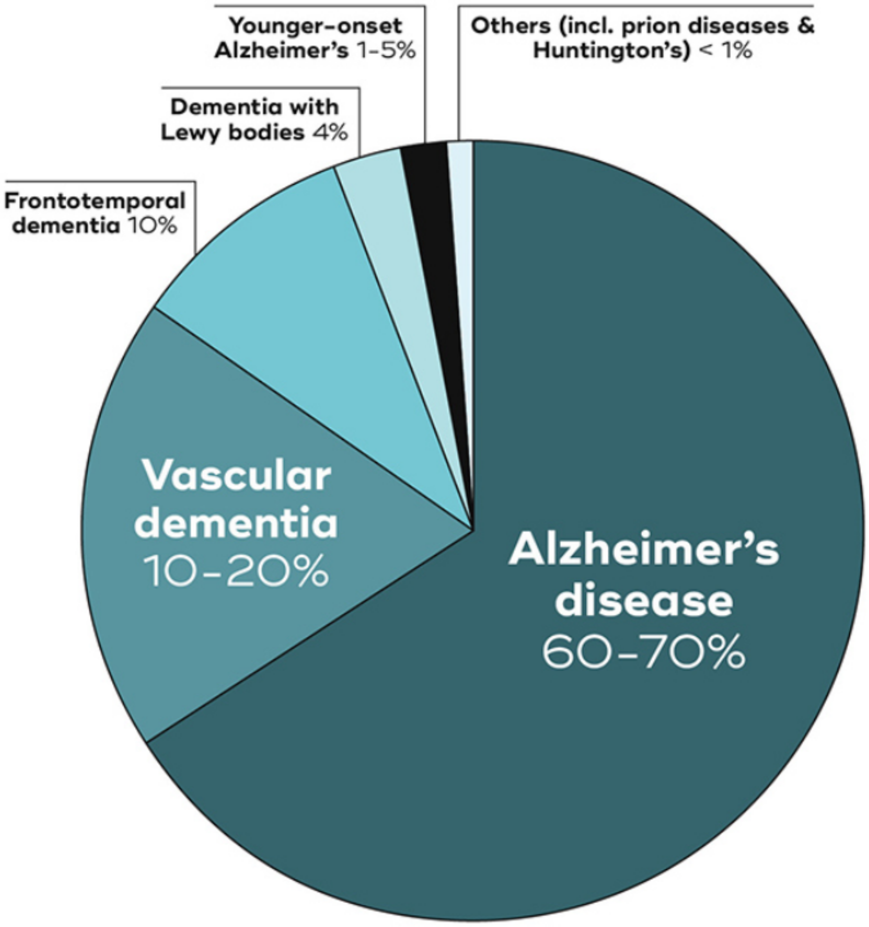


Figure 1: Distribution Dementia Types [1]

Among the listed types of dementia, one is most common than the others, as the graph from the University of Queensland reports (figure 1).
to understand if a patient has one of dementia diseases several different tests exist and they can be splitted into three macro categories : Neurological exams, Mental status tests and brain imagin exams. In mental status tests patients' thinking skills are tested during the resolution of simple problems.Two common types of mental status tests are the MMSE

and Mini-Cog test. This project used MMSE test consisting into a 30-point questionnaire. This method is extensively used in clinical research to measure cognitive impairment through simple targeted questions and small graphic tasks to test the domain of different brain functions, such as orientation (autopsychic and outward), memory, attention and calculation, the ability to recall certain acquisitions and language.

2.2 Alzheimer and OPTIMA DATASET

As showed into the figure 2.1, Alzheimer's disease is the most common cause of dementia. It is a disease that causes a progressive loss of memory and a lowering of cognitive abilities limiting the normal activities of daily life.

Alzheimer has no cure but just a set of treatments to slow down the disease are known. Scientists haven't found Alzheimer's causes but only a partial set of hypothesis.

Some studies claim that the most common risk factors for dementia amenable to prevention are those associated with cerebrovascular disease[12].

Another study says that the most common cause of dementia is β-amyloid:it states neural death appears to be associated with intracellular calcium homeostasis and subsequent production of free radicals by calcium-sensitive enzymes[13].

This project focused on trying to extend research studies of the causes of Alzheimer's and in general of all types of dementia. To developed it, the data has been provided by the Oxford Project to Investigate Memory and Ageing(OPTIMA) [14].

OPTIMA collect a huge amount of data on both cognitively impaired and cognitively normal elderly people during life.

Inside informations, can be splitted into the following subcategory:

- Medical history and physical examination.
- Neuropsychological assessments.
- Brain scans.
- Urine and csf samples.
- Histopathological information following brain donation.
- DNA

2.3 Key Performance Indicator(KPI)

A Key Performance Indicator (KPI) is a measurable value used to gauge the performance of a process and identify the best practices. It measures the trend of a business process and helps making decisions to improve the outcomes[15]. Each company can use a set of metrics to track the status of a specific process. With KPIs, the company focuses the attention to one or more metric to become a point of reference to identify progresses.

The administration is in charge of defining business goals for its company, then, a so-called business analyst will analyze business goals to define KPIs. However, not all processes can be analyzed with KPIs and in general, the opportunity to analyze a process using them is measured with a robustness scale, which considers[16]:

- Ease of understanding - KPIs must explain what it is measuring at the minimum possible grain level
- Cost of data - Collecting data to enable analysis on a KPI costs. These costs can include collecting data, structuring data and experts on interpret results
- Meaningfulness - KPIs must reflect only critical goals i.e. the ones which give the company an improvement on quality of processes
- Frequency of data change - the more informative the indicator is with time, the more robust will be

During the development of this thesis some performance indicators have been defined to optimize the analysis on the dataset and identify the best practices.

2.4 Data Science

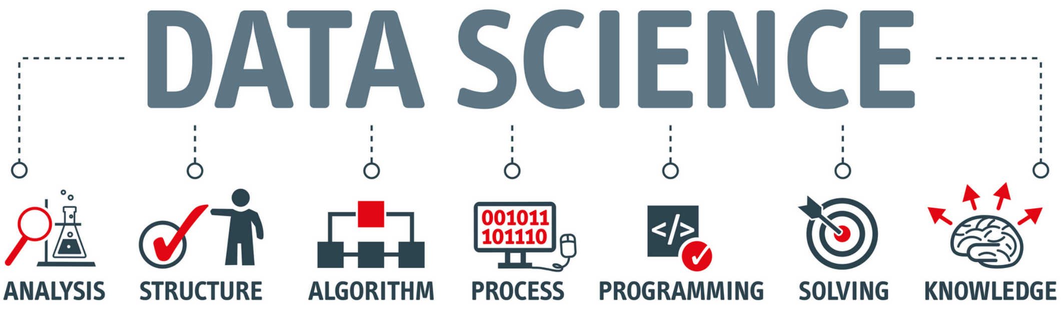


Figure 2: Data Science Steps Overview [2]

Data science is one of the scientific disciplines having an enormous growth during the 21st century. This field of science consists of using computer facilities, mathematics and statistics models to extract information and knowledge from unstructured and structured data. In its first application, data science has been used solely to building statistics but it has rapidly evolved becoming one of the pillars of today's society. We find this discipline in areas as artifical intelligence, machine learning and Internet of things. With the advent of Big Data has been applied in more traditional fields like medicine, engineering and social science.

Many different dates can be used to track Data Science's slow growth and its impact on the data management industry. In 1996, the term data mining became popular thank to an article called From Data Mining to Knowledge Discovery in Databases[17] where the overall process of discovering useful information from data has been explained.

As already mentioned, this multi-disciplinary field broke out during the early 2000s with the advent of the Internet era. In the age of Internet increased suddenly the amount of data and this phenomenon led to the creation of the term Big Data. The term Big data indicates the usage of large amount of information to perform analysis and extraction of new knowledge.

However, data are often stored without any planning for future analysis and leaving their potential knowledge hidden. Then if we want to extract this knowledge, we need to make a great effort to extract their real value. In the next sections, we will see that the most important operation during the data science project is the cleaning and standardization phases using the data wrangling procedure.

2.5 Data Science Project Development

Data Science Project has role to identify, analyze and extract the knowledge of a dataset. During this type of project, the main key is to have excellent communication among the client the company that commissiones the project and the data scientist team. To accomplish an excellent work the developer team need to understand perfectly the context, intermediate objectives and achieve the final goals.

Normally a Data Science project follows a common methodology called Cross-industry standard process for data mining(CISP-DM). CISP-DM is a technique conceived in 1996 defining a guideline to direct data mining efforts. It is composed of the steps shown in the figure below:

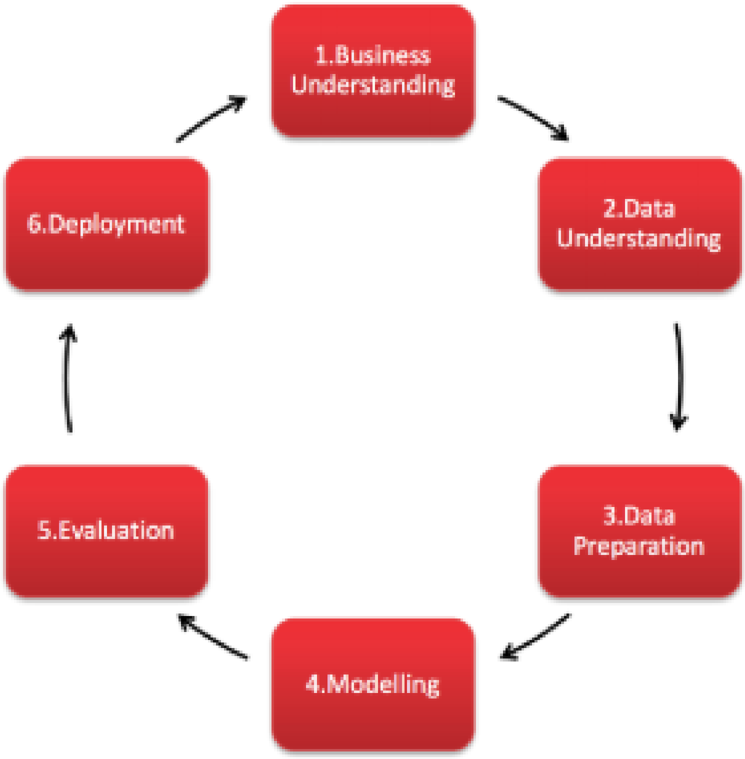


Figure 3: CRISP_DM phases and relationship against its [3]

A chief features of this technique is that it is no a linear process but a cycle process, it is as an idealised sequence of events. Tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain steps again.

2.5.1 Business Understanding

The primary objective of this phase of the project is to translate business goals into data mining goals. Business goals is part of the planning process and normally describe what a company want to accomplish using business terminology, whereas data mining goals describe the same goals but in technical terms.
To realize this translation, we need to follow this steps :

1. Define final output

   - Set Objectives - Definition and description of primary objective in business terms.
   - Project Plan Definition - Definition and description of the ways to achieve final goals and the phases this procedure is divided into. Project Plan fixes the tools and techniques to be used.
   - Business Success Criteria - Definition of criteria to understand if the project plan reaches fulfil the business goals.

2. Evaluate the initial situation

   - Inventory of Resources - List the available resources for the project.
   - Requirements, Assumptions and Constraints - List of all the project requirements .
   - Risks and Contingencies - Identification of possible risks and delays associated to them.
   - Terminology - Glossary of terminology relevant to the project.
   - Costs and Benefits - Analysis based on the relationship between costs and benefits. The idea is to understand if we have a set of final benefits upper than costs associated to develop the project.

3. Define data mining goals

   - Business success criteria - Description of criteria to consider outputs consistent with the achievement of the business objectives.

- Data mining success criteria - Description of criteria in order to consider outcomes successful for the projects in technical terms.

4. Produce project plan

   - Project plan - List of the stages to be executed in the project, together with their duration, required resources, inputs, outputs, and dependencies.
   - Initial assessment of tools and techniques - At the end of the first phase you should undertake an initial assessment of tools and techniques. Here, for example, you select a data mining tool that supports various methods for different stages of the process. It is important to assess tools and techniques early in the process since the selection of tools and techniques may influence the entire project.

2.5.2 Data Understanding

Data understanding phase includes three steps.
Data Collection listed data sources acquired. At the end of this phase it is necessary to develop a report for data description. At this scope, a preliminary descriptive analysis is needed, to extract information. In this project preliminary descriptive analysis is based on following the doctor's guidelines.
During Data Exploration a deep analysis of acquired data is fulfilled. To make this, we used reporting techniques and data visualization tools that are provided by the programming language used for the project.
Into Report for Data Exploration, consists of listing the results of the exploration of your information. This includes the initial hypothesis and states how they impact the development of the entire project.

2.5.3 Data Preparation

The third stage of the project is one of the most important because we have to decide which information we are going to use in the next phases. In all data science projects the first three phases (business understanding, data understanding and data preparation) account for more than 80% of the entire project efforts. 76% of data scientists say that

data preparation is the worst part of their job[18].
During data prepation, the following steps will be executed:

- Data Gathering - It defines the ways how data are provided from the company's data lake to perform the project. there are different ways to make this as using API, non-relational dataset or using a format like CSV. Into this project, data are provided CSV files. After getting the data, the goal is to identify the right columns in its. This happens using a data catolog or a data guideline where all tables of company's data lake are listed and for each table a features description is reported to help data scientist to identify right columns. Furthermore, data guideline shows which are data type used and their meaning.

- Data Cleaning - It is the most important step in data preparation phase and it is crucial for removing faulty data and filling in gaps. Often data provided, are unstructured then have to apply on it, techniques to manage missing values, errors and to build conforming data to a standardized pattern.
  Data cleaning has to strong impact on the future phases of the project because it defines the quality of data.

- Data Filtering - Data filtering refers to removing unnecessary information. the general idea is to remove the "noise" into the data defining specific constraints to apply. The final goal is to obtain a dataset where can perform a better analysis.

- Data Trasformation - It is a process where the main objective is to apply an updating the format where there are a converting on data types. Then, data transformation wants to achieve a well-defined result to obtain data more understood.

2.5.4 Modelling

The modelling phase is based on selecting the specific modelling technique to improve quality of data among the different available in the literature. Before we define a model we need to structure a scheduling procedure to test model's quality and validity.
Into machine learning world there are many different tools and procedures. The choise is associated to the goal that we wanto to obtain : Regression, Classification, Clustering, Dimensionality Reduction, Ensemble Methods, Neural Nets and Deep Learning, Transfer

Learning and Reinforcement Learning. After have defining the objective and have choosing the tool to use, the following iterative procedures will be followed:

- Algorithm implementation.
- Training - Data is fed into the algorithm to eventually learn patterns in data.
- Validation - The trained algorithm is checked against a validation data set to tune parameters that best fit our data.
- Testing - The algorithm is tested against new data to evaluate the abstraction power of the model.

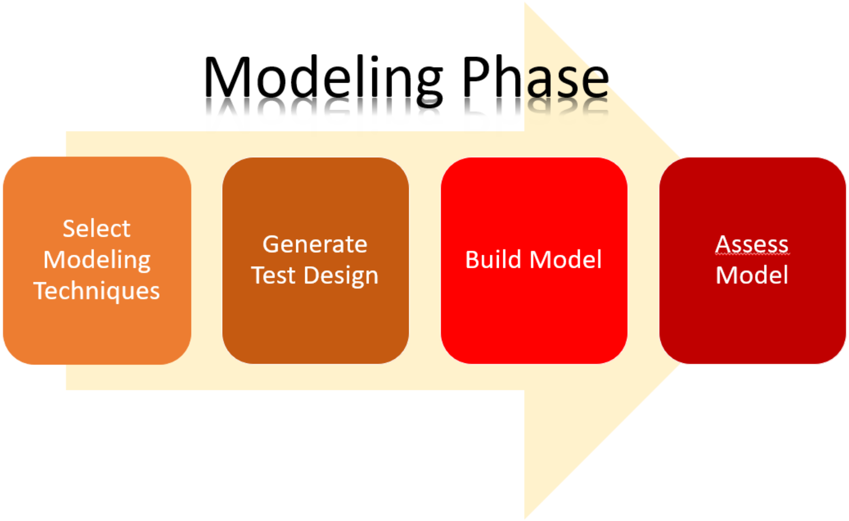


Figure 4: Modelling Steps Overview

2.5.5 Evaluation

Evaluation phase checks the obtained results using factors as accurateness, precision and computational effort. The idea is to understand if the model is able to meet the business goals defined during the business understanding phase.
We will do an evaluation of the data mining result considering the final business success standard and create a report on it. Then we will perform the process of reviewing where a more comprehensive review for the data mining process, is developed checking if every if task has been overlooked. The idea is to identify if some keys objective haven't been sufficiently considered.

2.5.6 Deployment

During this phase is defined as the structures throught which the data will be presented to the final client is built. The tools chosen to present the results are dashboard, API and web app.

3 Materials and Methods

In this chapter we explain every material and method used to develop this project. The chapter is splitted into two parts:

- The first one explains which data are used to create final dataset called Patient Final Data.
- The second one explains which programming languages them and packages are used to achieve the data science project proposing a general overview of them.

3.1 Raw Data

As mentioned in section 2.2, the data source used to develop this project is based on the OPTIMA data source that provides information on different fields: medical stuff, parent's history, general information and personal information. All the documents provided by OPTIMA are totally anonymous so that there is no way to have personal data like name, surname or date of birth.

OPTIMA data source have stored patient information since 2009. Every data have been provided in CSV format and it have the following basic features:

- ID -It identifies a unique patient.
- Study ID - Numerical value that identifies which are study:OPTIMA or Challenge.
- Gender - Gender of the patient.
- DOD - It identifies the date when data was collected.
- EPISODE DATE - It identifies the Date when episode happens.

With OPTIMA data source has been provided an excel file called Variables Guide.

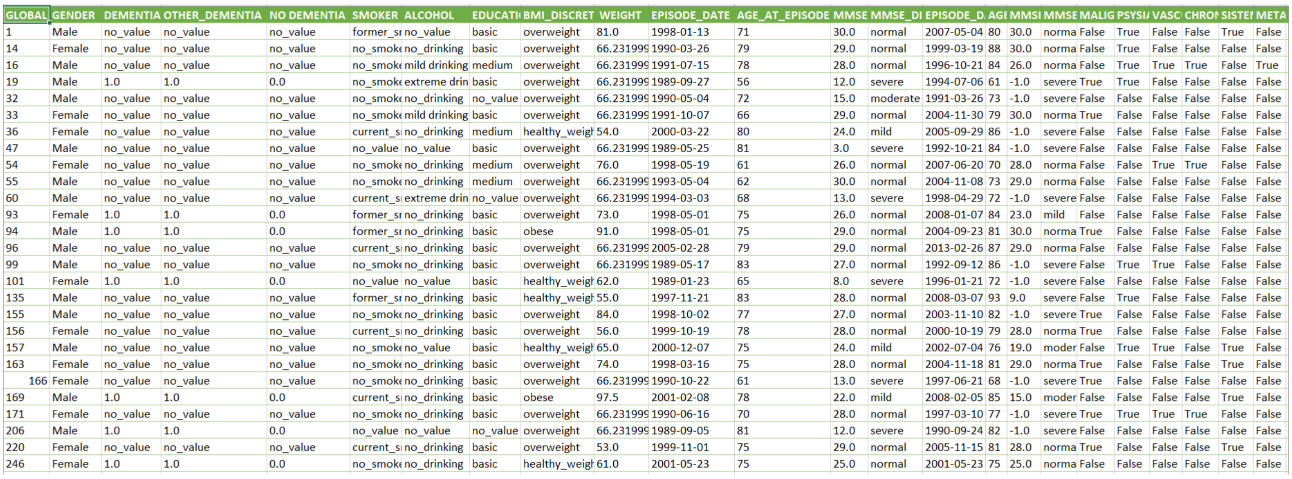
Variables Guide is a descriptive file providing us a complete overview about tables presenting the data source and a deep description of each columns. This description contains data types, which are value used and what are meaning. About Variables Guidelines there is an overview into section 4.3.1.

3.2 Patients Final Data

Patients Final Data will be the final output obtained after the first three phases of the CRISP-DM process.
To build Patient Final Data rules have been followed provided by a team of doctors that define in deep the features to use and how to combine then. In general we can say that Patients Final data is the output of the Data preparation phase and can be splitted into the following sub-category:

- General information - Global patient id, gender, episode date, age at episode.
- EHR information - body max index, weight, comorbdity types, APOE .
- Lifestyle information - education level, smoking level, alcohol level.
- Mini Mental Score information - mmmse for episode, mmse discretized, mmse final.
- Dementia type - alzheimer, other dementia, no dementia.

A deep description of the rules and filter applying are shown into the following chapter.



