POLITECNICO
MILANO 1863

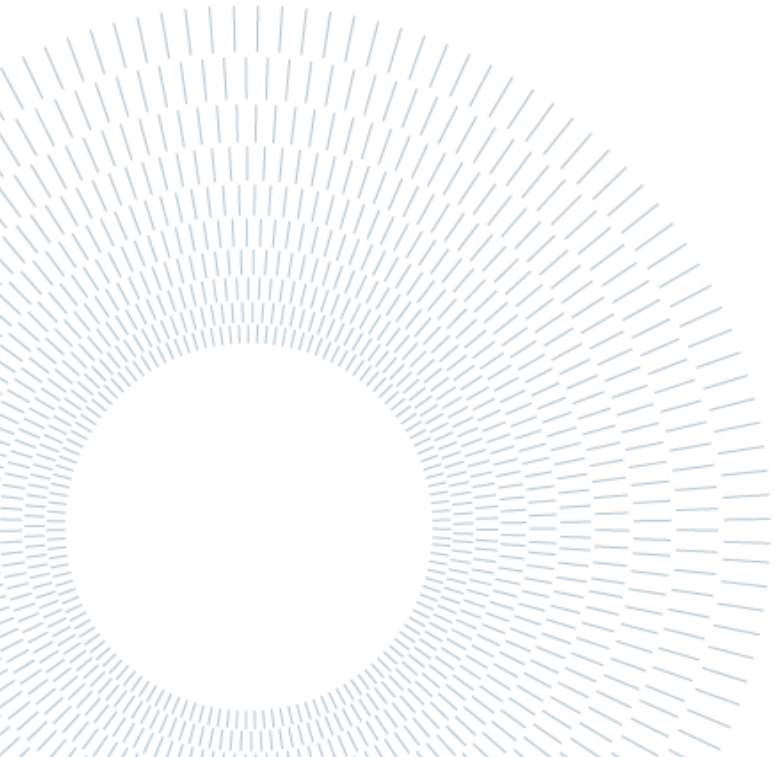
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Evaluation of statistical analysis methods and machine learning techniques to develop Personas for willingness of vaccination against COVID-19

TESI DI LAUREA MAGISTRALE IN
BIOMEDICAL ENGINEERING
INGEGNERIA BIOMEDICA

Author: **Francesca Onofri**

Student ID: 944174
Advisor: Enrico Gianluca Caiani
Co-advisor: Emanuele Tauro
Academic Year: 2021-22



Abstract

On January 30, 2020, the World Health Organization (WHO) declared a state of international public health emergency due to the Coronavirus epidemic in China. On 11 March 2020, the director general of the WHO defined the spread of Covid-19 as a pandemic throughout the planet. Scientists from all over the world immediately began studying the virus for a vaccine, and more than 150 vaccines were tested. At the end of December, the vaccination campaign officially began in Europe, initially intended for health workers and frail people, and then progressively extended to the entire population over 5 years of age. However, since vaccinations were not deemed mandatory in Italy, it was necessary to identify those that were resistant against getting vaccinated due to various reasons. Thus, the creation of Personas to identify the main characteristics of people willing to be vaccinated was deemed of interest.

The goal of this thesis is to compare multiple clustering and supervised machine learning methods in order to create Personas that stand out for their willingness, or lack thereof, to vaccinate against Covid-19.

A survey, created by the University of Milan to investigate people's psychophysical conditions and their willingness to get vaccinated, was administered online, directed to the Italian population, in the months of January and February 2021. The data obtained from the survey were analysed through agglomerative and hierarchical clustering techniques. Further analyses were performed using supervised machine learning methods, creating classification models. Through statistical analysis, the differences within the datasets obtained by the previous methods were investigated, dividing them into significantly different clusters.

Three Persona tables were thus obtained, which were compared with each other to evaluate their effectiveness in highlighting common characteristics among individuals who would not have been vaccinated against Covid-19.

This thesis shows that supervised machine learning methods are comparable, in results, to statistical analysis methods and clustering techniques in identifying the main characteristics of Personas if a target variable is available.

Key-words: Covid-19, vaccination, Personas, clustering, classification.