GLOBAL	GENDER	DEMENTIA	OTHER_DEMENTIA	NO DEMENTIA	SMOKER	ALCOHOL	EDUCATI	BMI_DISCRET	WEIGHT	EPISODE_DATE	AGE_AT_EPISODE	MMSE	MMSE_DI	EPISODE_D	AGI	MMSI	MMSE	MALIG	PSYSI	VASCI	CHRON	SISTEI	META
1	Male	no_value	no_value	no_value	former_s	no_value	basic	overweight	81.0	1998-01-13	71	30.0	normal	2007-05-04	80	30.0	norma	False	True	False	False	True	False
14	Female	no_value	no_value	no_value	no_smoke	no_drinking	basic	overweight	66.231999	1990-03-26	79	29.0	normal	1999-03-19	88	30.0	norma	False	True	False	False	False	False
16	Male	no_value	no_value	no_value	no_smoke	mild drinking	medium	overweight	66.231999	1991-07-15	78	28.0	normal	1996-10-21	84	26.0	norma	False	True	True	True	False	True
19	Male	1.0	1.0	0.0	no_smoke	extreme drin	basic	overweight	66.231999	1989-09-27	56	12.0	severe	1994-07-06	61	-1.0	severe	True	True	False	False	False	False
32	Male	no_value	no_value	no_value	no_smoke	no_drinking	no_value	overweight	66.231999	1990-05-04	72	15.0	moderate	1991-03-26	73	-1.0	severe	False	False	False	False	False	False
33	Female	no_value	no_value	no_value	no_smoke	mild drinking	basic	overweight	66.231999	1991-10-07	66	29.0	normal	2004-11-30	79	30.0	norma	True	False	False	False	False	False
36	Female	no_value	no_value	no_value	current_s	no_drinking	medium	healthy_weigh	54.0	2000-03-22	80	24.0	mild	2005-09-29	86	-1.0	severe	False	False	False	False	False	False
47	Male	no_value	no_value	no_value	no_value	no_value	basic	overweight	66.231999	1989-05-25	81	3.0	severe	1992-10-21	84	-1.0	severe	False	False	False	False	False	False
54	Female	no_value	no_value	no_value	no_smoke	no_drinking	medium	overweight	76.0	1998-05-19	61	26.0	normal	2007-06-20	70	28.0	norma	False	False	True	True	False	False
55	Male	no_value	no_value	no_value	no_smoke	no_drinking	medium	overweight	66.231999	1993-05-04	62	30.0	normal	2004-11-08	73	29.0	norma	False	False	False	False	False	False
60	Male	no_value	no_value	no_value	current_s	extreme drin	no_value	overweight	66.231999	1994-03-03	68	13.0	severe	1998-04-29	72	-1.0	severe	False	False	False	False	False	False
93	Female	1.0	1.0	0.0	former_s	no_drinking	basic	overweight	73.0	1998-05-01	75	26.0	normal	2008-01-07	84	23.0	mild	False	False	False	False	False	False
94	Male	1.0	1.0	0.0	former_s	no_drinking	basic	obese	91.0	1998-05-01	75	29.0	normal	2004-09-23	81	30.0	norma	True	False	False	False	False	False
96	Male	no_value	no_value	no_value	current_s	no_drinking	basic	overweight	66.231999	2005-02-28	79	29.0	normal	2013-02-26	87	29.0	norma	False	False	False	False	False	False
99	Male	no_value	no_value	no_value	no_smoke	no_drinking	basic	overweight	66.231999	1989-05-17	83	27.0	normal	1992-09-12	86	-1.0	severe	False	True	True	False	False	False
101	Female	1.0	1.0	0.0	no_value	no_value	basic	healthy_weigh	62.0	1989-01-23	65	8.0	severe	1996-01-21	72	-1.0	severe	False	False	False	False	False	False
135	Male	no_value	no_value	no_value	former_s	no_drinking	basic	healthy_weigh	55.0	1997-11-21	83	28.0	normal	2008-03-07	93	9.0	severe	False	True	False	False	False	False
155	Male	no_value	no_value	no_value	no_smoke	no_drinking	basic	overweight	84.0	1998-10-02	77	27.0	normal	2003-11-10	82	-1.0	severe	True	False	False	False	False	False
156	Female	no_value	no_value	no_value	current_s	no_drinking	basic	overweight	56.0	1999-10-19	78	28.0	normal	2000-10-19	79	28.0	norma	True	False	False	False	False	False
157	Male	no_value	no_value	no_value	no_smoke	no_value	basic	healthy_weigh	65.0	2000-12-07	75	24.0	mild	2002-07-04	76	19.0	moder	False	True	True	False	True	False
163	Female	no_value	no_value	no_value	no_smoke	no_drinking	basic	overweight	74.0	1998-03-16	75	28.0	normal	2004-11-18	81	29.0	norma	True	False	False	False	True	False
166	Female	no_value	no_value	no_value	no_smoke	no_drinking	basic	overweight	66.231999	1990-10-22	61	13.0	severe	1997-06-21	68	-1.0	severe	True	False	False	False	False	False
169	Male	1.0	1.0	0.0	current_s	no_drinking	basic	obese	97.5	2001-02-08	78	22.0	mild	2008-02-05	85	15.0	moder	False	False	False	False	True	False
171	Female	no_value	no_value	no_value	no_smoke	no_drinking	basic	overweight	66.231999	1990-06-16	70	28.0	normal	1997-03-10	77	-1.0	severe	True	True	True	True	False	False
206	Male	1.0	1.0	0.0	no_value	no_value	no_value	overweight	66.231999	1989-09-05	81	12.0	severe	1990-09-24	82	-1.0	severe	False	False	False	False	False	False
220	Female	no_value	no_value	no_value	current_s	no_drinking	basic	overweight	53.0	1999-11-01	75	29.0	normal	2005-11-15	81	28.0	norma	True	False	False	False	True	False
246	Female	1.0	1.0	0.0	no_smoke	no_drinking	basic	healthy_weigh	61.0	2001-05-23	75	25.0	normal	2001-05-23	75	25.0	norma	False	False	False	False	False	False

Figure 5: Final Patient Data Overview

3.3 Python and R(Programming Languages)

the main programming language used to develop this project is Python. It is a multi-paradigm language that has among its main objectives: dynamism, simplicity and flexibility [19]. It was conceived in 1980 by Guido van Rossum at Centrum Wiskunde Informatica (CWI) and it is an interpreted, high-level, general-purpose programming language. It is the most powerful language because it is able to integrate a huge variety of programming paradigms and presents a comprehensive standard library.

Python is an open-source language and there is a global community called CPython that provides reference implementation, a complete explanation about the libraries and how to use it.

Python is a language that provides a huge set of tools in the area of statistics. Being a open-source system, many different packages containing a variety of tools, methods and techniques are freely available.

Also there is a huge number of packages available for the machine learning sector and for the data mining world. in this project,many different packages have been used to help with data visualization and data manipulation stage. These packages have been explained in deep in the following paragraphs. Furthermore, during this project R programming language has been used to develop survival analysis. R programming language is also an open-source language that provide good tools to made analysis into a dataset.

3.4 Visualization Packages in Python

Matplotlib.Pyplot

Matplotlib.pyplot is a collection of command style functions having the role to create graphs, figures like Matlab's styles. Each pyplot function makes some changes to a figure such as creating a figure, creating a plotting area in a figure, figuring some lines in a plotting area, decorating the plot with labels and so on. It is much popular for the data visualization stage using elementary a graph's style.

Seaborn

Seaborn is a library to achieve statistic graphs based on matplotlib with strong integration with pandas data structure. Pandas is the main library that can be used to perform analysis with dataframes, datalakes and in general to use data mining and machine learning tools. Seaborn aims to make visualization a primary part of exploration and to understanding data defining semantic mapping and statistical aggregation necessary to build informative graphs.

3.5 Data Manipulation Packages in Python

Pandas

Pandas is a software library to perform operations for data manipulation and analysis. Pandas was created by Wes McKinney in 2008 when he needed to perform quantitative analysis on financial data. It contains data structures for tables and series manipulation and provide the possibility to import a huge variety of file formats such as csv, excel, JSN, ZIP, images and so on. Pandas library allows us to use SQL tools like groupby, merge, concatenation and also tools to make data cleaning and data filtering. It is the main instrument for data manipulation phase in Python language.

NumPy

NumPy is a library to manage multidimensional array objects with a set of routines collection. Furthermore, it allow us to manage a large collection of high-level mathematical functions. The name Numpy is a combination of the two words : num = Numerical Py = Python and was created by Jim Hugunin.

Math

Library that provides a huge set of mathematical functions coded in the C standard. This type of function can be used with all types of existing numbers(complex, real, irrational, etc).

Datetime

Python DateTime, TimeDelta, Strftime(Format) with Examples. In Python, date, time and datetime classes provides a number of function to deal with dates, times and time intervals. Date and datetime are objects in Python, so when you manipulate them, you are actually manipulating objects and not string or timestamps.

4 Proposed approach to extract pattern of dementia evolution

The following chapter explains in detail the whole procedure followed to achieve the final objectives. The central element is to follow the CRISP-DM methodology for the realization of the project. The procedure used to translate the business objectives into data mining objectives, what are the KPI set and partial data exploration of raw data, will be explained. Furthermore, a complete explanation about features putting in my final dataset has been presented. Many of these features have been setted using the doctors 's rules that are explained into section data extraction.

In general we can say that the entire data science project procedure have been followed. Into the following section chapter have been splitted : Bussines understanding, Data understanding, Data exploration, Data preparation and Data extraction.

4.1 Description of the goals

During the development of a Data Science project, a common misperception is that the main part of the work is focused to apply statistics algorithms using tools like deep learning or machine learning but this is often only the last step. In practice a data scientist spends more than 60% of its time to discuss and understand the data after that data scientist starts to apply the prediction methods mentioned before.

This procedure, called data wrangling, represents much of the analyst's professional process and it takes up most of its working day. Activities to apply are: to understand the informations are provided by the data, to choose the data to use, to understand how to combine multiple different data sources, to decide how to split the result in terms of size and shape that can drive the next steps of analysis.

The final goal of this thesis is to define methods and techniques that can been used to establish dementia patient information and how they evolve after the first episode to the last one. The idea is that a clinic will use this model to aggregate information and predict the likelihood of a rapid or slow disease progression. The rate of progression will be defined according to a specific metric derived from a measure of cognitive ability as mentioned in section 2.1, mini mental state examination(MMSE).

4.2 Business Understanding

In this project three different KPIs have been defined to understand and improve the knowledge about dementia and its causes: Multiple Filter Analysis, Identification of possible strong correlations between MMSE and patient information, How to evolve the rate of progression after the first diagnosis.

4.2.1 Multiple Filter Analysis

During this phase filter analysis have been performed on Patient Final Data based on different factors such as age, sex, comorbidities, lifestyle, family history (both maternal and paternal), age of onset, BMI and so on.

Process Outcome KP1 – Multiple Filter Analysis

This key indicator is composed by the following metrics :

- Main comorbidities associated to moderate level of dementia
- Main comorbidities associated to severe level of dementia.
- Number of patients that are strong smoker with at least one type of dementia.
- Number of patients that are strong drinker with at least one type of dementia.
- Number of patients getting worse.
- Number of patients remaining stable.
- Number of patients getting better.

4.2.2 Identification of possible strong correlations between MMSE and patient information

As stated before, this KPI is used to analyze any possible correlation between MMSE and patient information. In this section it has been tried to understand if some features have values above a fixed threshold using data analysis tools like heatmap and bar charts. Using the reference document, we can say that correlation coefficients are used in statistics to measure relationship between two variables. The formulas return a value between

-1 and 1, where 1 indicates a strong positive relationship, -1 indicates a strong negative relationship and zero indicates no relationship on its. [20].

Process Outcome KP2 - Identification of strong correlations between MMSE and patient information

The main metrics to achieve reasonable results are the following:

- Heatmap analysis.
- Identfication of features with high correlation ratio.
- Level of multi correlation in the Patients Final Data episode by episode.
- Bivariate analysis episode by episode.

4.2.3 How evolve the rate of progression after First diagnosis

Survival Analysis is used to estimate the lifespan of a particular population under study. The goal is to estimate the time for an individual or a group of individuals to experience an event of interest [21]. Into this project the two events used to build survival analysis are the first episode analyzed and the episode when the patient becomes a severe one. Then, it have been computed "time to event" how week among it.

Process Outcome KP3 - Rate of progression After First diagnosis

Into this section, main key indicator are:

- Time for Mini Mental State Examination to become severe.
- Identification of features that have p_value less than a specific threshold.
- Hazard Ratio Analysis for a set of features.

4.2.4 Technical Goals

Once business goals have been defined, it's time to translate the upon goals into a more technical objectives using the data mining language. The first stuff needed to do is to apply programming tools to make data ready. In this phase we have applied the following Preparation Goals:

1. **General Preparation Goals**

    - Preparation Goals 0.1 - Cleaning and standardization of available data.
    - Preparation Goals 0.2 - Extraction of patient's information from the OPTIMA documents.

2. **KPI 1 Main Goals** In order to obtain insights to fulfil the KPI 1, the following tasks are required:

    - Preparation Goals 1.1 - Classification of patients into a fixed set of subgroups in order to enable more precise analysis.
    - Preparation Goals 1.2 - Definition of filter analysis on comorbidities type
    - Preparation Goals 1.3 - Definition of filter analysis based on APOE genotype to understand if there is a more important one.
    - Preparation Goals 1.4 - a complete analysis of patient's lifestyle trying to identify if exists a correlation between behaviour and mini mental state examination.
    - Preparation Goals 1.5 - Definition of filter analysis on three type of dementia.
    - Preparation Goals 1.6 - Setting a new feature called Mixed Status that is a combination of more than one features present into Patient Final Data.

3. **KPI 2 Main Goals** In order to obtain insights to fulfil the KPI 2 the following tasks are required:

    - Preparation Goals 2.2 - Setting a constraint that limits patients in the dataset at only those than having at least 5 different episodes.
    - Preparation Goals 2.2 - As explained in paragraph 5.3.3, Setting a threshold that defines the minimum number of patients for an episode.

4. **KPI 3 Main Goals** In order to obtain the outcome KP 3 the following steps are required:

    - Preparation Goals 3.1 - Identification of features that have p_value less than 0.05.

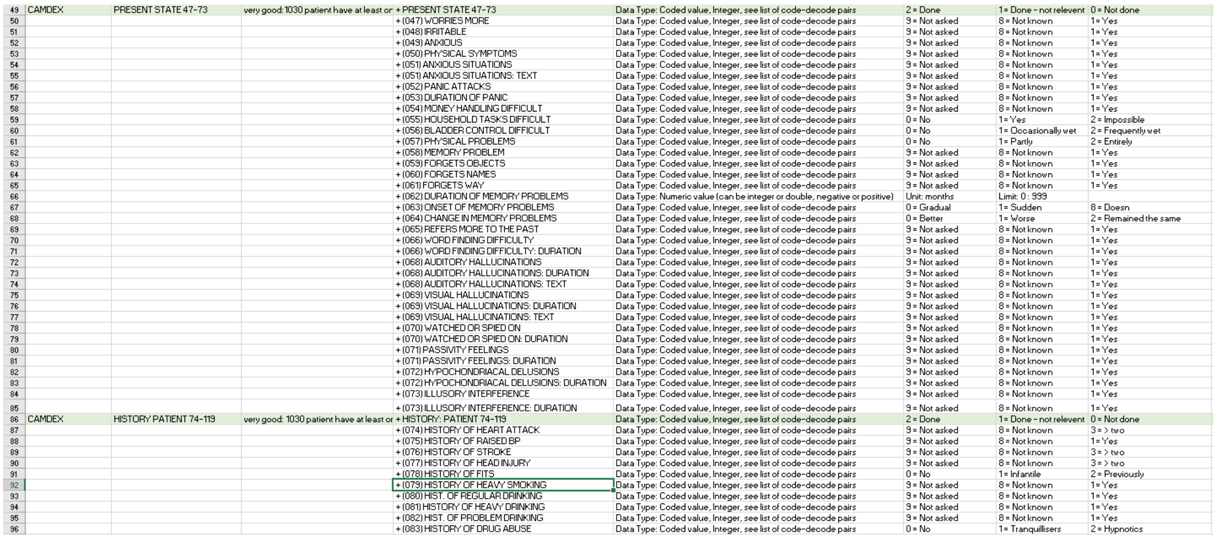
- Preparation Goals 3.2 - Using feautures found into PG3.1, applying on it Cox Proportional Hazard Models.

4.3 Data Understanding

The intention of this section is to fulfil a first preliminary exploration on data. To develop this project the main dataset,called Final Patient Data has been created, that combine a subset of information taken of the OPTIMA.

4.3.1 Description of OPTIMA Data Source

Optima Data Source is a huge Data Source that contains many different Data Files used to take informations. Into this Data Source the principal file is Variable Guide-UPM.xls, a file containing guidelines for the variables. For each variable there is a brief description,data types and information about strange or incorrect values. For example, some tables presented 9999 values that means null value.



49	CAMDEX	PRESENT STATE 47-73	very good:1030 patient have at least or	+ PRESENT STATE 47-73	Data Type: Coded value, Integer, see list of code-decode pairs	2 = Done	1= Done - not relevent	0 = Not done
50				+ (047) WORRIES MORE	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
51				+ (048) IRRITABLE	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
52				+ (049) ANXIOUS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
53				+ (050) PHYSICAL SYMPTOMS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
54				+ (051) ANXIOUS SITUATIONS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
55				+ (051) ANXIOUS SITUATIONS: TEXT	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
56				+ (052) PANIC ATTACKS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
57				+ (053) DURATION OF PANIC	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
58				+ (054) MONEY HANDLING DIFFICULT	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
59				+ (055) HOUSEHOLD TASKS DIFFICULT	Data Type: Coded value, Integer, see list of code-decode pairs	0 = No	1= Yes	2 = Impossible
60				+ (056) BLADDER CONTROL DIFFICULT	Data Type: Coded value, Integer, see list of code-decode pairs	0 = No	1= Occasionally wet	2 = Frequently wet
61				+ (057) PHYSICAL PROBLEMS	Data Type: Coded value, Integer, see list of code-decode pairs	0 = No	1= Partly	2 = Entirely
62				+ (058) MEMORY PROBLEM	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
63				+ (059) FORGETS OBJECTS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
64				+ (060) FORGETS NAMES	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
65				+ (061) FORGETS WAY	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
66				+ (062) DURATION OF MEMORY PROBLEMS	Data Type: Numeric value (can be integer or double, negative or positive)	Unit: months	Limit: 0 : 999	
67				+ (063) ONSET OF MEMORY PROBLEMS	Data Type: Coded value, Integer, see list of code-decode pairs	0 = Gradual	1= Sudden	8 = Doesn
68				+ (064) CHANGE IN MEMORY PROBLEMS	Data Type: Coded value, Integer, see list of code-decode pairs	0 = Better	1= Worse	2 = Remained the same
69				+ (065) REFERS MORE TO THE PAST	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
70				+ (066) WORD FINDING DIFFICULTY	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
71				+ (066) WORD FINDING DIFFICULTY: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
72				+ (068) AUDITORY HALLUCINATIONS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
73				+ (068) AUDITORY HALLUCINATIONS: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
74				+ (068) AUDITORY HALLUCINATIONS: TEXT	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
75				+ (069) VISUAL HALLUCINATIONS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
76				+ (069) VISUAL HALLUCINATIONS: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
77				+ (069) VISUAL HALLUCINATIONS: TEXT	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
78				+ (070) WATCHED OR SPIED ON	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
79				+ (070) WATCHED OR SPIED ON: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
80				+ (071) PASSIVITY FEELINGS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
81				+ (071) PASSIVITY FEELINGS: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
82				+ (072) HYPOCHONDRIACAL DELUSIONS	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
83				+ (072) HYPOCHONDRIACAL DELUSIONS: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
84				+ (073) ILLUSORY INTERFERENCE	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
85				+ (073) ILLUSORY INTERFERENCE: DURATION	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
86	CAMDEX	HISTORY PATIENT 74-119	very good: 1030 patient have at least or	+ HISTORY: PATIENT 74-119	Data Type: Coded value, Integer, see list of code-decode pairs	2 = Done	1= Done - not relevent	0 = Not done
87				+ (074) HISTORY OF HEART ATTACK	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	3 = > two
88				+ (075) HISTORY OF RAISED BP	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
89				+ (076) HISTORY OF STROKE	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	3 = > two
90				+ (077) HISTORY OF HEAD INJURY	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	3 = > two
91				+ (078) HISTORY OF FITS	Data Type: Coded value, Integer, see list of code-decode pairs	0 = No	1= Infantile	2 = Previously
92				+ (079) HISTORY OF HEAVY SMOKING	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
93				+ (080) HIST. OF REGULAR DRINKING	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
94				+ (081) HISTORY OF HEAVY DRINKING	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
95				+ (082) HIST. OF PROBLEM DRINKING	Data Type: Coded value, Integer, see list of code-decode pairs	9 = Not asked	8 = Not known	1= Yes
96				+ (083) HISTORY OF DRUG ABUSE	Data Type: Coded value, Integer, see list of code-decode pairs	0 = No	1= Tranquillisers	2 = Hypnotics

Figure 6: Overview Variable_ Guide

4.3.2 Description of the Data Files

To build **Patients Final Data** an huge different files taken from the OPTIMA project have been used. Files used are the following :

- **CAMDEXSCORES.xlsx**
  To obtain MMSE(Mini Mental State Examination) and gender for each patient episodes.

- **optimaDX.xlsx**
  To classify if the patient has Alzheimer, other type of dementia or no dementia.

- **Data Request Jan 2019 (Optima, Lead, Challenge) final history.xlsx and nart duration smoking.xlsx**
  To classify patients into one of the following categories : current smoker, former smoker and no smoker.

- **Data Request Jan 2019 (Optima, Lead, Challenge) final history.xlsx**
  To define Alcohol Consumption and Education Level of each patient.

- **Clinical background.xlsx**
  To obtain weight and BMI for each patient.

- **Data Request Jan 2019 (Optima, Lead, Challenge) final history.xlsx**
  To define Malignant Comorbidity, Physiatric Comorbidity, Vascular Comorbidity, Sistemic Comorbidity, Sistemic Comorbidity.

- **DiseasesORcomorbidities - comorbidities.xlsx and optima patients CUIs.xlsx**
  To identify if the patient has Chronic Infectin Comorbidity, Metabolic Comorbidity, Respiratory Comorbidity, K Comorbidity, Inflammatory Comorbidity.

- **APOE.xlsx**
  To obtain APOE for each patient.

Using these files a main database with 30 features has been build. An overview of the characteristics features is given in the table below:

Patient Final Data Columns		
Variable Name	Variable Type	Column Description
GLOBAL PATIENT DB ID	continuos	unique key dataset
GENDER	categorical	male or female
DEMENTIA	categorical	1=yes, 0=no or no value
OTHER DEMENTIA	categorical	1=yes, 0=no or no value
NO DEMENTIA	categorical	1=yes, 0=no or no value
PETERSON MCI	categorical	1=yes, 0=no or no value
SMOKER	categorical	current smoker, ex smoker, no smoker or no value
ALCOHOL	categorical	extreme drinking, mild drinking, no drinking or no value
EDUCATION	categorical	medium, basic, high, no value
BMI NUMERIC	continuos	body max index number
BMI DISCRETIZED	categorical	healthy weight, obese, overweight or underweight
WEIGHT	continuos	numeric value in kg
EPISODE DATE	date	standard gg/mm/aaaa
AGE AT EPISODE	continuos	numeric value
MMSE FOR EPISODE	continuos	score in numeric value
MMSE DISCRETIZED INIT	categorical	mild, moderate, normal, severe
EPISODE DATE FINAL	date	standard gg/mm/aaaa
AGE AT FINAL EPISODE	continuos	numeric value
MMSE FINAL	continuos	score in numeric value
MMSE DISCRETIZED FINAL	categorical	mild, moderate, normal, severe
MALIGNANT COMORBIDITY	categorical	1=yes, 0=no or no value
PSYSIATRIC COMORBIDITY	categorical	1=yes, 0=no or no value
VASCULAR COMORBIDITY	categorical	1=yes, 0=no or no value
CHRONIC INFECTIN COMORBIDITY	categorical	1=yes, 0=no or no value
SISTEMIC COMORBIDITY	categorical	1=yes, 0=no or no value
METABOLIC COMORBIDITY	categorical	1=yes, 0=no or no value
RESPIRATORY COMORBIDITY	categorical	1=yes, 0=no or no value
K COMORBIDITY	categorical	1=yes, 0=no or no value
IMFLAMMATORY COMORBIDITY	categorical	1=yes, 0=no or no value
APOE	categorical	all types of APOE genotype

Table 1 : Overview of Final Patient Data columns

4.4 Exploratory Analysis

In the first three paragraphs some information about the raw data and then some distribution graphs about the main features into Patients Final Data have been showed.

4.4.1 Raw Data Exploration

A general overview of my raw Data is shown. The general idea has been to make some exploration analysis into three specific dataset: **CAMDEXSCORES.xlsx**, **APOE.xlsx** and **Data Request Jan 2019 (Optima, Lead, Challenge) final history.xlsx**.

Important Features CAMDEXSCORES file	
Variable Name	Variable Type
GLOBAL PATIENT DB ID	Continuos
GENDER	Categorical
EPISODE DATE	Date
CAMDEX SCORES: MINI MENTAL SCORE	Continuos
AGE AT EPISODE	Continuos
Number of patients	**total rows**
1035	9565

Table 2 : Overview of CAMDEXSCORES.xlsx

Important Features Data Request Jan 2019 final history file	
Variable Name	Variable Type
GLOBAL PATIENT DB ID	Continuos
GENDER	Categorical
EPISODE DATE	Date
AGE AT EPISODE	Continuos
HISTORY PATIENT 74-119: (079) HISTORY OF HEAVY SMOKING	Categorical
HISTORY OF PATIENT 74-119:(080): HIST. OF REGULAR DRINKING	Categorical
HISTORY OF PATIENT 74-119:(081): HISTORY OF HEAVY DRINKING	Categorical
HISTORY OF PATIENT 74-119:(081): HISTORY OF DRINKING PROBLEM	Categorical
HISTORY PATIENT 74-119:(115)HISTORY CANCER	Categorical
Number of patients	**total rows**
1035	9565

Table 3 : Overview of Data Request Jan 2019 (Optima, Lead, Challenge) final history.xlsx

Important Features APOE file	
Variable Name	Variable Type
GLOBAL PATIENT DB ID APOE EPISODE DATE	Continuos Categorical Date
Number of patients	**total rows**
1035	9565

Table 4 : Overview of APOE.xlsx

Following rules provided by doctor's team patients having null value for **CAMDEX SCORES: MINI MENTAL SCORE** feature into CAMDEXSCORES file and null value for **APOE** feauture into APOE have been removed. During the first analysis into CAMDEXSCORES.XLSX has been seen that we have **9565** tuple splitting for **1035** different patients. If we apply filters to remove null value we obtain the following results:

Rows and Number of Patient with Raw Data	
Number of Rows	Number of Patients
9565	1035
Rows and Number of Patient with Clean Data	
Number of Rows	Number of Patients
8172	1013

Table 5 : Comparison between Raw Data and Clean Data

Then 1393 rows and 22 patients.In order too obtain these results the join tools have been used merging the previous dataframes. the keys used to make inner join operation is the pair GLOBAL PATIENT DB ID, EPISODE DATE.

4.4.2 Patient Final Data Exploration

In this section, using Data Visualization tools, the distribution of the most important features in Patient Final Data has been shown to understand if dataset is homogeneous or not and it if is possible to see specific trend or something like that.

4.4.3 Distribution of Patients by AGE AT FIRST EPISODE

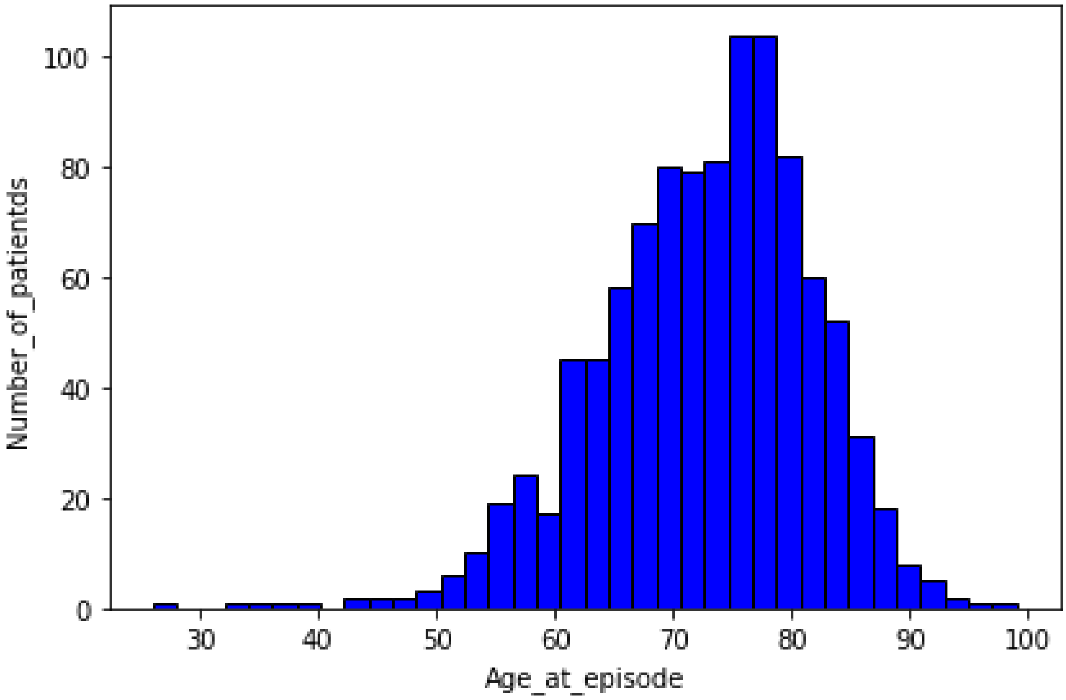


Figure 7: Distribution of Patients by AGE

Data Source: Patients File

In figures 7 the distribution of Patients into the OPTIMA Data Source based on Age has been shown. The histogram shows that more than 75% of the people are more than 65 years old. This results make sense because this type of disease is associated with older people. It is good to remember that this graph shows people during the first episode. The peak is in the range between 75 and 80 years old.

4.4.4 Distribution of Patients by MMSE SCORE

Data Source: Patients File

In figure 11 the distribution of MMSE SCORE has been shown for the first episode and for the last one.

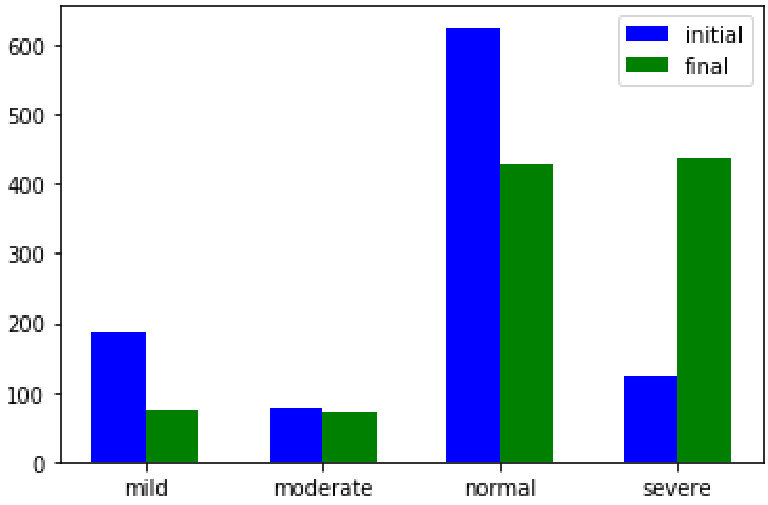


Figure 8: Distribution of Patient by MMSE SCORE

It is possible to comprehend that normal status and severe status follow an inverse trend. The number of normal status decreases of about 33% (200 units less) while severe status increase of about 300% (300 units more).

These result are inline with scientific research that says risk of dementia diseases grows up in proportion of seniority. MMSE is provided to us as a numerical value then have been applied a code to categorize it to build a categorical analysis like the previous one.

The applied rules, are shown in the following code:

```
list_mmse = MMSE_data['CAMDEX SCORES: MINI MENTAL SCORE'].tolist()
list_mmse_categorical =[]
for i in range(len(list_patient)) :
   if int(list_mmmse[i]) > 24 :
      list_mmse_categorical.append('normal')
   if int(list_mmse[i]) >=19 and int(list_mmse[i]) <= 24:
      list_mmse_categorical.append('mild')
   if int(list_mmse[i]) >=14 and int(list_mmse[i]) < 19:
      list_mmse_categorical.append('moderate')
   if int(list_mmmse[i]) < 14 :
      list_mmse_categorical.append('severe')
```

The Mini-Mental state examination is a cognitive test scoring on a scale of 0-30 where cognitive impairment is valutated in this way:

- Normal cognitive impairment: >24 (It means no relevant problem with memory, language and in general no cognitive problem).
- Mild cognitive impairment: 19-24 (It defines problems with memory, language, thinking and judgment at a greater level if it is compared with normal age-related changes. In general, it does not interfere notably with activities of daily life [22]).
- Moderate cognitive impairment: 14-18 (During this stage are visible the first dementia symptoms, such as difficulties with language and problem-solving).
- Severe cognitive impairment: <14 (A severe cognitive impairment is defined as a deterioration or loss in intellectual capacity that places a person in jeopardy of harming him or herself or others and, therefore, the person requires substantial supervision by another person).

4.4.5 Distribution of Patients by ALLELE

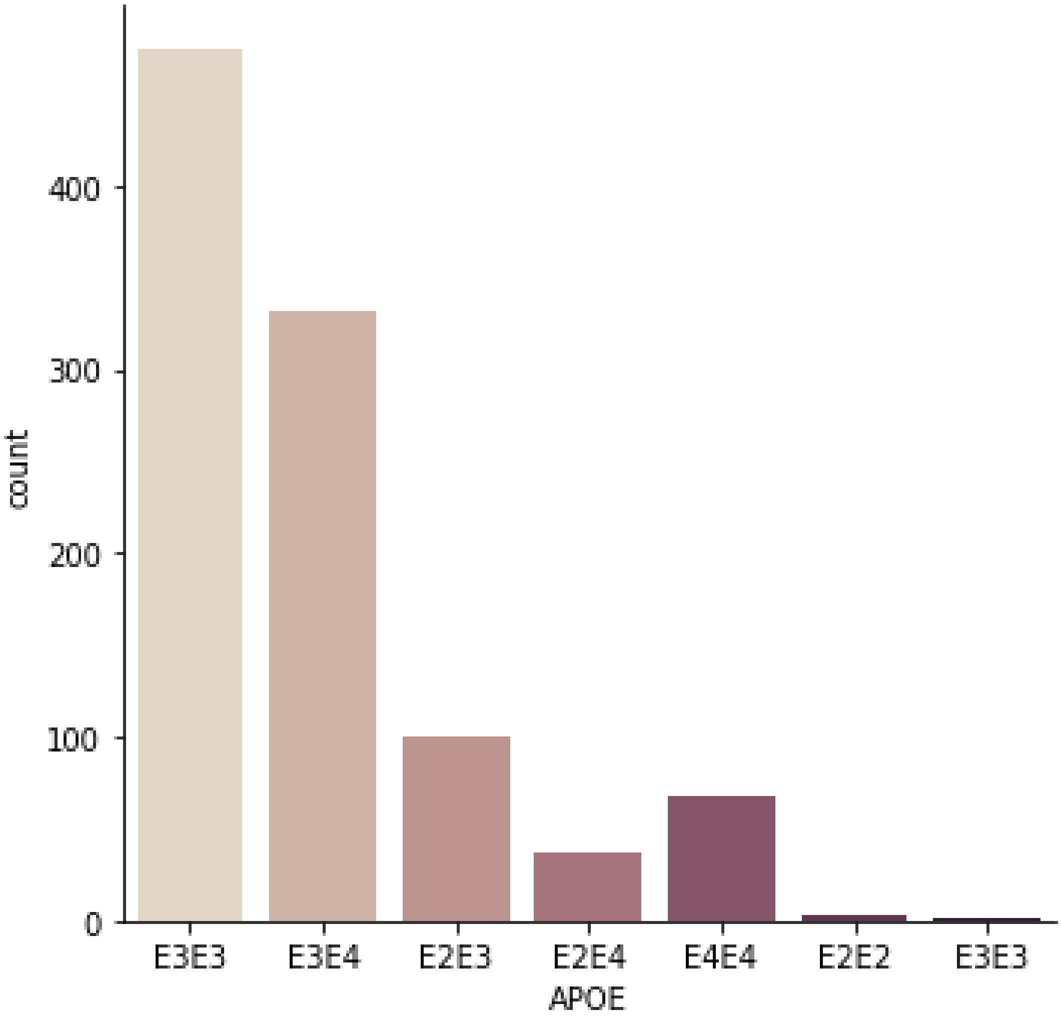


Figure 9: Distribution of Patients by APOE

Data Source: Patients File
In figure 8 the distribution of APOE has been shown.

What's APOE?
""The APOE gene provides instructions for making a protein called apolipoprotein E. This protein combines with fats (lipids) in the body to form molecules called lipoproteins[23]"".

Three different type of alleles of APOE gene exist: E2,E3,E4. Everyone has one pair of the gene and the possible combination are the following : E3E3, E3E4, E2E3, E2E4, E4E4, E2E2, E3E3. Many studies prove that APOE4 allele increase the risk for Alzheimer's and in general for dementia diseases. Having one copy of E4 (E3/E4) can increase your risk by 2 to 3 times while two copies (E4/E4) can increase the risk by 12 times.[9]

Into OPTIMA dataset patients are distributed in this way: APOE with at least one gene like E3 are more than 82% of samples.Furthermore the graph shows that APOE with at least one gene as E4 are more than 42%.

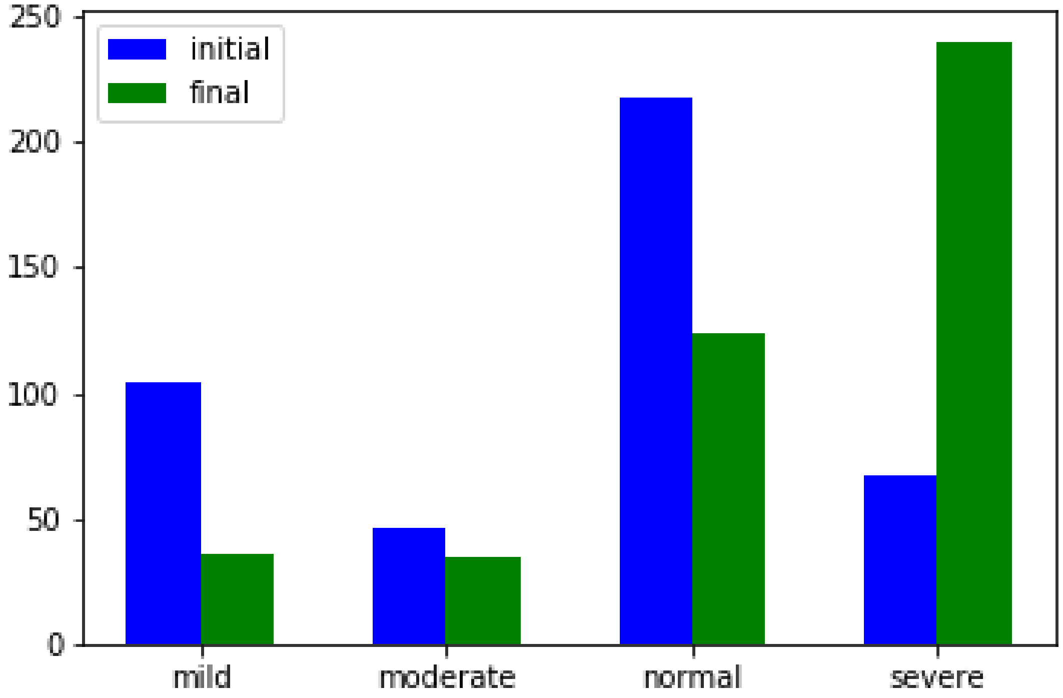


Figure 10: Distribution of Patient by MMSE only for patients that have at least one allele E4

The figure above show how evolve the Mini Mental Score from the first test to the last one considering only patients having at least one gene like E4 (427 patients).

Using this filter, we noticed that figure 9 clearly shows that during the first episode more than 52% of patients have normal MMSE status and only 16% have severe status. While in the last episode more than 56% have a severe status and less than 30% have normal status. This can be considered a good indicator meaning the progression rate of Dementia's disease has a reliable correlation with the E4 gene.

4.4.6 Distribution of Patients by GENDER

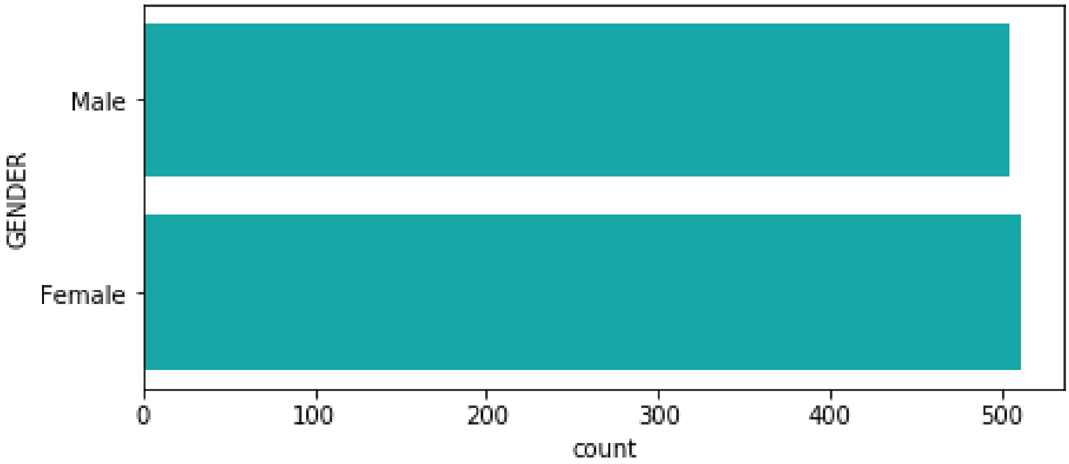


Figure 11: Distribution of Patient by GENDER

Data Source: Patients File

In figure above distribution of patients based on gender has been shown. The distribution based on gender is the following: number of women 510 and number of men 503. Then the distribution of patients based on gender is perfectly homogeneous.

4.4.7 Distribution of Patients by COMORBIDITIES

Data Source: Patients File

in this section the most common comorbidities for patients into the OPTIMA data source have been analyze. In this stage the goal is to make a preliminary analysis based only on the first episode. Into Patients Final Data there are 9 different types of comorbidity for this reason only a subset of these have been considered:

- **Malignant Comorbidity - Diseases list** : all types of cancer disease based on patient history.
- **Vascular Comorbidity - Diseases list** : diabetes, hypertension, heart disease, hypercholesterolemia, and peripheral vascular disease.
- **Psysiatric Comorbidity - Diseases list** : The most prevalent disease into this category are anxiety disorders, substance use disorders and other depressive disorders.

- **Chronic Infectin Comorbidity - Diseases list** : Burkitt lymphoma, cervical cancer disease, chronic Lyme arthritis disease, neruborreliosis disease, kaposi sarcoma disease, bacillary anglomatosis disease.

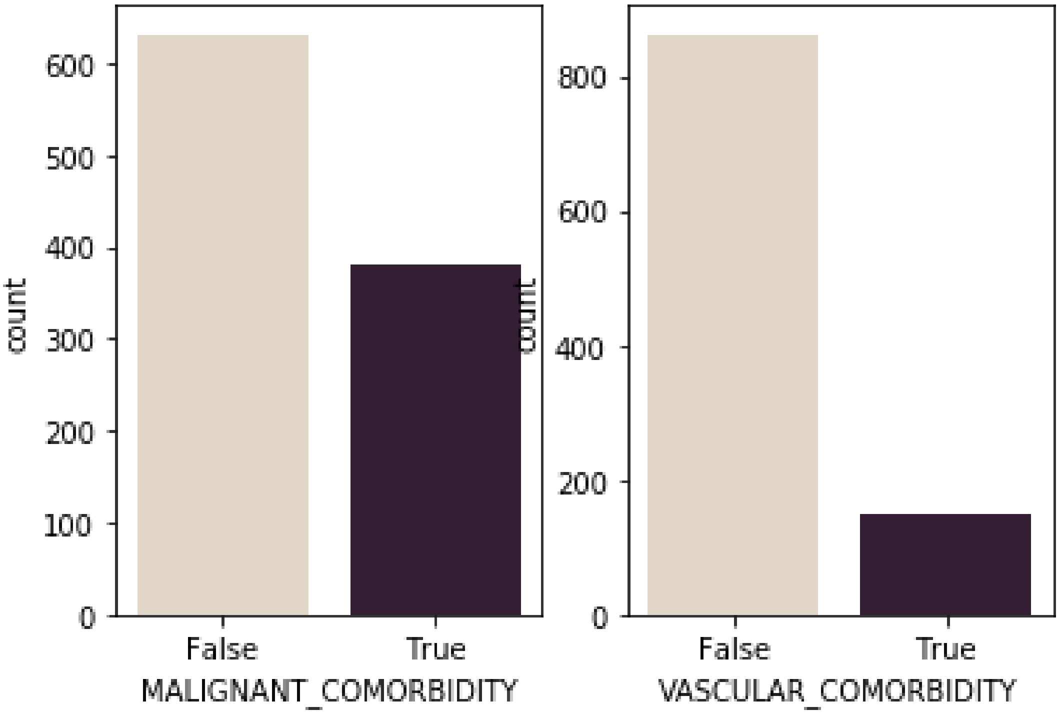


Figure 12: Distribution patient by Comorbidities: Malignant and vascular comorbidities

The first graph in figure 13 show that during the first episode, people having Malignant Comorbidity are about 40% of the samples. This trend is in line with some studies that claim cancer is more common in people that have more than 70 years where 46% of them being dead.[24]. The second graph show distribution patient based on vascular comorbiditity during the first episode where only 16.8% of samples has the disease.

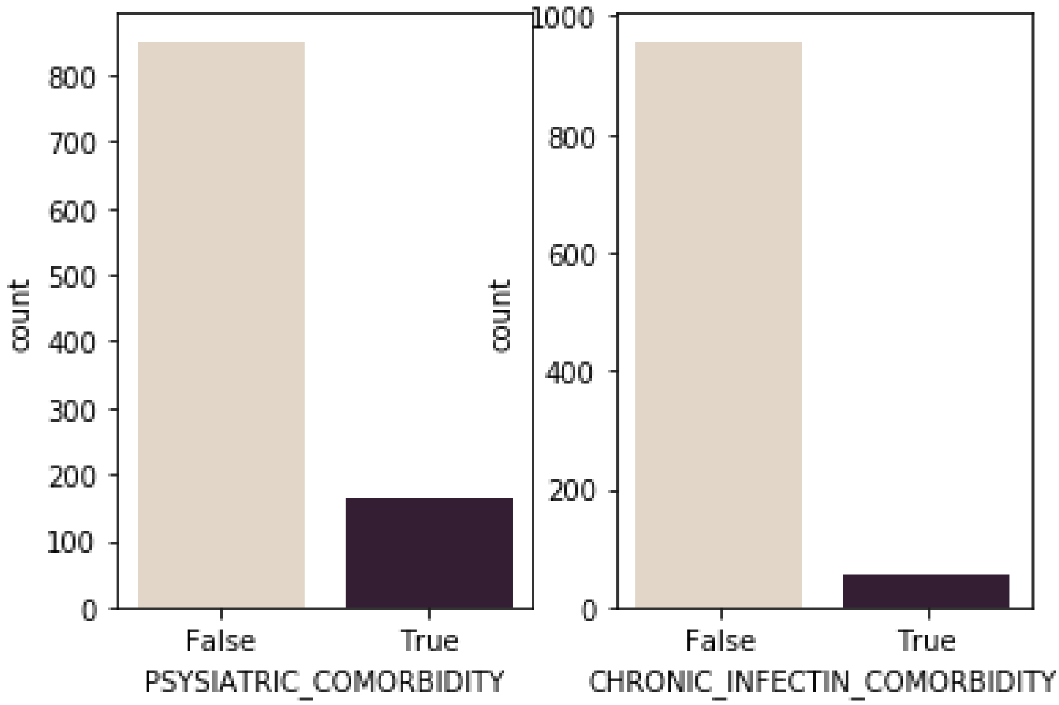


Figure 13: Distribution patient by Comorbidities: Psysiatric and chronic infectin comorbidities

Graphs above show that about 80% of people doesn't have chronic infectin comorbidity whereas about 66% of samples doesn't have psysiatric comorbidity .

4.5 Data Preparation

OPTIMA Data Source is a repository filled by a Doctor's Team using simple tools, like Excel file. During this phase, doctors doesn't have a good knowledge of data analysis standard and then they made mistakes in filling the tables. The first task has been data cleaning operation.
In order to enable operations like aggregations, filtering and search on a data cleaning stage is needed.
Applying this task is also helpful for the future studies that will continue this project.
Defining a rigorous data cleaning procedure is an essential task.

4.5.1 Section Data Cleaning

Into this section the data cleaning procedure used to obtain a clean data from a raw data has been explained.

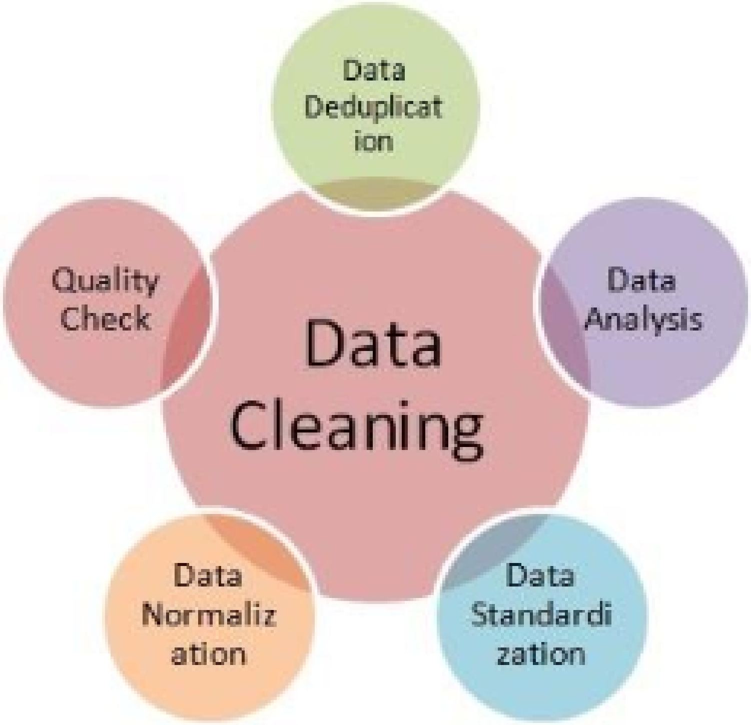


Figure 14: data_cleaning

Documents File Cleaning

Goals: Cleaning subcategory fields:

- Removing special characters.
- Removing possible double spacing between words.
- Removing possible duplicated data that are no used.
- Removing rows that have PM(post mortem) wording.

Section File Cleaning

File Cleaning is a huge procedure having the aim to remove all the informations that are wrong or senseless. All the cleaning tasks that are explained below are applied on the entire amount of columns that have been used to build Final Patient Data

Input File: Section File

Goals:Cleaning fields

- Filtering sections by type, selecting only the ones used for our scope.
- Special characters removal excluding characters involved in the dates format
- Fixing words spacing and leaving single space between words.
- Removing all value into column used for our scope that have value no consistent.
- Removing rows patient that have no value for mmse score.

Results:

All texts correctly cleaned and formatted. They are now ready to be used for information extraction.

4.5.2 Data Extraction

Data Extraction is a process that has the role of retrieval of data from various sources. This procedure is the most important process during the data analysis project called ETL(Extraction, Transform and Load). In this project, we used a Full Extraction approach. In full extraction, data is copied from the source system in its entirety.

The paragraphs below have been split into two parts: the former explains which is the final result that we want to obtain and the latted one explains in detail the algorithm used to achieve the above objective. To build a dataset that shows the evolution of the EHR(electronic health record) parameters, lifestyle parameters, MMSE and comorbidities episode by episode, episode date are used as a joining key so that we have a consistent dataset. This type of merge is applicable because all the tables that have been used, have an episode date column with consistent data.

To define raw columns that have been taken to create feutures for Final Patients Data, the idea is looking for related keywords inside Variable Guide-UPM.xls. For example, if we want to identify raw columns for the education level feature ,we will seek keywords as school, education, university, degree and so on.

4.5.2.1 Smoking Level Extraction

Final Target: Building a column called SMOKING that has four different values : current smoker, former smoker, no smoker, no value.

To identify Smoking Level of each patient two different tables have been taken nart duration smoking.xlsx and Data Request Jan 2019 (Optima, Lead, Challenge) final-history.xlsx. Inside this two tables there is a subset of columns of interest: "SMOKING: SMOKING" and "SMOKING: SMOKING:" TEXT inside nart duration smoking.xlsx and "HISTORY PATIENT 74-119: (079) HISTORY OF HEAVY SMOKING" inside Data Request Jan 2019 (Optima, Lead, Challenge) final-history.xlsx.

Raw Columns			
File Source Name	Raw Column Name	Type	N. Patient
nart duration smoking	SMOKING: SMOKING	Categorical	130
nart duration smoking	SMOKING: SMOKING :TEXT	Text	130
final history	HISTORY OF HEAVY SMOKING	Categorical	1035

Table 6 : Overview of files using to Smoking Level

"SMOKING: SMOKING" has four distinct values: 1 indicates people that are current smoker, 2 indicates people that are former smoker, 0 indicates never smoker and 9 indicates that the information about this features are missing.
"SMOKING: SMOKING : TEXT" is a text field where the previous column has a text defining how long this patient has quitted smoking if it is equal to 2. Inside this field the text format is no unique how it is possible to see in the following tables.

SMOKING: SMOKING : TEXT
40 years ago
7 years ago("Never been a heavy smoker")
Regular smoker but stopped 18 years ago
Stopped 22 years ago
Stopped smoking 20 years ago
Stopped smoking regularly 14 years ago

Table 7 : Different text inside SMOKING: SMOKING : TEXT

Therefore it is necessary to define a rules to extract the information of interest from this column. Rules are explained into the algorithm section below.
"HISTORY OF HEAVY SMOKING" defines if a patient has smoked more than 20 cigarettes per day for a long time(more than 15 years). This column has 4 values: 1 is equal to YES, 0 is equal to NO, 9 is equal to NOT ASKED, 8 is equal to NOT KNOWN. For my objective "HISTORY OF HEAVY SMOKING" is more useful than the other two columns because it has at least one value for 1035 patients.

Algorithm:

STEP 1: First of all, only the columns of interest have been taken from the two tables :GLOBAL PATIENT DB ID and EPISODE DATE that are the key of each tables and the three columns mentioned before.

STEP 2: On "SMOKING: SMOKING: TEXT" a substring of a text field has been extracted using a pandas tools : numeric value indicating how long patient has quitted smoking.

Therefore, before applying the extraction rules the result have been stored into a new column called num_ year.

Then this algorithm has been applied :

```
for index,row in list_smoker2.iterrows() :
   if  SMOKING:SMOKING == 2 and num_year > 5 :
      patient_categorization_smoke.append('former_smoker')
      patient_id.append(GLOBAL_PATIENT_DB_ID)
   else if SMOKING:SMOKING == 2 and num_year <= 5 :
      patient_categorization_smoke.append('current_smoker')
      patient_id.append(GLOBAL_PATIENT_DB_ID)
```

The code explain that if "SMOKING:SMOKING" is equal to 2 a patient is a former smoker. But if he/she quitted smoking no more than five years ago he/she is labeled as a current smoker. In other case he/she is labeled as a former smoker. Therefore, with the application of this constraint we have increased the number of current smokers compared to those of former smokers

STEP 3: Taken data from "SMOKING : SMOKING" column using the following algorithm:

```
for index,row in list_smoker2.iterrows() :
if  SMOKING:SMOKING == 1 and GLOBAL_PATIENT_DB_ID not in patient_id :
   patient_categorization_smoke.append('current_smoker')
   patient_id.append(GLOBAL_PATIENT_DB_ID)
if  SMOKING:SMOKING == 2 and GLOBAL_PATIENT_DB_ID not in patient_id :
  patient_categorization_smoke.append('former_smoker')
  patient_id.append(GLOBAL_PATIENT_DB_ID)
if  SMOKING:SMOKING == 0 and GLOBAL_PATIENT_DB_ID not in patient_id :
  patient_categorization_smoke.append('no_smoker')
  patient_id.append(GLOBAL_PATIENT_DB_ID)
```

The only patient that have been analyze are only patient that are no taken yet.

STEP 4: Last step, taken data from "HISTORY OF HEAVY SMOKING" column. It is the column that has more consistent data because it is has 9565 rows for 1035 patients.

```
for index,row in list_smoker.iterrows() :
  if  HISTORY PATIENT 74-119:(081)HISTORY OF HEAVY SMOKER == 1 or
   GLOBAL_PATIENT_DB_ID not in patient_categorization_smoke:
     patient_categorization_smoke.append('current_smoker')
     patient_id.append(GLOBAL_PATIENT_DB_ID)
  if  HISTORY PATIENT 74-119:(081)HISTORY OF HEAVY SMOKER == 0 or
  row['GLOBAL_PATIENT_DB_ID'] not in patient_categorization_smoke:
  patient_categorization_smoke.append('no_smoker')
  patient_id.append(GLOBAL_PATIENT_DB_ID)
```

The only patient that have been taken are only patient that are no taken yet.

4.5.2.2 Education Level Extraction

Final Target:It defines a column called EDUCATION that has four different values: basic, medium, high and no value.To build this column,the number of education years during each patients life has been considered.

To identify education level of each patient informations have been taken from Data Request Jan 2019 (Optima, Lead, Challenge) final-present.xlsx. Inside this table the subset of columns is the following : "Age Left School" and "Years in Further education".

Algorithm:

STEP 1: First of all, only the columns of interest have been taken from the two tables : GLOBAL PATIENT DB ID and EPISODE DATE are the key of each tablesand the two columns above.

STEP 2: Using "Age Left School" and "Years in Further education" columns the formula below has been applied :

$$education = (AgeLeftSchool + YearsInFurtherEducation) - 5 \tag{1}$$

The idea is to create a new column called education that is a sum of the two previous columns minus the age when people start to go to school tipically (5 years old).

STEP 3: During the last step, an if condition has been set to split patients education level into one of the category defined into the target features. The following code explain how are set ranges:

```
for i in range(len(education_patient)) :
   if education_patient <= 16 :
      education_categorization.append('basic')
      patient_id.append(GLOBAL_PATIENT_DB_ID);
   if education_patient > 16 and education_patient > 22 :
         education_categorization.append('medium')
         patient_id.append(GLOBAL_PATIENT_DB_ID);
   if education_patient >= 22 :
    education_categorization.append('high')
    patient_id.append(GLOBAL_PATIENT_DB_ID);
```

4.5.2.3 Drinking Level Extraction

Final Target:It defines a column called DRINKING that has four different values: no drinking, extreme drinking, mild drinking and no value.

To define Drinking level of each patient information on Data Request Jan 2019 (Optima, Lead, Challenge)_final-history.xlsx. have been used. This table contains 1030 different patient with at least one record.

For my objective, the subset of interest are the following: HISTORY PATIENT 74-119: (080) HIST. OF REGULAR DRINKING, HISTORY PATIENT 74-119: (081) HISTORY OF HEAVY DRINKING and HISTORY PATIENT 74-119: (082) HIST. OF PROBLEM DRINKING.

Algorithm:

STEP 1: First of all, only the columns of interest has been taken from the table:GLOBAL PATIENT DB ID and EPISODE DATE that are the key of each tables and the three columns above.

STEP 2: During this step the way informations are used inside the columns have been defined. Into the following table the possible values have been showed.

Raw Columns final history	
Raw Column Name	Data Type
(080) HIST. OF REGULAR DRINKING	9=NotAsked 8=NotKnown 0=No 1=Yes
(081) HISTORY OF HEAVY DRINKING	9=NotAsked 8=NotKnown 0=No 1=Yes
(082) HIST. OF PROBLEM DRINKING	9=NotAsked 8=NotKnown 0=No 1=Yes

Table 8 : Tables used to build DRINKING LEVEL columns

STEP 3: An if condition has been set splitting information into one of the category define into the target feature. the following code explain how are setted ranges:

```
for index,row in list_alcohol.iterrows() :
   if  HISTORY PATIENT 74-119:(081)HISTORY OF HEAVY DRINKING == 1 or
       HISTORY PATIENT 74-119:(081)HIST. OF PROBLEM DRINKING == 1 :
      patient_categorization.append(’extreme drinking’)
   else if HISTORY PATIENT 74-119:(081)HIST. OF REGULAR DRINKING == 1 :
      patient_categorization.append(’mild drinking’)
   else :
      patient_categorization.append(’no drinking’)
```

The algorithm above shows the way data are split. We can notice that is set as extreme drinking, patients that have value 1 for heavy drinking column and for problem drinking column, as mild drinking, people that have value 1 for regular drinking column and no drinking people that have 0 value for the three columns mentioned.

4.5.2.4 Comorbidity Extraction

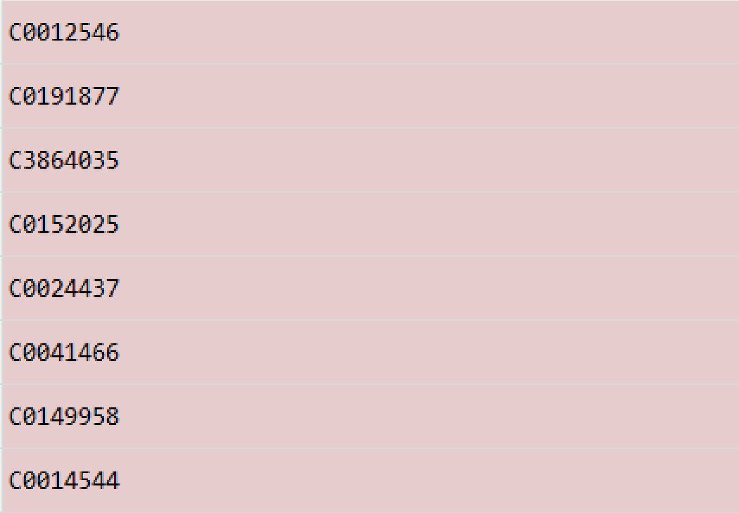
Final Targets:It defines 9 columns that describe different comorbidities. Each comorbidity column has three different value: True, False, no value.

During this task has been used to compute these 9 columns three different tables : Data Request Jan 2019 (Optima, Lead, Challenge)_final-history.xlsx, diseasesORcomorbidities - comorbidities.xlsx. and optima_patients_CUIs.xlsx

If we want to use the information inside the last two tables we need to apply a specific procedure :

Procedure comorbidity_identify:

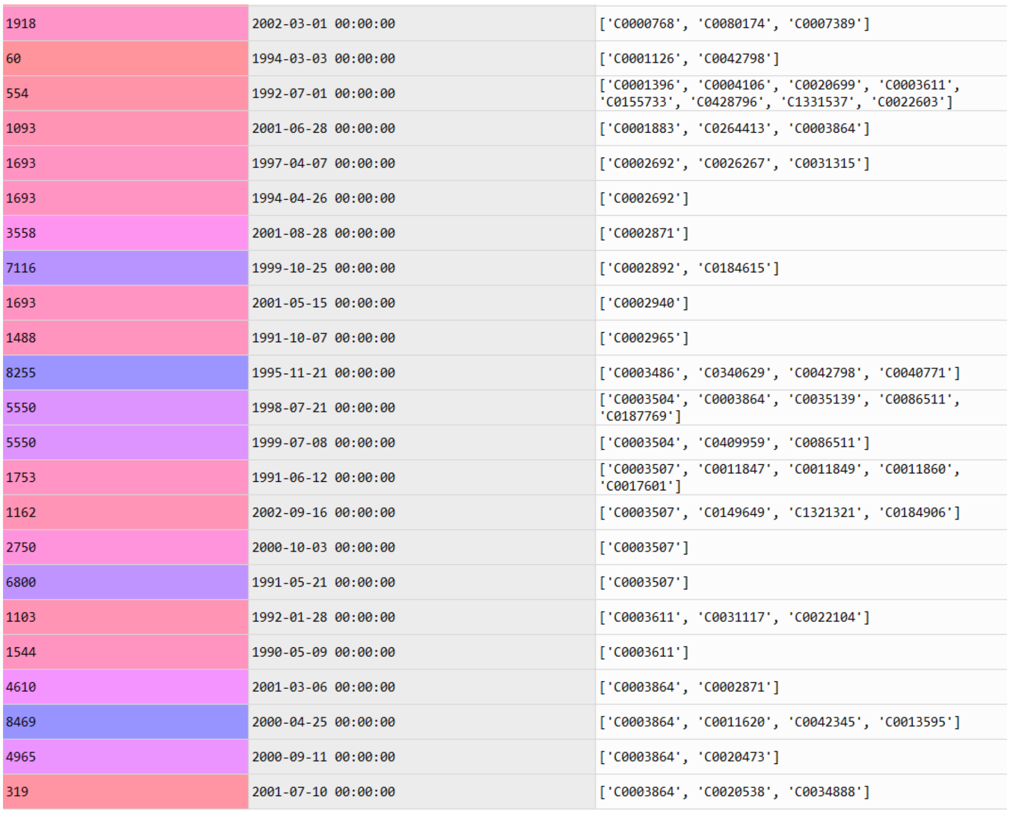
1. Definition of **comorbidity_type** variable using table diseasesORcomorbidities - comorbidities.xlsx. This excel file contains a list of different types of diseases and their classification in a specific column called TYPE. It is used to label the disease into one of the 9 comorbidities. For example, if we set TYPE = N that is linked to the physiatric comorbidity, we obtain the following result:



C0012546
C0191877
C3864035
C0152025
C0024437
C0041466
C0149958
C0014544

Figure 15: Overview list of CUI for psysiatric_comorbidity

2. Definition of variable **comorbidity_patients_optima** based on optima_patients_CUIs.xlsx to obtain a list of CUI afflicting each patient.



1918	2002-03-01 00:00:00	['C0000768', 'C0080174', 'C0007389']
60	1994-03-03 00:00:00	['C0001126', 'C0042798']
554	1992-07-01 00:00:00	['C0001396', 'C0004106', 'C0020699', 'C0003611', 'C0155733', 'C0428796', 'C1331537', 'C0022603']
1093	2001-06-28 00:00:00	['C0001883', 'C0264413', 'C0003864']
1693	1997-04-07 00:00:00	['C0002692', 'C0026267', 'C0031315']
1693	1994-04-26 00:00:00	['C0002692']
3558	2001-08-28 00:00:00	['C0002871']
7116	1999-10-25 00:00:00	['C0002892', 'C0184615']
1693	2001-05-15 00:00:00	['C0002940']
1488	1991-10-07 00:00:00	['C0002965']
8255	1995-11-21 00:00:00	['C0003486', 'C0340629', 'C0042798', 'C0040771']
5550	1998-07-21 00:00:00	['C0003504', 'C0003864', 'C0035139', 'C0086511', 'C0187769']
5550	1999-07-08 00:00:00	['C0003504', 'C0409959', 'C0086511']
1753	1991-06-12 00:00:00	['C0003507', 'C0011847', 'C0011849', 'C0011860', 'C0017601']
1162	2002-09-16 00:00:00	['C0003507', 'C0149649', 'C1321321', 'C0184906']
2750	2000-10-03 00:00:00	['C0003507']
6800	1991-05-21 00:00:00	['C0003507']
1103	1992-01-28 00:00:00	['C0003611', 'C0031117', 'C0022104']
1544	1990-05-09 00:00:00	['C0003611']
4610	2001-03-06 00:00:00	['C0003864', 'C0002871']
8469	2000-04-25 00:00:00	['C0003864', 'C0011620', 'C0042345', 'C0013595']
4965	2000-09-11 00:00:00	['C0003864', 'C0020473']
319	2001-07-10 00:00:00	['C0003864', 'C0020538', 'C0034888']

Figure 16: CUI list overview for pair GLOBAL PATIENT DB ID and episode date

To retrieve comparable data with **comorbidity_type** we need to apply text cleaning operation to remove comma or special character like round bracket and square bracket.

3. At the end, using the two previous variables defined, a procedure code has been apply to obtain list of patients with that comorbidity under analysis. For example, the next code show how the procedure on psysiatric comorbidity has been carried out:

```
for index,row in comorbidity_patients_optima:
   if row['diagnosis_CUIs'] in comorbidity_type_psysiatric
      patient_with_comorbidity.append(row['GLOBAL_PATIENT_DB_ID'])
```

Where comorbity_patients_optima is a list containing GLOBAL PATIENT DB ID, episode date and diagnosis_CUI afflicting patient during episode date and comorbidity_type is a list of CUI for psysiatric comorbidity. Then if we find a match between diagnosis_CUI and CUI we can put True on psysiatric comorbidity for that patient on that episode date.

Algorithm:

STEP 1: First of all, from the three tables only the columns of interest have been taken : GLOBAL PATIENT DB ID and EPISODE DATE that are the key of each tables and the other columns need to compute all types of comorbidities.

STEP 2: in this step it has been explained which are the columns taken and how to use them to calculate features.

1. MALIGNANT_COMORBIDITY

   - It uses variable HISTORY OF CANCER from Data Request Jan 2019 (Optima, Lead, Challenge)_final-history.xlsx file, we describe if there is incidence of cancer into family history patient.
     Possible value: 0=no, 1=yes, 9=not known, 8=not asked

     ```
     if HISTORY OF CANCER == 1 :
          MALIGNANT\_COMORBIDITY = TRUE;
     if HISTORY OF CANCER == 0 :
          MALIGNANT\_COMORBIDITY = FALSE;
     if HISTORY OF CANCER != 1 and HISTORY OF CANCER !=0 :
          MALIGNANT\_COMORBIDITY = no value;
     ```

2. CHRONIC_INFECTIN_COMORBIDITY

   - It takes information from diseasesORcomorbidities - comorbidities.xlsx where TYPE = V and optima_patients_CUIs.xlsx using procedure comorbidity_identify explained above.

3. PSYSIATRIC_COMORBIDITY

    - It takes info from HISTORY OF PSYCH. ILLNESS From Data Request Jan 2019 (Optima, Lead, Challenge)_final-history.xlsx.
      Possible value: 0=no, 1=yes, 9=not known, 8=not asked
    - It Holds information inside diseasesORcomorbidities - comorbidities.xlsx where TYPE = N and optima_patients_CUIs.xlsx using procedure comorbidity_identify explain above.

4. METABOLIC_COMORBIDITY

    - It takes information from diseasesORcomorbidities - comorbidities.xlsx where TYPE = M and optima_patients_CUIs.xlsx using procedure comorbidity_identify explained above.

5. SISTEMIC_COMORBIDITY

    - It Uses information inside Data Request Jan 2019 (Optima, Lead, Challenge)_final-history.xlsx .The Data are inside columns called (075) HISTORY OF RAISED BP and HISTORY OF DIABETES.
      Possible value: 0=no, 1=yes, 9=not known, 8=not asked.

6. RESPIRATORY_COMORBIDITY

    - It takes information from diseasesORcomorbidities - comorbidities.xlsx where TYPE = R and optima_patients_CUIs.xlsx using procedure comorbidity_identify explained above.

7. VASCULAR_COMORBIDITY

    - It Uses information inside Data Request Jan 2019 (Optima, Lead, Challenge)_final-history.xlsx. The taken columns ,are inside columns called (075) HISTORY OF HEART ATTACK and HISTORY OF STROKE
      Possible value: 0=no, 1=yes, 9=not known, 8=not asked

- It Holds information inside diseasesORcomorbidities - comorbidities.xlsx where TYPE = V and optima_patients_CUIs.xlsx using procedure comorbidity_identify explain above.

8. IMFLAMMATORY_COMORBIDITY

   - It takes information from diseasesORcomorbidities - comorbidities.xlsx where TYPE = V and optima_patients_CUIs.xlsx using procedure comorbidity_identify explain above.

9. K_COMORBIDITY

   - It takes information from diseasesORcomorbidities - comorbidities.xlsx where TYPE = K and optima_patients_CUIs.xlsx using procedure comorbidity_identify explain above.

4.5.2.5 APOE Extraction

Final Target:It builds categorical features that show what is the APOE genotype for each patient.

Before setting APOE's column we need to remove all the wrong values inside the original raw colum.

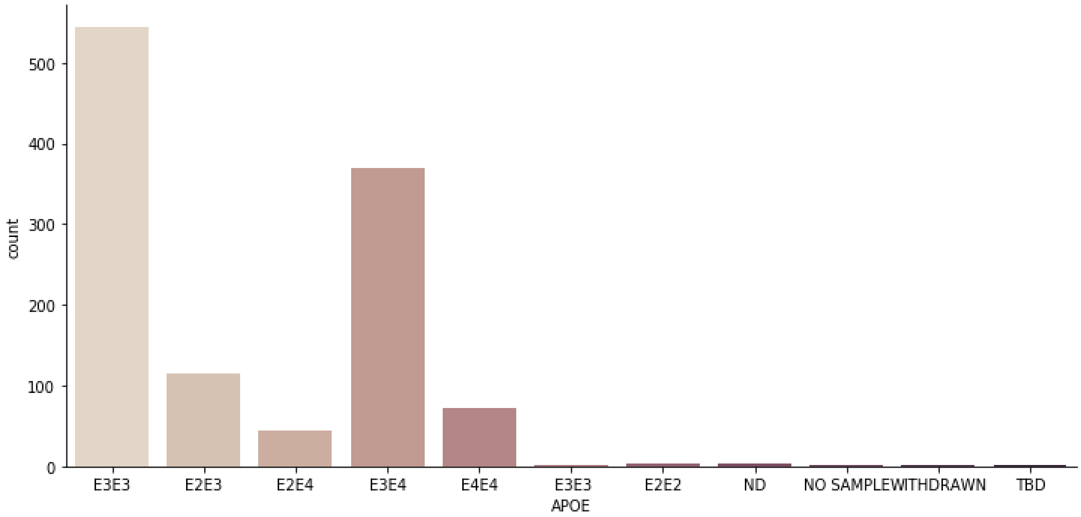


Figure 17: Distribution Patient by APOE

APOE data	
Raw Data	Clean data
number of distinct patient 1023	number of distinct patient 1013

Table 9 : Comparison between APOE Raw Data and Clean Data

Algorithm:

STEP 1: First of all, from the two tables only the columns of interest have been taken : GLOBAL PATIENT DB ID and EPISODE DATE that are the key of each tables and the APOE column above.

STEP 2: During this step, the only thing that has been done is to take the values found in the original table after the data cleaning operation.

4.5.2.6 Dementia Type Extraction

Final Targets: It Defines 3 columns that describe if the patient has Dementia, Other Dementia or no Dementia the first one includes only Alzheimer's disease while other dementia is a huge group of diseases like vascular disease, Parkinson disease, lewy body dementia and so on.

To define these three columns data inside optimaDX.xlsx have been manipulated , using 13 different columns. To correctly use the information in this table, a set of rules has been provided by the team of doctors. Doctor's rules define a set a constraints to categorize patients into the correct group.

dementia(Alzheimer's disease), other dementia, no dementia into the table below are respectively D, OD, ND. Rules defined by doctor's team will be explain into Step 2 of algorithm section.

Algorithm:

STEP 1: First of all, only the columns of interest have been taken from the two tables : GLOBAL PATIENT DB ID and EPISODE DATE that are the key of each tables that are considered and the other columns need to compute all types of comorbidities.

STEP 2: On this section, doctor's rules have been applied to build dementia, other demetia and no value columns. In the following table every constraints used to define dementia type are explain:

Doctor's rules to define dementia		
Variable Name	possible values	doctor's rule
COGNITIVE IMPAIRMENT	1=Yes 0=No 8=NotKnown 9=NotAsked	if 1 then D=1, OD=1, ND=0
PETERSEN MCI	1= Yes 0=No 8=NotKnown 9=Not Asked	if 1 then D=0, OD=0, ND=1

Table 10: Overview of columns used to build dementia, other dementia, no dementia columns 1/2

Doctor's rules to define dementia		
Variable Name	possible values	doctor's rule
VCI	1=Yes 0=No 8=Not Known 9=Not Asked	if 1 then D=0, OD=1, ND=0
DEMENTIA PRESENT	0=No 1=Mild 2=Moderate 3=Severe	if 1,2,3 or 4 then D=1, OD=1, ND=0
AD (NINCDS-ADSDA)	0=Negative 1=Possible 2=Probable 8=Not Asked	if 1,2,3 or 4 then D=1, OD=0, ND=0
MIXED DEMENTIA TYPE	1=Yes 0=No 9=Not Asked	if 1,3 then D=1, OD=1, ND=0
VASCULAR DEMENTIA	0=Negative 1=Possible 2=Probable 8=Not Asked	if 1,2,3 or 4 then D=0, OD=1, ND=0
DEMENTIA OTHER	1=Yes 0=No	if 1 then D=0, OD=1, ND=0
PARKINSON DISEASE	1=Yes 0=No	if 1 then D=0, OD=0, ND=1
PROGRESSIVE SUPRA-NUCLEAR PALSY	1=Yes 0=No	if 1 then D=0, OD=1, ND=0
FRONTO-TEMPORAL DEMENTIA	0=Negative 1=Possible 2=Probable 8=Not Asked	if 1,2 then D=0, OD=1, ND=0
CORTICO-BASAL DEGENERATION	0=Negative 1=Possible 2=Probable 8=Not Asked	if 1,2 then D=0, OD=1, ND=0
LEWY-BODY DISEASE	1=Yes 0=No 8=NotKnown 9=NotAsked	if 1 then D=0, OD=1, ND=0

Table 11 : Overview of columns used to build dementia, other dementia, no dementia columns 2/2

The table optimaDX has 9565 rows and at least one row for 1030 patients but if we do more accurate analysis, we can notice that there are many patients that have only values like 9, 8 and nan for each episode. For these subgroup of patients is impossible to used doctor's rules "no_value" string has been used to point out patients that have no consistent values.

The following figure shows the way patients distribute between consistent and unconsistent data:

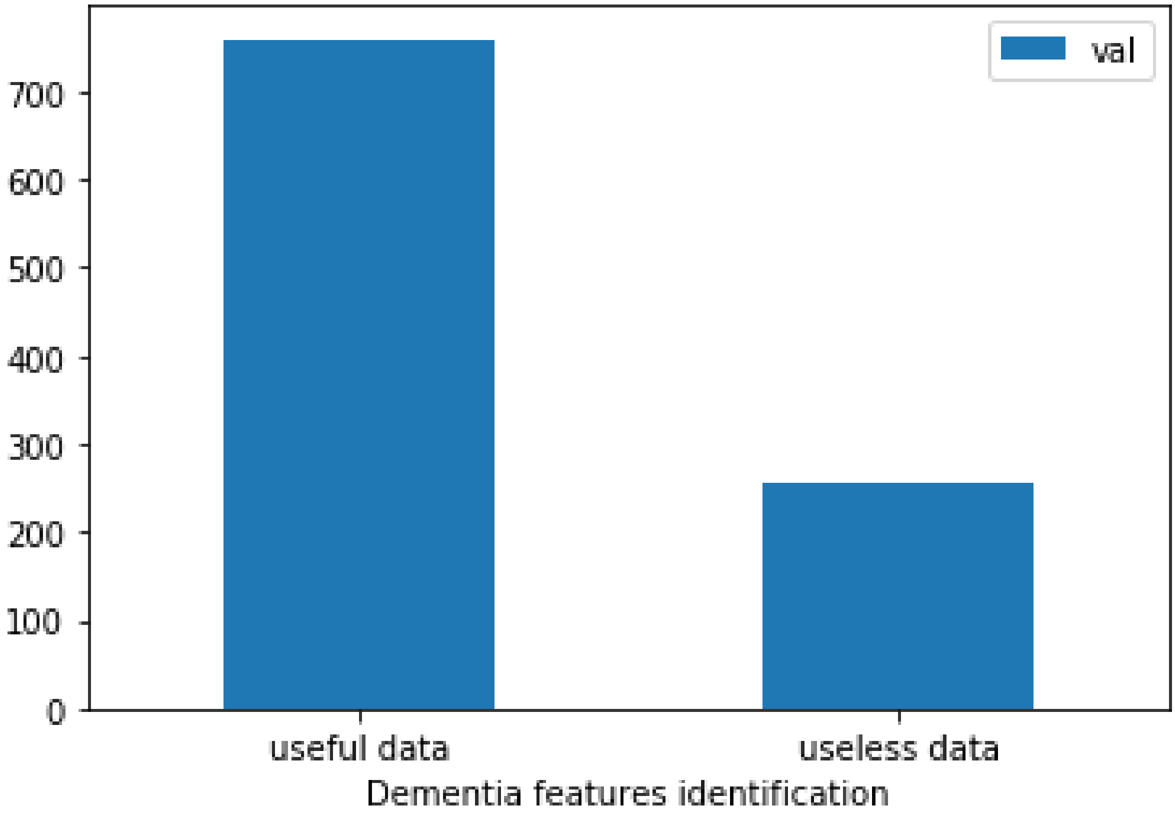
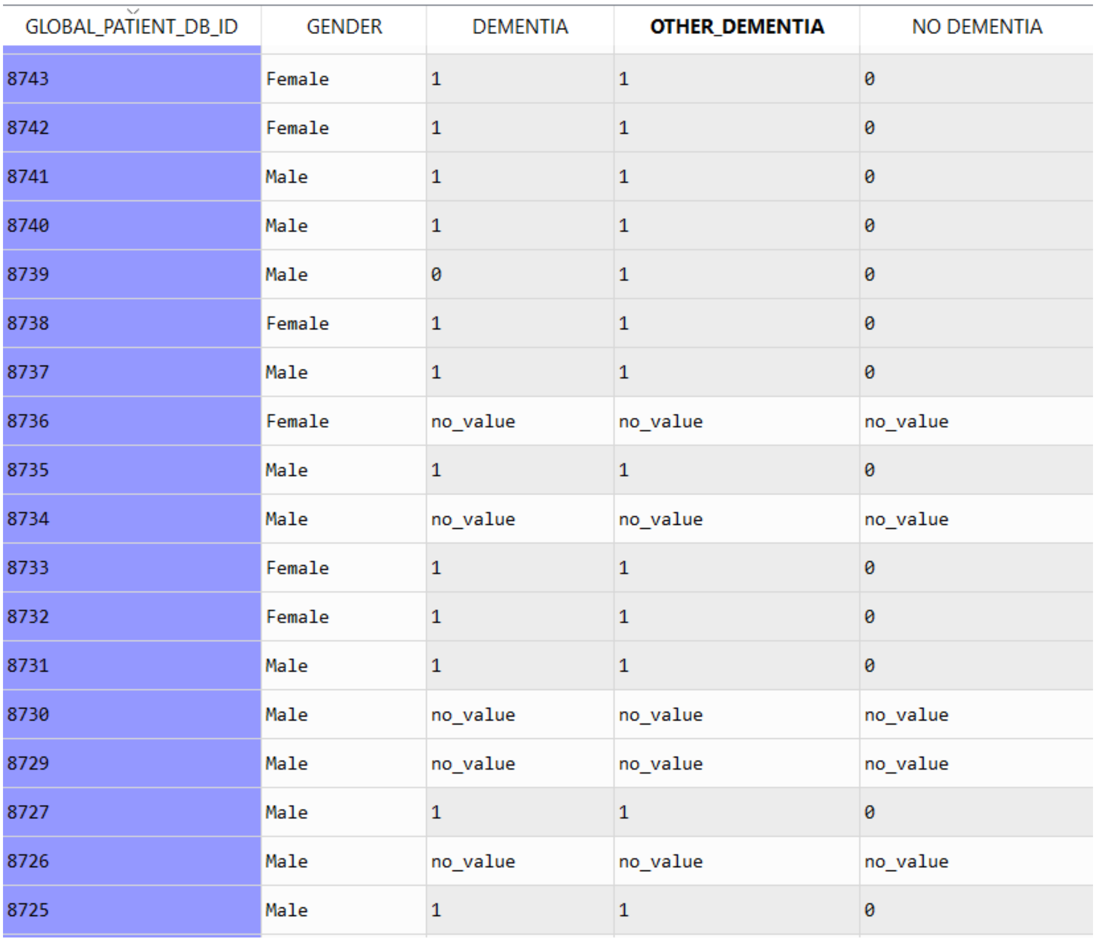


Figure 18: Overview of data distribution on optimaDX

The graph shows that more than 25% of patients has no data for either column for all episode under analysis, then for these patients, it is impossible to have information about Alzheimer or another type of dementia.

STEP 3: During this stage doctor's rules have been used to build the columns mentioned above.

The Final result into the Final Patient Data is the following:



GLOBAL_PATIENT_DB_ID	GENDER	DEMENTIA	OTHER_DEMENTIA	NO DEMENTIA
8743	Female	1	1	0
8742	Female	1	1	0
8741	Male	1	1	0
8740	Male	1	1	0
8739	Male	0	1	0
8738	Female	1	1	0
8737	Male	1	1	0
8736	Female	no_value	no_value	no_value
8735	Male	1	1	0
8734	Male	no_value	no_value	no_value
8733	Female	1	1	0
8732	Female	1	1	0
8731	Male	1	1	0
8730	Male	no_value	no_value	no_value
8729	Male	no_value	no_value	no_value
8727	Male	1	1	0
8726	Male	no_value	no_value	no_value
8725	Male	1	1	0

Figure 20: Overview of DEMENTIA,OTHER,DEMENTIA,NO DEMENTIA into Final Patient Data

5 Discussion of Results

Chapter can be splitted into two part:

1. An exploaration on the final dataset Final Data Patients.
2. The development of the KPI defined during the previous chapter. During this stage many techniques have been applied(heatmap, stacked bar chart, survival analysis, distribution analysis) to achieve a complete and consistent analysis of them.

5.1 Exploratory Analysis of Output Data

Once we processed the raw data to extract information from Optima Data Source , it is time to develop the key performance indicator mentioned in section 4.2.

5.1.1 Exploring Final Data Patients

Table Structure: in this table there are 8134 rows associated to 1013 distinct patients. As already discussed in section 4.3.2, dataset contains 30 columns that we can split into the following subcategories : general information, dementia information, episode data, lifestyle patient information and comorbidity information.

5.1.1.1 Number of Episode per Patients

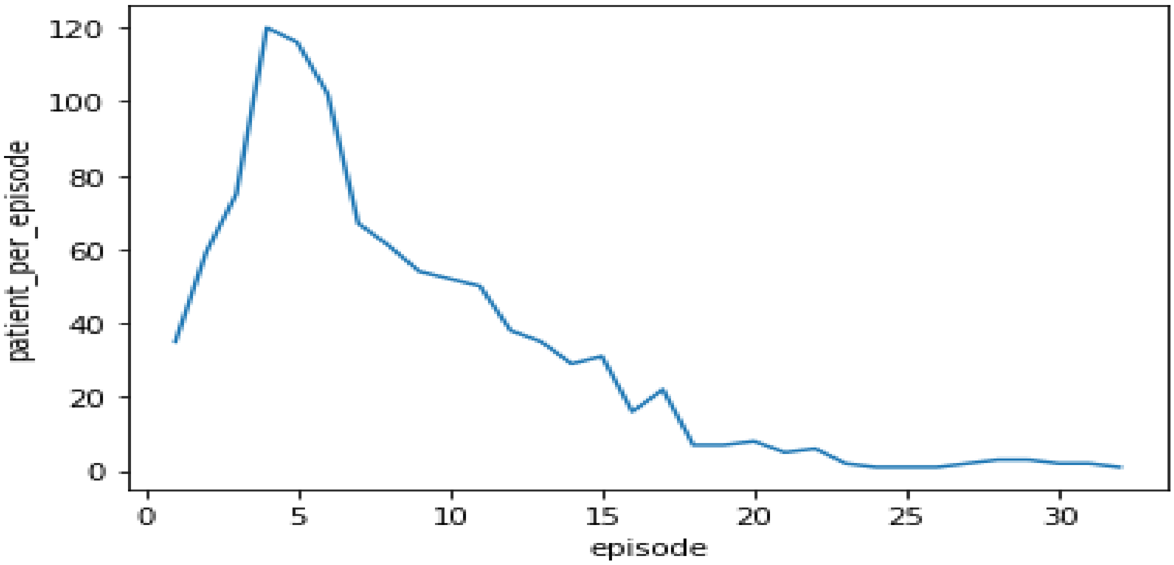


Figure 21: Number of epiode for patients

The figure 21 shows that on average, each patients has an average value of 8 episodes with minimun number of episode equal to 1 and maximum number of episode equal to 35.

Graphs below show the chief statistics about episode per Patients

Final Patient Data Statistics	
Statistics Type	Episode Per Patients
count	1013
mean	8.02962
std	5.33049
min	1
25%	4
50%	6
75%	11
max	35

Table 12 : Final Patient Data Statistics Episode Number

5.1.1.2 Distribution of Patients by Dementia Type

Using Doctor's rules, as discussed on section 4.5.3.6, these four subcategories have been setted. During the first episode patients that have no value are more than 50% whereas in the last episode patients that has no value are less than 50% .

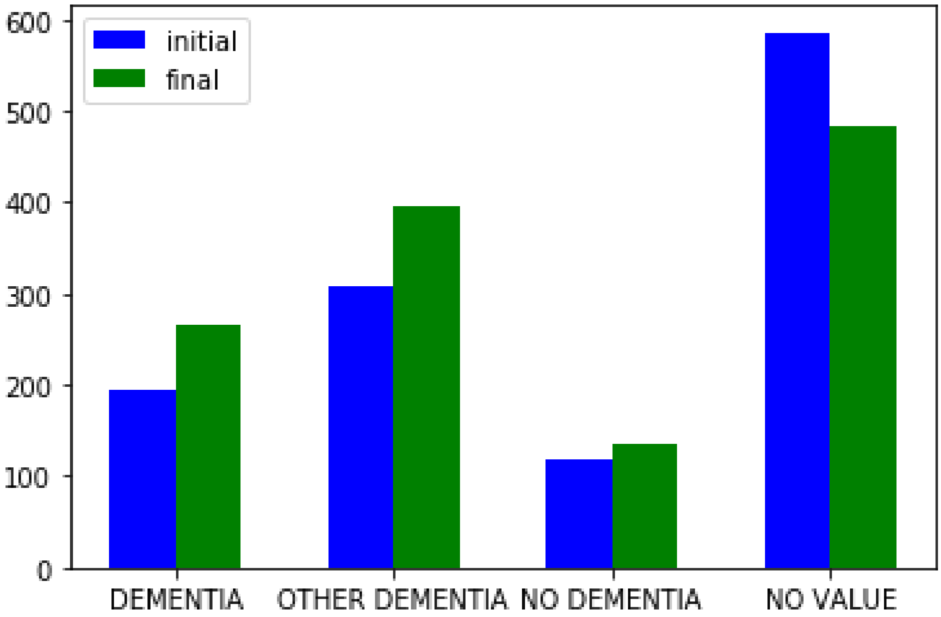


Figure 21: Overview dementia distribution

An important consideration we come out from the previous chart: every patients that has

1 value about dementia column also have 1 value in the other dementia column. This means that the constraints defined by the doctors to filter any patient with only dementia (alzheimer disease) don't work maybe because they have nan or not asked value.
Using the variations of the graph from the first to the last episode we can define a starting point to perform Multiple Filter Analysis operations.

5.2 KPI 1 - Multiple Filter Analysis

In this section different and various Multiple Filter Analysis have been done. The general scope of this task is to apply some filtering to find new type of information and to identify possible new ways to read the data. To do this, some advice was provided by the doctors after their analysis of Final Patient Data contents . The following Multiple Filter Analysis have been developed:

1. Main comorbidities associated with moderate level of MMSE
2. Main comorbidities associated with severe level of MMSE.
3. Number of patients that are strong smoker when have at least one types of dementia.
4. Number of patients that are strong drinker when have at least one types of dementia.
5. Number of patients getting worse.
6. Number of patients remaining stable.
7. Number of patients getting better.
8. Mixed Status Column Creation.
9. Remove patients that no change their MMSE status from initial episode to final episode.
10. Remove all patients that no have at least five episodes.

5.2.1 Main comorbidities associated to moderate or severe level of MMSE

Moderate level of Mini Mental State Examination

Patients Number on the last episode: 282

Distribution of Comorbidities		
Comorbidity Type	True	False
malignant comorbidity	54	228
psysiatric comorbidity	26	256
sistemic comorbidity	45	237
vascular comorbidity	24	258
chronic infectin comorbidity	4	278
k comorbidity	4	278
respiratory comorbidity	0	282
imflammatory comorbidity	4	278
metabolic comorbidity	4	278

Table 13 : Distribution moderate MMSE Comorbidities

Analyzing patient having moderate level of MMSE on the last episode, it possible to notice that the more common comorbidities are malignant and psysiatric. Malignant comorbidity is present on 19% of the samples whereas sistemic comorbidity on 15% of the samples.

Severe level of Mini Mental State Examination

Patients Number on the last episode: 567

Distribution of Comorbidities		
Comorbidity Type	True	False
malignant comorbidity	11	556
psysiatric comorbidity	7	560
sistemic comorbidity	17	550
vascular comorbidity	14	553
chronic infectin comorbidity	7	560
k comorbidity	1	566
respiratory comorbidity	0	567
imflammatory comorbidity	7	560
metabolic comorbidity	0	567

Table 14 : Distribution severe MMSE Comorbidities

The table below shows that patients reaching severe status during the last episode have no specific comorbidities more common than others.

5.2.2 Number of patients that are strong smoker or strong drinker when have at least one types of dementia.

Starting from Final Patient Data, a filter has been applied to remove patients that no have any type of dementia and then on it an analysis to identify strong smokers and strong drinkers have been developed. The goal is to understand if smoke or alcohol influence the development of dementia.

<table>
<tr><td colspan="2">Original Final Patient Data</td></tr>
<tr><td>Number of Rows</td><td>Number of patient</td></tr>
<tr><td>8595</td><td>1013</td></tr>
<tr><td colspan="2">Filtering Final Patient Data</td></tr>
<tr><td>Number of Rows</td><td>Number of patient</td></tr>
<tr><td>2398</td><td>556</td></tr>
</table>

Table 15 : Comparison number of rows and patient between the original dataset and the filtering dataset

The two graphs below shows only patient that during the last episode have flag Alzheimer disease and other dementia diseases equal to 1.

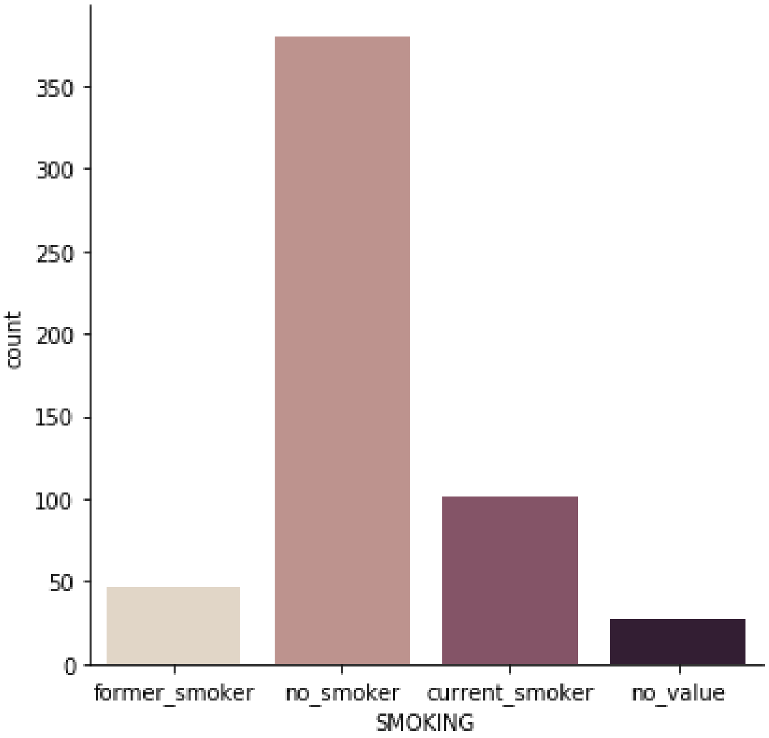


Figure 22: Distribution of Patients by SMOKING level

SMOKING level graphs shows that more than 65% of the patients never smoked regularly in their lives. Furthermore more then 25% are current or former smoker.
Relying on statistics just obtained, no meaningful conclusions could be drawn from the analysis provided.

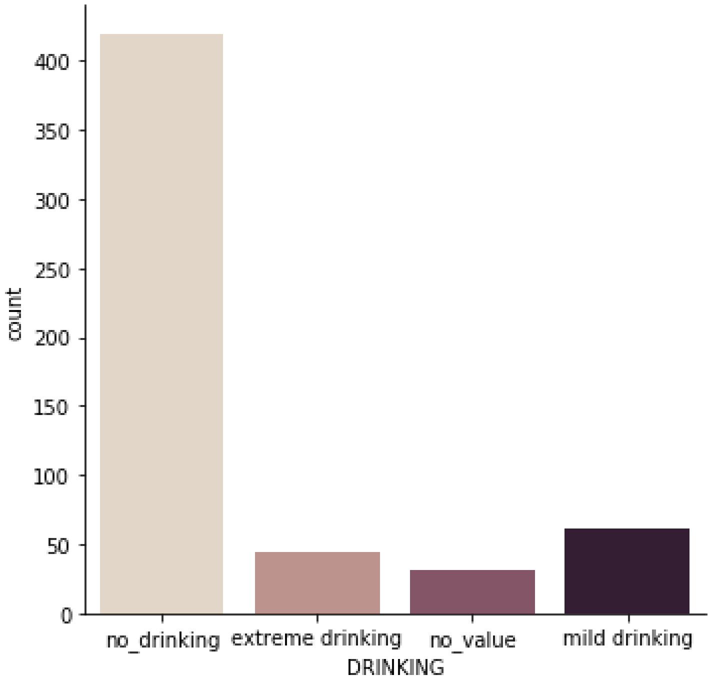


Figure 23: Distribution of Patients by ALCOHOL level

Instead figure 27 show that more than 75% of the patient are no drinking then meaningful conclusions could be drawn: high consumption of alcohol haven't strong correlation with dementia's disease.

5.2.3 Mixed Status Column and Patients Evolution

After an overview of the doctor's team, has been defined a new column combining two column present into Final Patients Data, DEMENTIA and MMSE_DISCRETIZED column. In this way we obtained a more complex analysis incorporate different but correlate features.

```
if DEMENTIA == 1 and MMSE_DISCRETIZED =='mild':
   mixed_status = 'mild'
elif DEMENTIA == 1 and MMSE_DISCRETIZED =='moderate':
   mixed_status = 'moderate'
elif DEMENTIA == 1 and MMSE_DISCRETIZED =='severe':
   mixed_status = 'severe'
elif PETERSEN MCI == 1 :
   mixed_status = 'mci_value'
else
    mixed_status = 'no_value'
```

Distribution Mixed_status during first episode				
mild	moderate	severe	mci_value	no_value
44	12	21	69	867
Distribution Mixed_status during last episode				
56	34	66	76	781

Table 16 : Comparison mix status between first and last episode

Table below shows the distribution of patients during the first episode and the last episode. We can observe that number of patients in each category is growing. The goal is to identify patients that change status and define if is correlative to some other features.

The following table show change status:

From MCI_value to Severe	14
From MCI_value to Moderate	6
From no_value to Mild	34
From no_value to MCI_value	57
From no_value to Severe	41
From no_value to Moderate	6
From MCI_value to Mild	8
From Mild to Moderate	9
From Mild to Severe	7

Table 17 : Overview of the mix status value from the first episode to the last

From MCI_value to Severe on Mixed Status

Analyzing this subcategory associated with 14 different patients that have an average of 13 episodes each one, we obtain the following results from which subsequent considerations for future studies can be made.

1. 72% of the patients have at least one genotype E4.
2. 75% of the patients have at least one of the comorbidities defined into Final Patient Data.
3. 92% of the patients are not a former smoker or current smoker.

From no_value to Severe on Mixed Status

Patients inside this subcategory are 41 with an average of 12 episodes each one.

1. 41% of the patients inside this subcategory are current smoker while the remaining part 59% are no smoker.
2. The following graph show distribution based on APOE type:

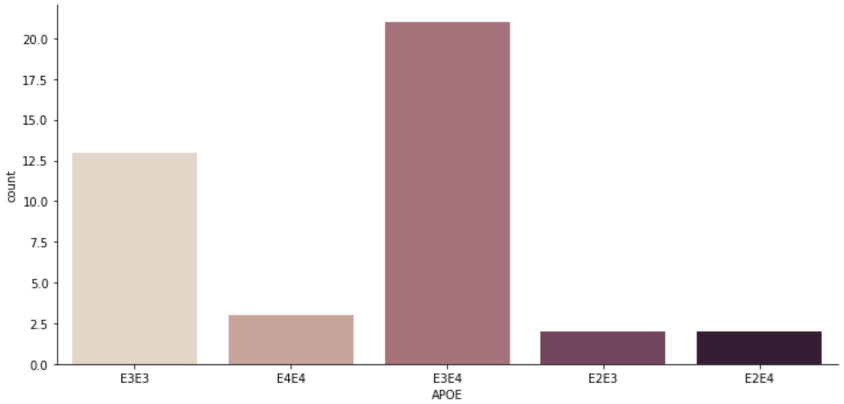


Figure 24: Overview of APOE for subcategory from no_ value to Severe

It is possible to note that, also in this subcategory, the percentage of patients who have at least one e4 genotype is high, equal to 63%.

From no_value to MCI_value on Mixed Status

Number of patients equal to 57 with an average of 7 episodes per patient.

1. 55% of patients in this subgroup suffered from malignant comorbidity.

2. 19 patients on 57 have no dementia flag equal to 1 during the first episode and have dementia(Alzheimer's disease) and other dementia flag equal to 1 during the last episode.

5.2.4 Patients that changes your MMSE status and with at least 5 episode

On Final Patient Data have been applied a filter considering only patient that changes their mini mental state examination score from the first to the last episode and furthermore all the patients that no have at least 5 episode have been removed .

```
list_patient_5 = final_list[4]['GLOBAL_PATIENT_DB_ID']
list_patient_5 = list_patient_5.tolist()
for i in range(len(final_list)) :
  for index,row in final_list[i].iterrows():
    if row['GLOBAL_PATIENT_DB_ID'] not in list_patient_5 :
      final_list[i] = final_list[i].drop(index)
```

With filter just shown, the number of total patients has fallen by 284 units, becoming equal to 724 and number of total rows becomes equal to 7276.

```
patient_status_not_change = []
patient_status_change = []
for index,row in final_list[0].iterrows():
   if row[’MMSE_DISCRETIZED_INIT’] == ’normal’ and
      row[’MMSE_DISCRETIZED_FINAL’] == ’normal’ :
     patient_status_not_change.append(row[’GLOBAL_PATIENT_DB_ID’])
   else :
     patient_status_change.append(row[’GLOBAL_PATIENT_DB_ID’])
for i in range(len(final_list)):
   for index,row in final_list[i].iterrows():
     if row[’GLOBAL_PATIENT_DB_ID’] not in patient_status_change :
        final_list[i] = final_list[i].drop(index)
```

Further, the application of this new filter above reduces the number of patients in Final Patient Data: 449 patients have changed status from the first to the last episode and the total lines are 4935.

5.3 KPI 2 - Correlation Analysis on Patient Final Data

In this section graphical representations have been carried out to analyze the correlations between Mini Mental State Examination and EHR patients, Mini Mental State Examination and general information. The tools used are Stacked bar plots and heatmap which give the possibility to obtain a complete view of the possible correlations. To compute correlation matrix Pearson’s correlation has been used.

Correlation analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables. A high correlation means that variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related [25].

For data scientists, checking correlations is an important part of the exploratory data analysis process. This analysis is used to decide which features affect the target variable the most. To do that, heatmap is used because visualization is generally easier to understand than reading tabular data, heatmaps are typically used to visualize correlation matrices. All the proposed analyzes were carried out on Final Patient Data after applying of **KPI 1**.

5.3.1 Correlation Analysis Between MMSE and general information

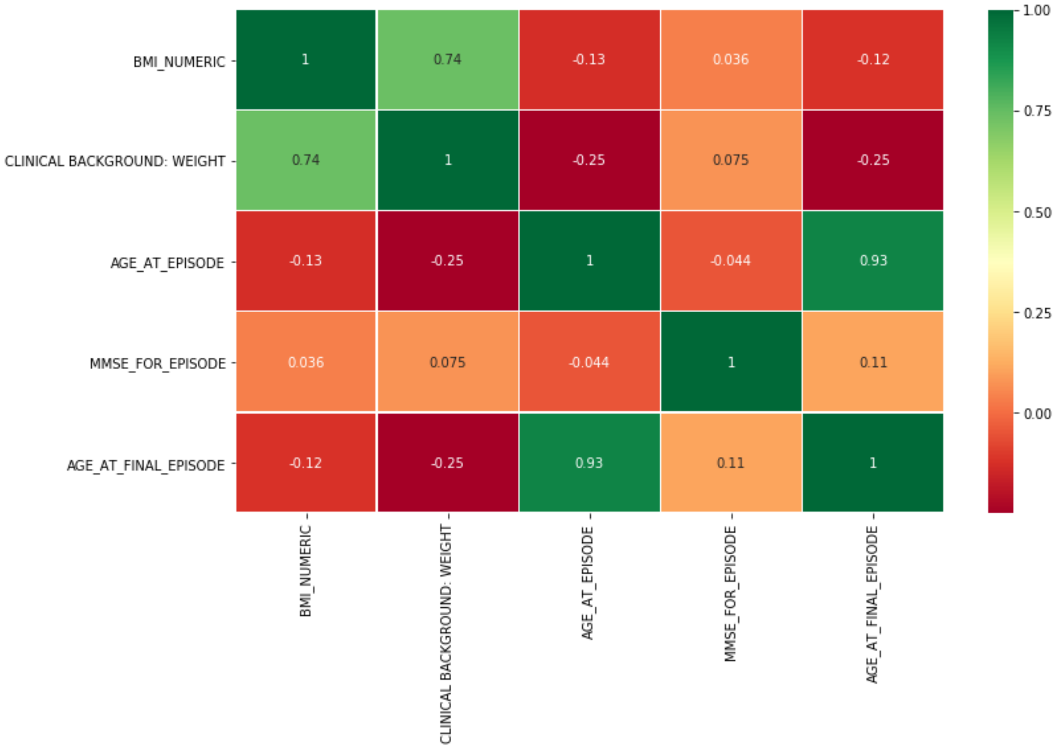


Figure 25: HeatMap MMSE and general information

Into the above figure, have been showed correlation between Mini Mental State Examination and general information(bmi, age, weight).
Looking HeatMap graph we can notice that there are no good positive or negative correlation between mmse and these feautures. Obviously the only strong correlations present are between features directly related to each other, such as body mass index and weight.

5.3.2 Correlation Analysis Between MMSE and comorbidity information

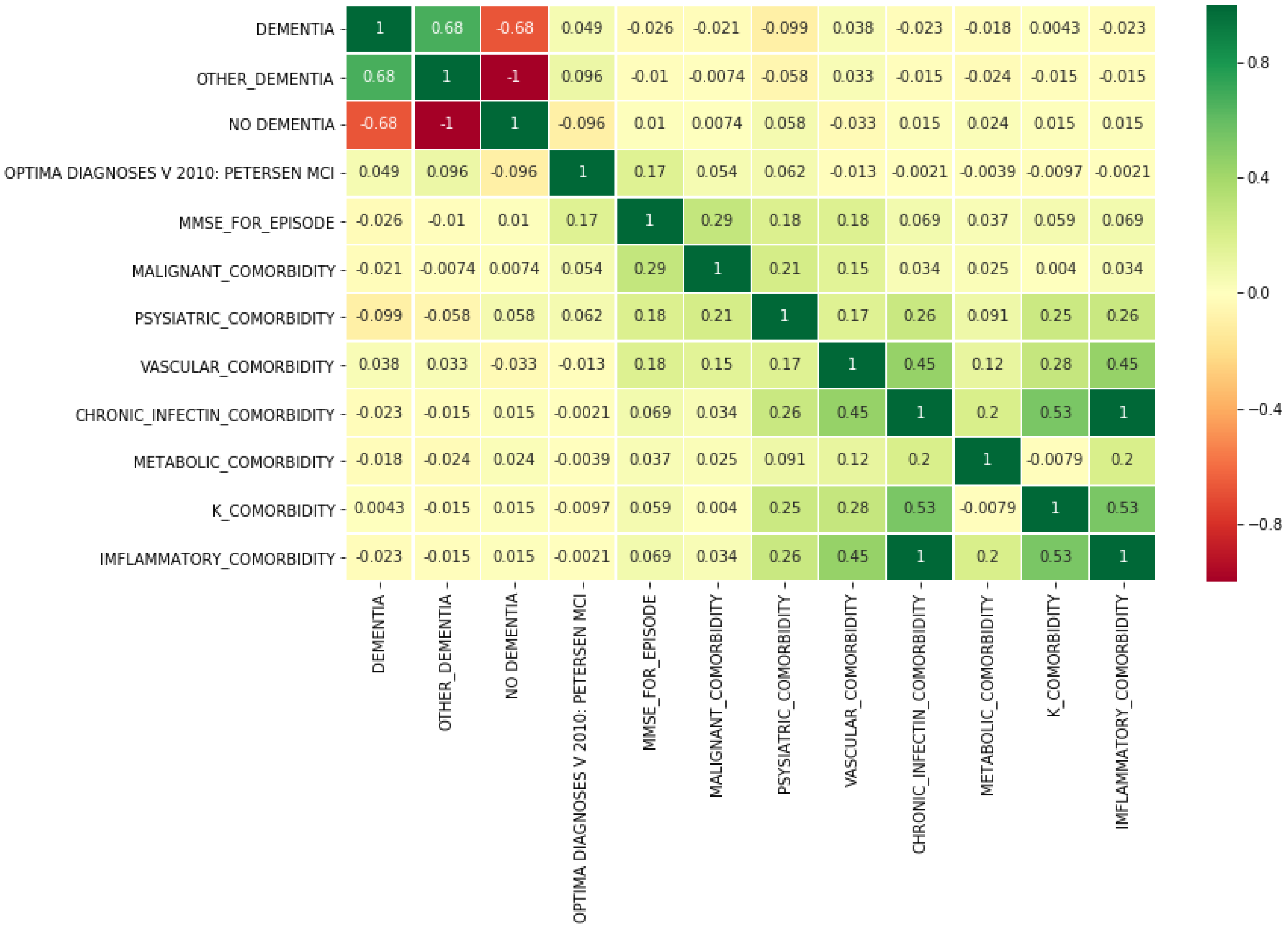
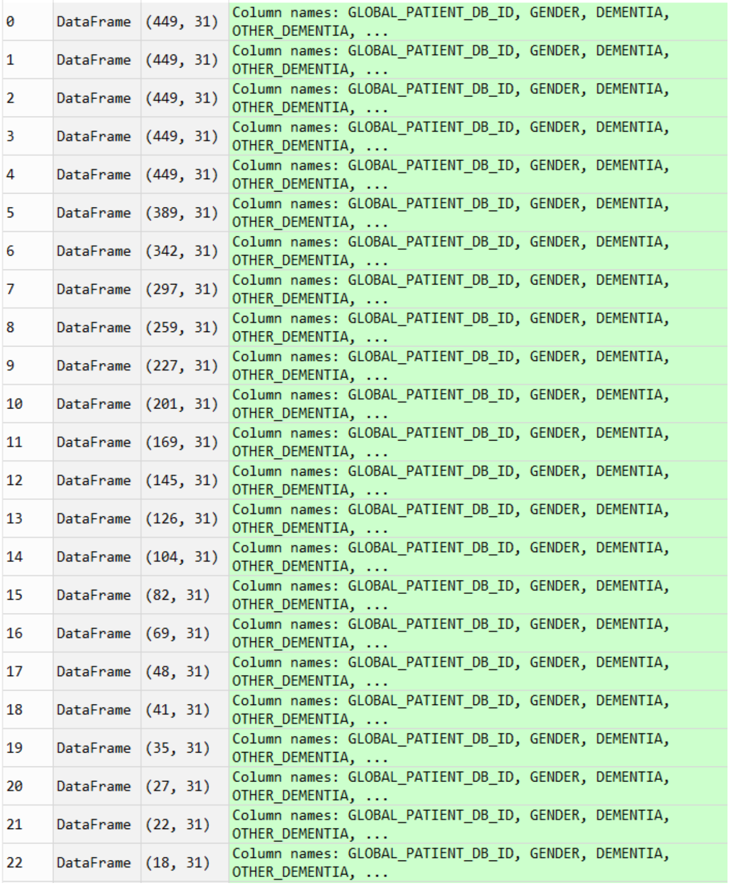


Figure 26: HeatMap MMSE and comorbidity information

Figure 26, show correlation between comorbidity information and Mini Mental State Examination. The two previous analyzes show us that using the pearson coefficient in search of strong correlations for our case study is totally useless.

5.3.3 Multicorrelation Analysis episode by episode

Multi correlation analysis is the most helpful tools to define a sensible study of this KPI. It has been used to show the progression of some features episode by episode, after applying the filters mentioned in section 5.2.2. The following schema shows how much patients have values for a episode. For example, the first five episodes have data for 449 different patients whereas the sixth episode have data for 389 different patient.



0	DataFrame	(449, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
1	DataFrame	(449, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
2	DataFrame	(449, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
3	DataFrame	(449, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
4	DataFrame	(449, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
5	DataFrame	(389, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
6	DataFrame	(342, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
7	DataFrame	(297, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
8	DataFrame	(259, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
9	DataFrame	(227, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
10	DataFrame	(201, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
11	DataFrame	(169, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
12	DataFrame	(145, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
13	DataFrame	(126, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
14	DataFrame	(104, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
15	DataFrame	(82, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
16	DataFrame	(69, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
17	DataFrame	(48, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
18	DataFrame	(41, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
19	DataFrame	(35, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
20	DataFrame	(27, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
21	DataFrame	(22, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...
22	DataFrame	(18, 31)	Column names: GLOBAL_PATIENT_DB_ID, GENDER, DEMENTIA, OTHER_DEMENTIA, ...

Figure 27: Overview Number of Patient for each episode

After the fourteenth episode it possible to notice that number of patients decreases considerably. Then, for the following stacked bar have been considering only the first 15 episodes. In the next bar graphs, achievement progression of MMSE and some comorbidities types are being compared. The left (vertical) axis represents patients in percentage, and the right (horizontal) axis represents achievement episodes. About the following stacked bar graphs we have to point out that number of patients isn't costant from episode 1 to epiosde 15, as mentioned in figure 28. To obtain a more clear understanding of the following stack bar graphs is presented for each of them a table showing:

- number of patients for each patient.
- the distribution of MMSE episode by episode.
- the distriubtion of feature under analysis episode by episode.

5.3.3.1 Mini Mental Score Eximation and Malignant Comorbidity

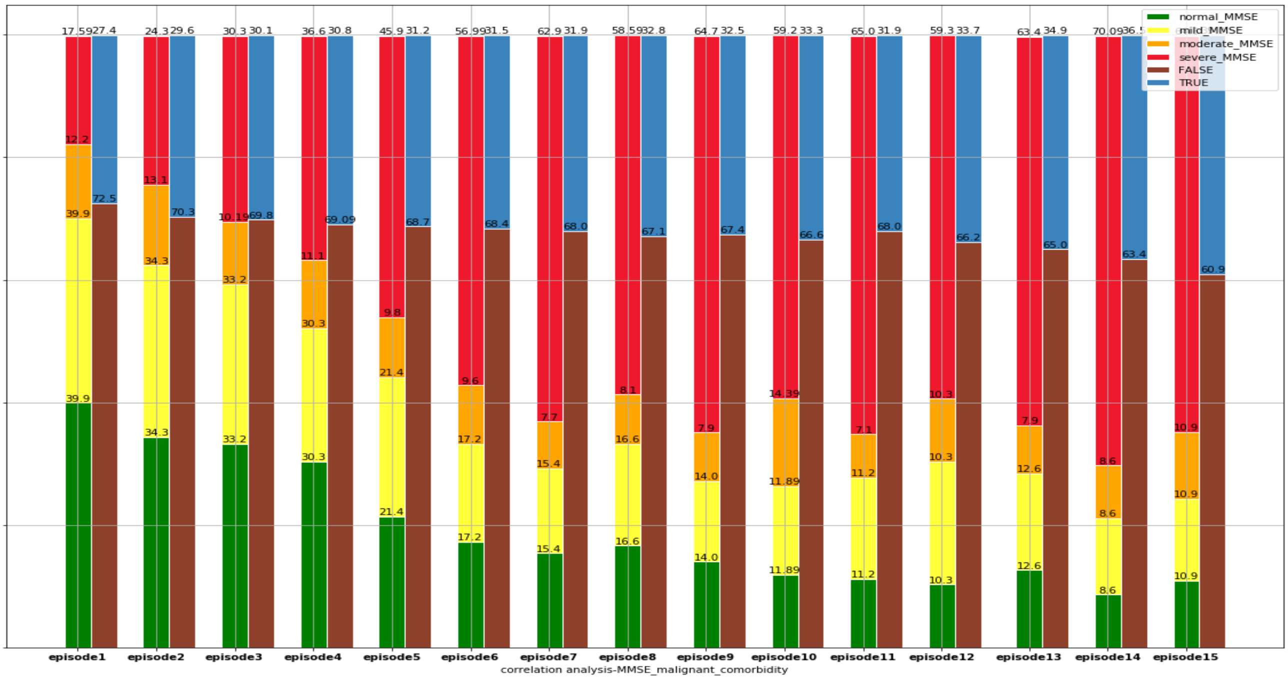


Figure 28: Stack Bar MMSE and Malignant Comorbidity

As explained in the previous section, patients are lost from episode to episode because the number of the episode that has been recorded is no unique for all the patients. About the first five episodes where the number of patients under analysis is homogeneous, the following line chart with normalized data has been defined to obtain more clear representation.

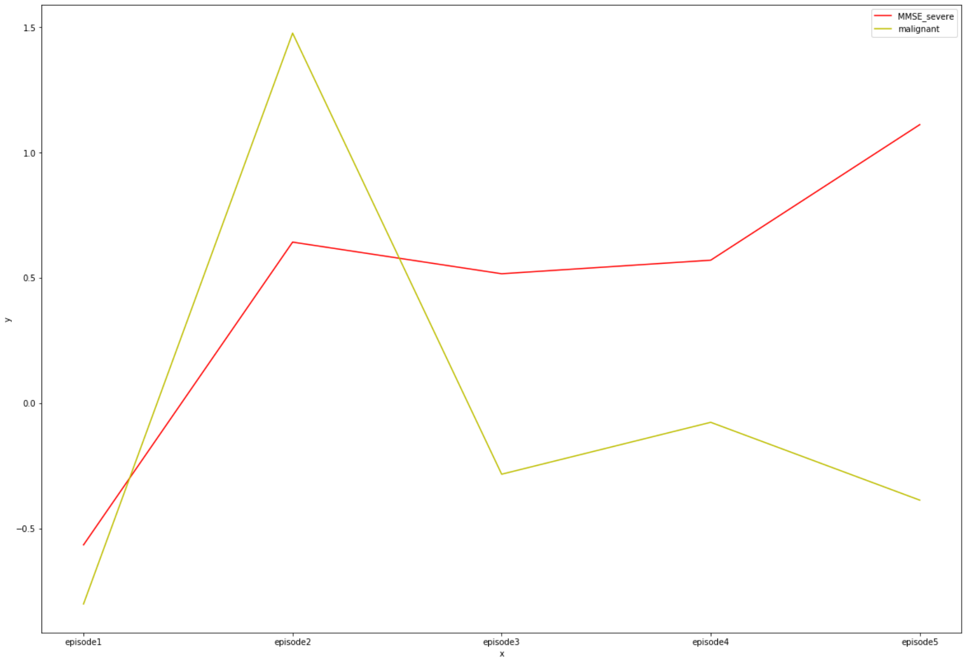


Figure 29: Progression mmse and malignant comorobidity for the first 5 episodes

We can see that as people with severe cognitive impairment increase, the number of people with malignant comorbidity increases but is no enough to say that are related. For the other episodes has been used the following table to exhibit some statistics and analysis

Malignant Comorbidity Evolution		
Number of Patient	Distribution Data	Episode
389	malignant comorbidity: true=108 false =234 mmse: severe=195 mod=33 mild=55 normal=59	episode6
342	malignant comorbidity: true=95 false =202 mmse: severe=187 mod=46 mild=41 normal=23	episode7
297	malignant comorbidity: true=85 false=174 mmse: severe=152 mod=43 mild=43 normal=21	episode8
259	malignant comorbidity: true=74 false =153 mmse: severe=147 mod=32 mild=30 normal=18	episode9
227	malignant comorbidity: true=67 false =134 mmse: severe=119 mod=29 mild=29 normal=24	episode10

Table 18 : Stack bar malignant-mmse statistics 1/2

Malignant Comorbidity Evolution		
Number of Patient	Distribution Data	Episode
201	malignant comorbidity: true=54 false=115 mmse: severe=110 mod=28 mild=19 normal=12	episode11
169	malignant comorbidity: true=96 false=49 mmse: severe=86 mod=29 mild=15 normal=15	episode12
145	malignant comorbidity: true=82 false=44 mmse: severe=80 mod=10 mild=20 normal=16	episode13
126	malignant comorbidity: true=38 false=66 mmse: severe=73 mod=9 mild=13 normal=9	episode14
104	malignant comorbidity: true=32 false=50 mmse: severe=53 mod=9 mild=11 normal=9	episode15

Table 19 : Stack bar malignant-mmse statistics 2/2

5.3.3.2 Mini Mental Score Eximation and Vascular Comorbidity

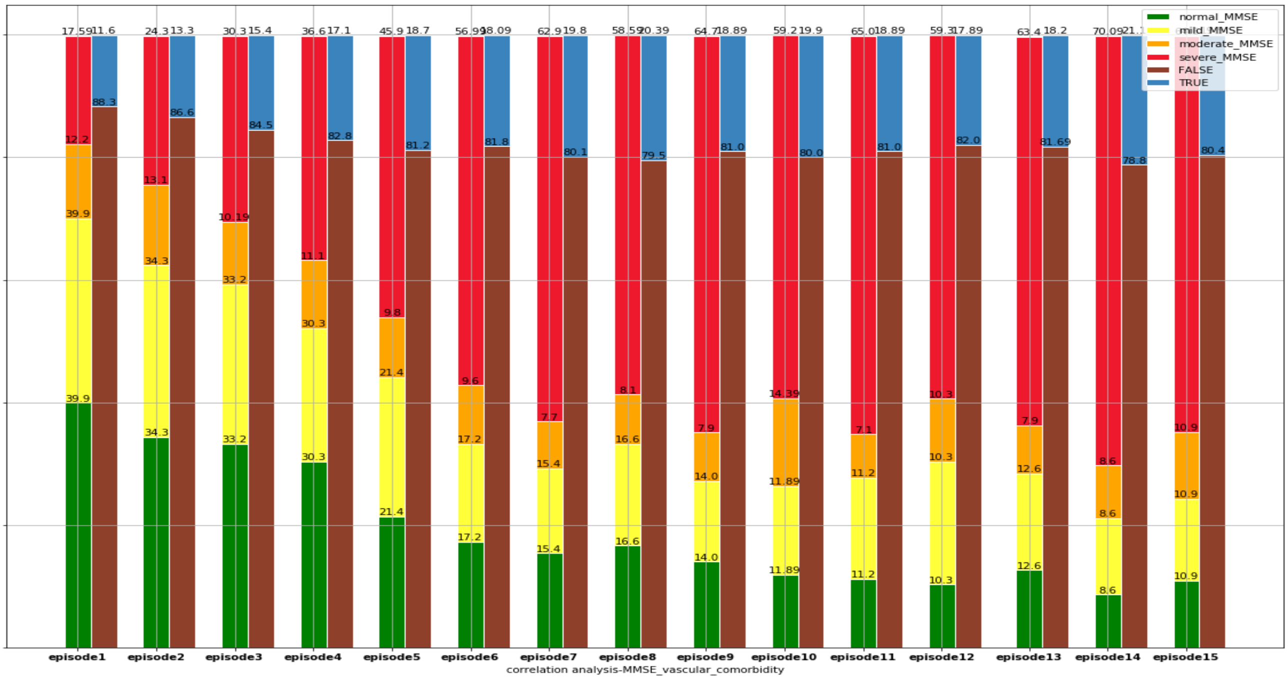


Figure 30: Stack Bar MMSE and Vascular Comorbidity

Figure 31 showing the progression of vascular comordibity and mini mental score examination where we can notice that during the first five episode progression is not proportional between them.

Vascular Comorbidity Evolution		
Number of Patient	Distribution Data	Episode
389	vascular comorbidity: true=62 false=280 mmse: severe=195 mod=33 mild=55 normal=59	episode6
342	vascular comorbidity: true=59 false=238 mmse: severe=187 mod=46 mild=41 normal=23	episode7
297	vascular comorbidity: true=53 false=20 mmse: severe=152 mod=43 mild=43 normal=21	episode8
259	vascular comorbidity: true=43 false=184 mmse: severe=147 mod=32 mild=30 normal=18	episode9
227	vascular comorbidity: true=40 false=161 mmse: severe=119 mod=29 mild=29 normal=24	episode10

Table 20 : Stack bar vascular-mmse statistics 1/2

Vascular Comorbidity Evolution		
Number of Patient	Distribution Data	Episode
201	vascular comorbidity: true=32 false=137 mmse: severe=110 mod=28 mild=19 normal=12	episode11
169	vascular comorbidity: true=119 false=26 mmse: severe=86 mod=29 mild=15 normal=15	episode12
145	vascular comorbidity: true=23 false=103 mmse: severe=80 mod=10 mild=20 normal=16	episode13
126	vascular comorbidity: true=22 false=82 mmse: severe=73 mod=9 mild=13 normal=9	episode14
104	vascular comorbidity: true=16 false=66 mmse: severe=53 mod=9 mild=11 normal=9	episode15

Table 21 : Stack bar vascular-mmse statistics 2/2

5.3.3.3 Mini Mental Score Eximation and Sistemic Comorbidity

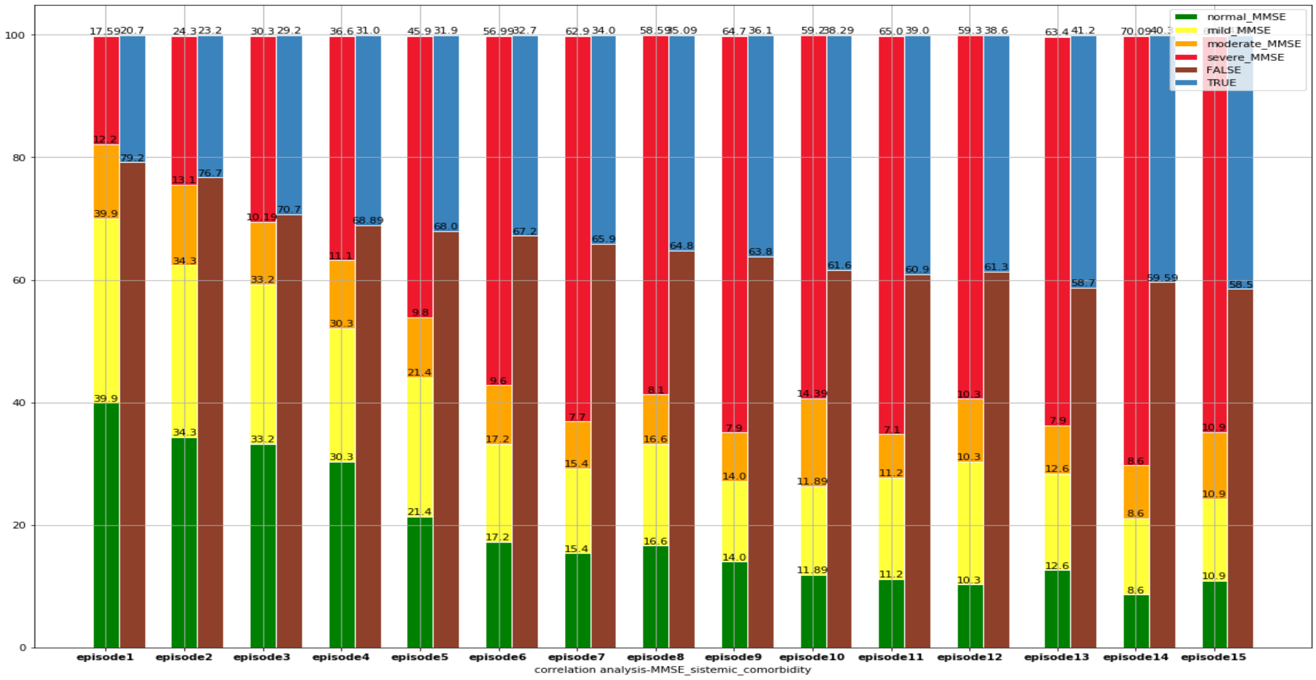


Figure 31: Stack Bar MMSE and sistemic comorbidity

Sistemic Comorbidity Evolution		
Number of Patient	Distribution Data	Episode
389	sistemic comorbidity: true=112 false=230 mmse: severe=195 mod=33 mild=55 normal=59	episode6
342	sistemic comorbidity: true=101 false=196 mmse: severe=187 mod=46 mild=41 normal=23	episode7
297	sistemic comorbidity: true=91 false=168 mmse: severe=152 mod=43 mild=43 normal=21	episode8
259	sistemic comorbidity: true=82 false=145 mmse: severe=147 mod=32 mild=30 normal=18	episode9
227	sistemic comorbidity: true=77 false=124 mmse: severe=119 mod=29 mild=29 normal=24	episode10

Table 22 : Stack bar sistemic-mmse statistics 1/2

Sistemic Comorbidity Evolution		
Number of Patient	Distribution Data	Episode
201	vascular sistemic: true=66 false=103 mmse: severe=110 mod=28 mild=19 normal=12	episode11
169	sistemic comorbidity: true=56 false=89 mmse: severe=86 mod=29 mild=15 normal=15	episode12
145	sistemic comorbidity: true=74 false=52 mmse: severe= mod= mild= normal=	episode13
126	sistemic comorbidity: true=42 false=62 mmse: severe=73 mod=9 mild=13 normal=9	episode14
104	sistemic comorbidity: true=34 false =48 mmse: severe=53 mod=9 mild=11 normal=9	episode15

Table 23 : Stack bar sistemic-mmse statistics 2/2

5.4 KPI 3 - Rate of progression after first diagnosis

As mentioned in section 4.3, survival analysis is a "time to event" technique. The general idea is to study the progress of a dataset in relation to time and an event of interest that is seen as our target features. In my case the event of interest is a variable that will be created for the purpose in this section.
Into the following step have been explain how are identify event of interest that is called several status and time.

```
for i in range(len(patient_group)):
   for index,row in patient_group[i].iterrows():
      if row['MMSE_DISCRETIZED_INIT'] == 'severe' :
         patient_that_became_severe.append(row['GLOBAL_PATIENT_DB_ID'])
         date_initial.append(row['EPISODE_DATE'])
         date_when_became_severe.append(row['EPISODE_DATE_FINAL'])
         several_status.append(True)
```

code below show how has been computed several status. The objective is to locate date when the patient becomes severe, if there is, and then create a variable called severe status that is 1 if the patient becomes severe otherwise 0.
Furthermore, if a severe status has been found the respective date is saved otherwise if a severe status is not identified for that patient, the last episode is taken as the date.

```
    psysiatric_type = psysiatric_type[psysiatric_type['TYPE']=='N']
   for n in range(len(psysiatric_type)) :
      if another_patient_with_psysiatric_problem[n] not in
          list_patient_with_psyatric_problem:
```

To the temporal division, it has been decided to evaluate the difference between the first episode and the event where the patient becomes severe. The difference between the two dates is calculated as weeks rooting.
Set these two variables a typical nonparametric survival analysis tools was used : Kaplan-Meier Estimator.

Kaplan-Meier Estimator: Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. The time starting from a defined point to the occurrence of a given event, for example death is called as survival time and the analysis of group data as survival analysis. This can be affected by subjects under study that are uncooperative and refused to be remained in the study or when some of the subjects may not experience the event or death before the end of the study, although they would have experienced or died if observation continued, or we lose touch with them midway in the study. We label these situations as censored observations. The Kaplan-Meier estimate is the simplest way of computing the survival over time in spite of all these difficulties associated with subjects or situations. The survival curve can be created assuming various situations. It involves computing of probabilities of occurrence of event at a certain point of time and multiplying these successive probabilities by any earlier computed probabilities to get the final estimate. This can be calculated for two groups of subjects and also their statistical difference in the survivals[26].

All the graphs associated with the Kaplan-Meier Estimator are composed of three subgraphs: the first describes the Kaplan-Meier curve where on the x-axis we have the time intervals and on the y-axis we have the probability that a patient will become severe and into the graph is present p_value.

In statistics, the p-value is the probability of obtaining results as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis[27].

The second graph divides the patients into sub-groups in relation to the feature used and analyzes the trend throughout the time interval, documenting with a percentage the number of patients who did not register the event or who were censored compared it to the initial total. Finally, the third graph shows the number of censured for each subgroup during the whole time interval considered.

5.4.1 APOE survival analysis

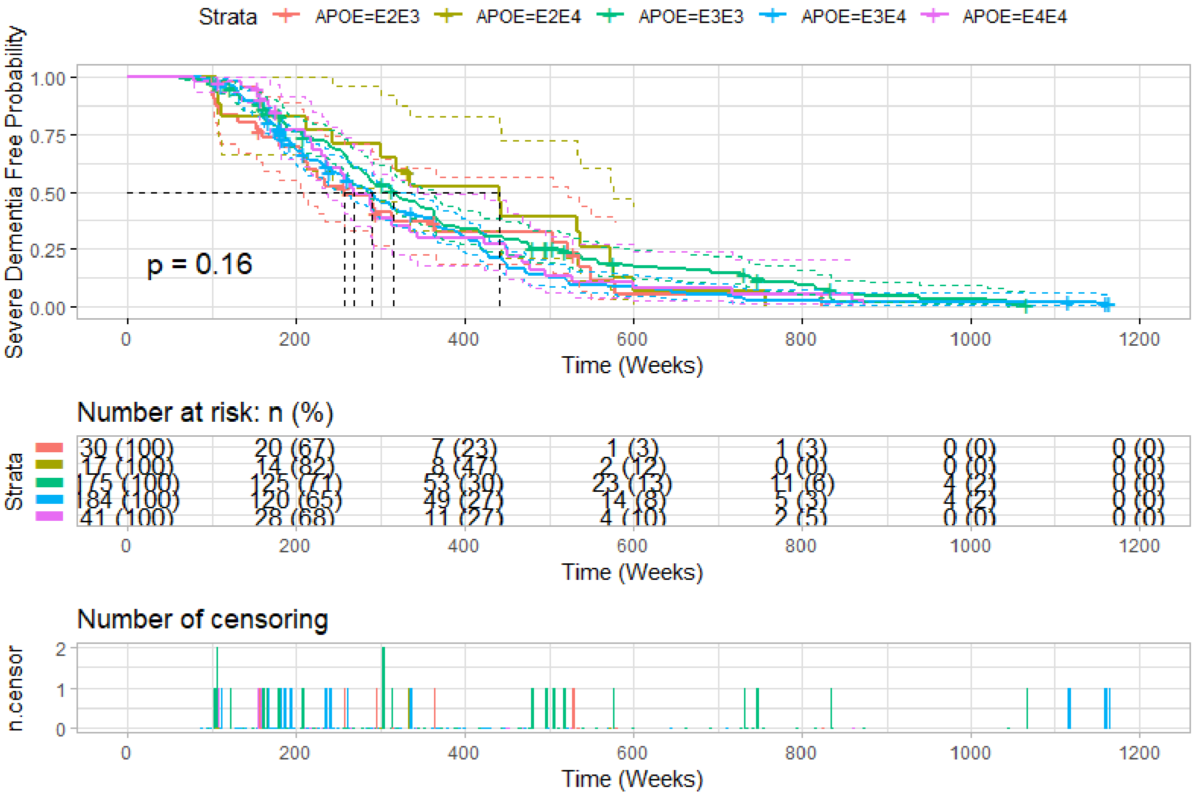


Figure 32: APOE survival analysis

Analyzing the previous graph we see that the patients who have APOE e2e4 are those who have 50% likelihood of becoming severe after 440 weeks on average (8.5 years). We also note that patients who have APOE e2e3 and e4e4 are those who after only 260 weeks (5 years) go to severe status. We also note that patients who become severe or who have been censored between 900 and 1500 weeks belong to APOE e3e4 or e3e3.

In the following chart we will consider a survival analysis associated only with patients who have Alzheimer's disease.

The graph presented confirms the studies mentioned in section 4.4.4. individuals with at

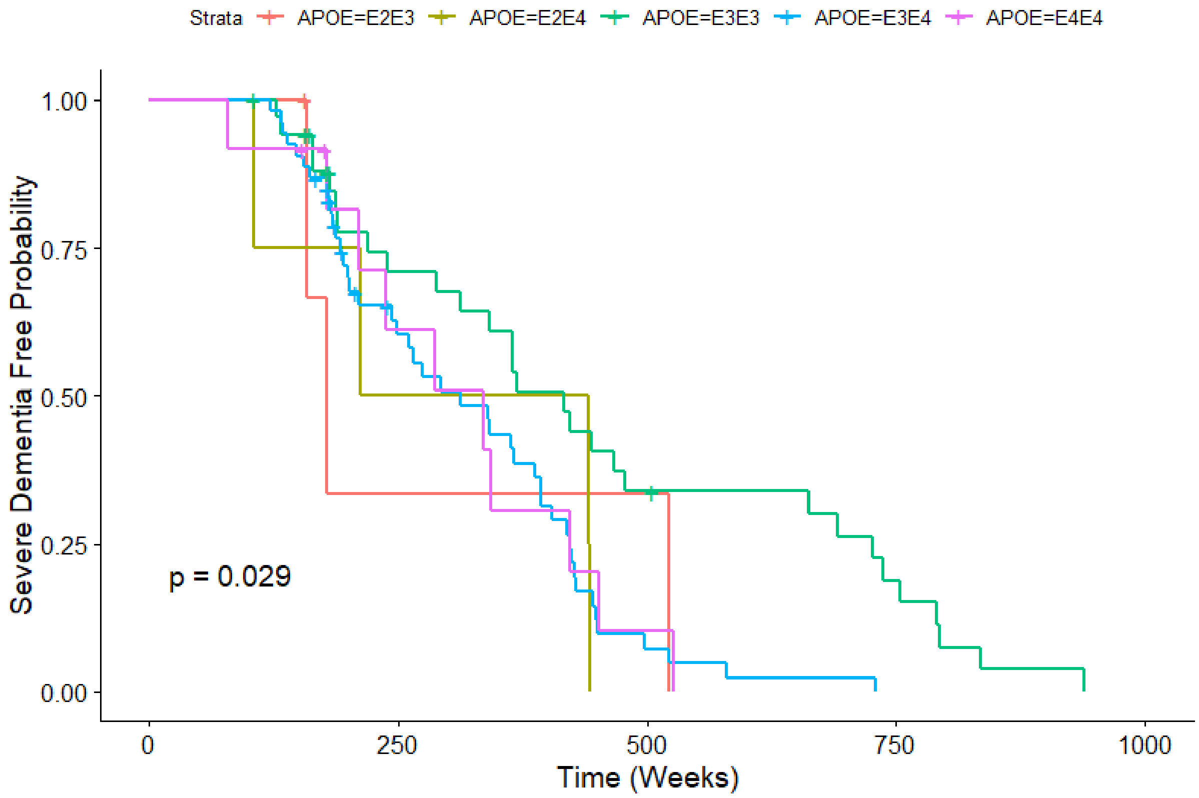
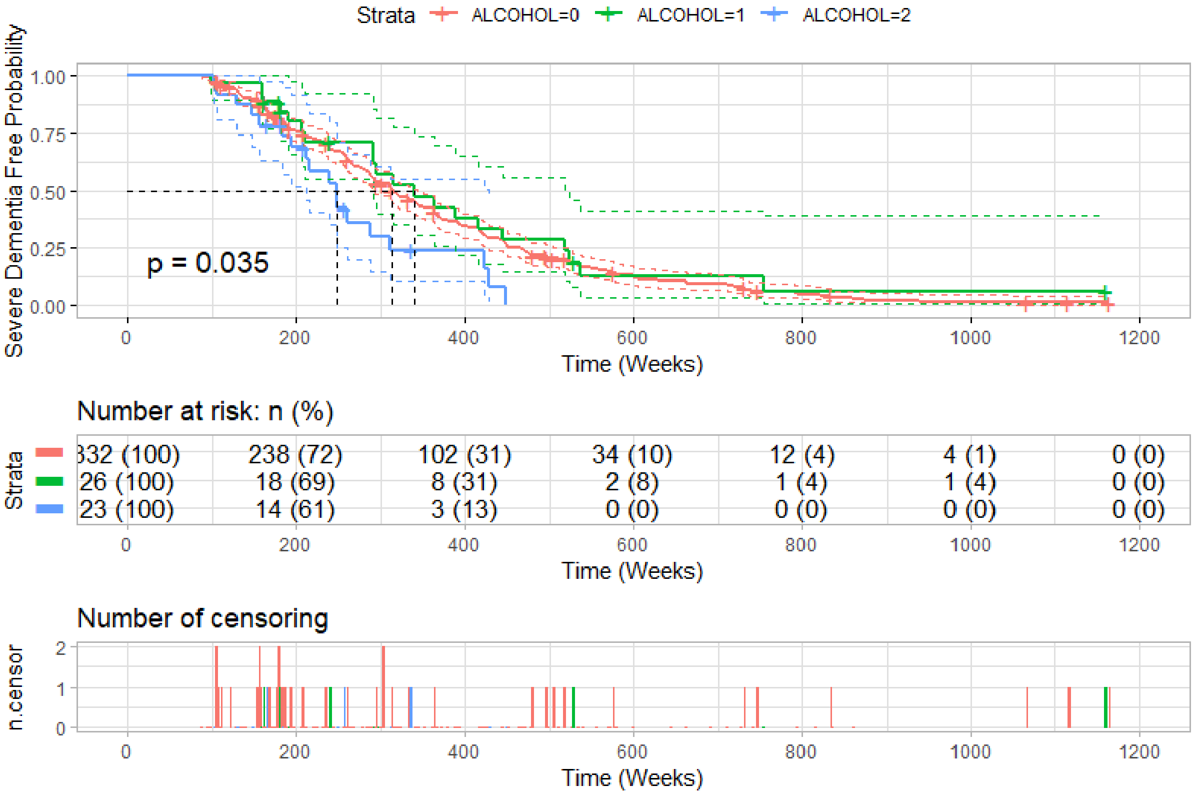


Figure 33: APOE survival analysis on patient with Alzheimer's disease

least one E4 genotype have a higher risk of going to a severe status than other patients. From this graph, the patients who go to severe status as late as possible are those with the E3E3 genotype

5.4.2 ALCOHOL survival analysis



Strata	0	200	400	600	800	1000	1200
🟥	332 (100)	238 (72)	102 (31)	34 (10)	12 (4)	4 (1)	0 (0)
🟩	26 (100)	18 (69)	8 (31)	2 (8)	1 (4)	1 (4)	0 (0)
🟦	23 (100)	14 (61)	3 (13)	0 (0)	0 (0)	0 (0)	0 (0)

Figure 34: ALCOHOL survival analysis

Figure 41 show survival analysis for patients into Final Patient Data associated to alcohol consumption.

In particolare :

- ALCOHOL = 0 represent patients that are no drinking
- ALCOHOL = 1 represent patients that are mild drinking
- ALCOHOL = 2 represent patients that are extreme drinking

The graph clearly shows that patients who have a history of extreme drinking have a significantly better likelihood of moving to severe status in a short time. In fact, the graph shows that the probability of becoming severe is 50 % already after 250 weeks (4.7 years) and 25 % after 300 weeks (5.7 years).

5.4.3 VASCULAR COMORBIDITY survival analysis

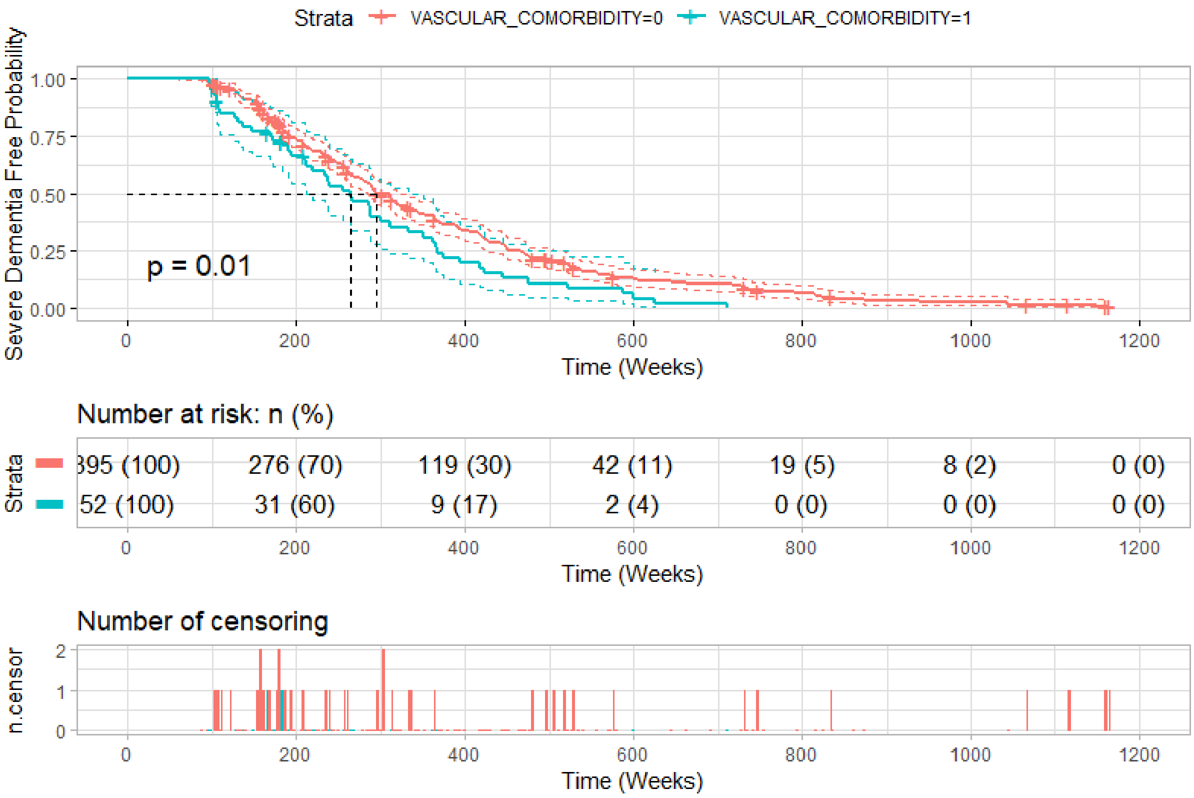


Figure 35: VASCULAR comorbidity survival analysis

In the graph above, VASCULAR_COMORBIDITY = 0 indicates patients who do not have that disease while VASCULAR_COMORBIDITY = 1 indicates patients who present it. It is noted that in relation to this comorbidity there is no big difference between the two categories of patients.

This maker can also be seen in the percentage of number at risk where the ratio remains unchanged from one interval to another.

5.4.4 IMFLAMMATORY COMORBIDITY survival analysis

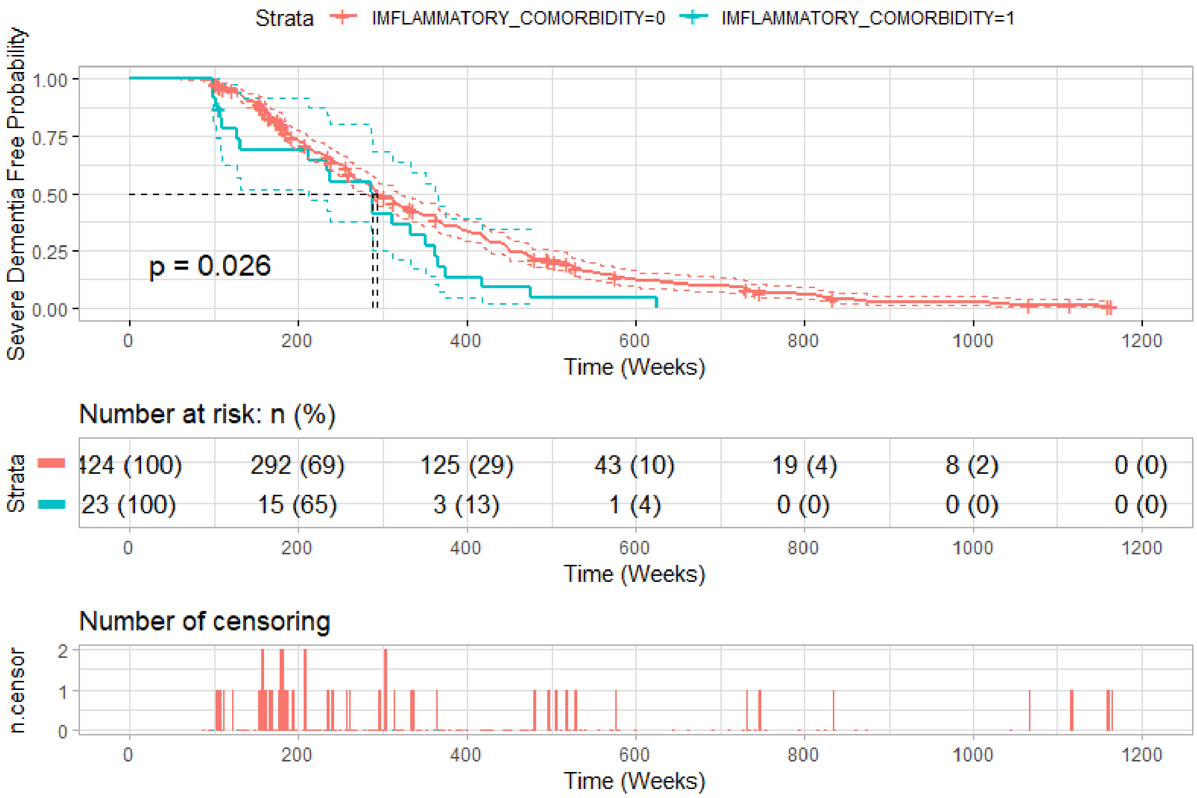


Figure 36: IMFLAMMATORY comorbidity survival analysis

The graph above shows us that 5 % of patients have Imflammatory comorbidity. We also note that the probability of becoming severe is 50 % after 290 weeks while it drops sharply to 25 % after 350 weeks.

In relation to the data present in Final Patient Data we can say that the likelihood of passing to severe status is greater for patients who have inflammatory comorbidity.

6 Conclusions and Perspectives

This master thesis work makes it possible to transform raw data into a structured form that can be used to extract knowledge that can help to improve Key Performance Indicators and a set of methods and techniques have been applied in order to discover pattern in a medical data. To archieve this goals a main dataset called Patients Final Data has been developed containing all the information needed.
The main challenge that have been faced can be splitted into:

1. **preparation stage** - Noise removal in medical data, filtering section by type, understanding medical data to define which are suitable data for the proposal purposes, classification of patients into a fixed set of subgroups in order to enable more precise analysis.

2. **business understanding** - Translation of business requirement on a data mining terms, the definition of which are columns that the clinical want on Final Patient Data, identify KPIs and the relative metrics.

3. **data extraction** - the way to use raw data to obtain final features, how raw data have been combined to obtain features and an overview about the algorithm used to compute comorbidity features, drinking feature, education feature, smoking feature, APOE feature and dementia type features.

4. **several approach to extract new information** - Development of the Key Performance Indicators defined and identification of a new points of view usable in future studies.

To developed challange proposed several tools have been used : histogram, bar chart, stacked bar chart, heatmap, survival analysis techniques, Kaplan-Meir Estimator.
The analysis of Patients Final Data corroborates some knowledge presenting on several studies about alzheimer's and dementia disease but, at the same time, develops new prospectives of analysis.
The project shows us that the likelihood to develop a dementia disease with severe status is more common for patients that have at least one e4 allele. Patients having e3e3 APOE develop the disease slowly and they often don't pass to the severe status.

Furthermore, the project shows us that smoking is not to be considered as a possible cause of disease while it is possible to say that strong drinking problem have a good relationship with the likelihood to become severe and then to developed dementia disease. Finally, no strong correlation was identified between one of the comorbidity analyzed in this project and the result provided by the cognitive tests MMSE. These results obtained aren't consistent because raw data used to build Final Patient Data comorbidities features have a lot of missing or unknown data. To have a more clear understanding of these relationships we could increase the number of samples and data robustness using the final Patiemt data structure as a start point. We can consider this as a perspective for future studies because this project is the first part of a more ambitious project that wants to identify causes and possible solutions to dementia disease using machine learning techniques. In conclusion, for each KPIs, the best methods and techniques have been applied to obtain it despite result of the clinical point of view has no been validated.

Bibliography

[1] “Types of dementia.” https://qbi.uq.edu.au/brain/dementia/types-dementia.

[2] “Improving your odds with data science hiring.” https://www.datanami.com/2018/09/17/improving-your-odds-with-data-science-hiring/.

[3] “What is the crisp-dm methodology?.” https://www.sv-europe.com/crisp-dm-methodology/.

[4] “data.” https://en.wikipedia.org/wiki/Data.

[5] “The digitization of the worldfrom edge to core - available at https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf,” 2018.

[6] “Oxford project to investigate memory and ageing (optima).” https://www.ndcn.ox.ac.uk/research/centre-prevention-stroke-dementia/resources/optima-oxford-project-to-investigate-memory-and-ageing.

[7] W. H. Organization *et al.*, “Risk reduction of cognitive decline and dementia: Who guidelines,” in *Risk reduction of cognitive decline and dementia: WHO guidelines*, pp. 401–401, 2019.

[8] E. S. Sharp and M. Gatz, “The relationship between education and dementia an updated systematic review,” *Alzheimer disease and associated disorders*, vol. 25, no. 4, p. 289, 2011.

[9] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, “Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy,” *Nature Reviews Neurology*, vol. 9, no. 2, p. 106, 2013.

[10] T.-B. Chen, S.-Y. Yiao, Y. Sun, H.-J. Lee, S.-C. Yang, M.-J. Chiu, T.-F. Chen, K.-N. Lin, L.-Y. Tang, C.-C. Lin, *et al.*, “Comorbidity and dementia: a nationwide survey in taiwan,” *PLoS One*, vol. 12, no. 4, 2017.

[11] L. Gustafson, “What is dementia?,” *Acta Neurologica Scandinavica*, vol. 94, pp. 22–24, 1996.

[12] D. G. Munoz and H. Feldman, “Causes of alzheimer’s disease,” *Cmaj*, vol. 162, no. 1, pp. 65–72, 2000.

[13] C. Hölscher, “Possible causes of alzheimer’s disease: amyloid fragments, free radicals, and calcium homeostasis,” *Neurobiology of disease*, vol. 5, no. 3, pp. 129–141, 1998.

[14] “Oxford project to investigate memory and ageing (optima).” https://www.ndcn.ox.ac.uk/research/centre-prevention-stroke-dementia/resources/optima-oxford-project-to-investigate-memory-and-ageing.

[15] “Kpi basics - available at https://kpi.org/KPI-Basics,”

[16] “Kpi characteristics - available at https://entrinsik.com/5-characteristics-effective-kpis/,”

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[18] “What is data preparation?.” https://www.talend.com/resources/what-is-data-preparation/.

[19] “Python.” https://it.wikipedia.org/wiki/Python.

[20] G. G. Koch, “A basic demonstration of the [-1, 1] range for the correlation coefficient,” *The American Statistician*, vol. 39, no. 3, pp. 201–202, 1985.

[21] “Survival analysis — part a.” https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e.

[22] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, *et al.*, “Mild cognitive impairment,” *The lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[23] “Apoe gene.” https://ghr.nlm.nih.gov/gene/APOE.

[24] M. Roser and H. Ritchie, “Cancer,” *Our World in Data*, 2020. https://ourworldindata.org/cancer.

[25] “Correlation analysis.” https://www.sciencedirect.com/topics/medicine-and-dentistry/correlation-analysis.

[26] M. K. Goel, P. Khanna, and J. Kishore, “Understanding survival analysis: Kaplan-meier estimate,” *International journal of Ayurveda research*, vol. 1, no. 4, p. 274, 2010.

[27] “P-value definition.” https://www.investopedia.com/terms/p/p-value.asp.