Abstract in italiano

Il 30 Gennaio 2020 l'Organizzazione Mondiale della Sanità (OMS) ha dichiarato lo stato di emergenza internazionale di salute pubblica dovuta all'epidemia di Coronavirus in Cina. L'11 Marzo 2020 il direttore generale dell'OMS ha definito la diffusione del Covid-19 una pandemia in tutto il pianeta. A fine dicembre in Europa è iniziata ufficialmente la campagna vaccinale, inizialmente destinata ad operatori sanitari e persone fragili, per poi estendersi progressivamente a tutta la popolazione sopra i 5 anni. Tuttavia, poiché in Italia il vaccino non è stato reso obbligatorio, si è cercato di identificare i motivi che hanno spinto le persone a decidere di non vaccinarsi. Pertanto la creazione di Personas è stata considerata di interesse per identificare le caratteristiche principali delle persone disposte a vaccinarsi.

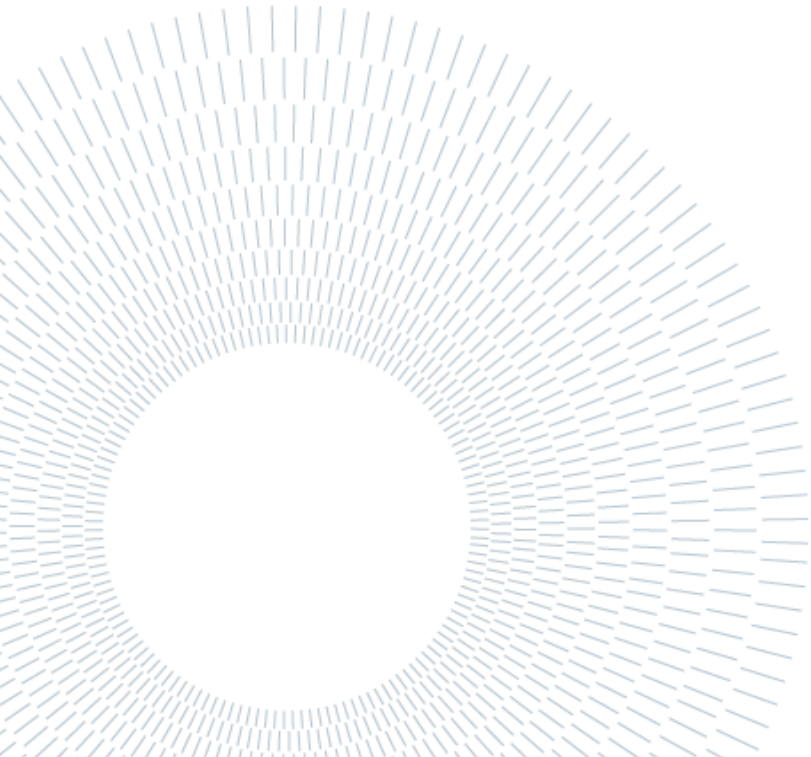
L'obiettivo di questa tesi è quello di confrontare più metodi di clustering e machine learning supervisionato al fine di creare Personas che si distinguano per la loro volontà di vaccinarsi contro il Covid-19.

Un questionario, appositamente creato dall'Università di Milano per indagare sulle condizioni psicofisiche delle persone e sulla loro volontà di vaccinarsi, è stato somministrato online, diretto alla popolazione italiana, nei mesi di Gennaio e Febbraio 2021. I dati ottenuti dal questionario sono stati analizzati tramite tecniche di clustering agglomerativo e gerarchico. Ulteriori analisi sono state eseguite tramite metodi di machine learning supervisionato, creando modelli di classificazione. Attraverso l'analisi statistica si sono andate ad approfondire le differenze all'interno dei dataset ottenuti dai precedenti metodi, suddividendoli in clusters significativamente differenti tra loro.

Si sono così ottenute tre clusterizzazioni, che sono state confrontate tra loro per valutarne l'efficacia nell'evidenziare caratteristiche comuni tra gli individui che non si sarebbero vaccinati.

Questa tesi mostra come i metodi di machine learning supervisionato sono comparabili, nei risultati, alle tecniche di clustering e analisi statistica nell'identificare le caratteristiche principali delle Personas, se è presente una variabile target.

Parole chiave: Covid-19, vaccini, Personas, clustering, classificazione.



Contents

Introduction

On December 31, 2019, the Wuhan Municipal Health Commission (China) reported to WHO a cluster of pneumonia cases of unknown aetiology in the city of Wuhan. On 30th January 2020, the World Health Organization (WHO) Emergency Committee declared a global health emergency based on growing case notification rates in locations all over the world [1]. On 11th March 2020 the WHO had officially declared Covid-19 a pandemic, based on its worldwide spread and extremely fast contagion rates.

Many scientific laboratories worldwide have worked on an unprecedented timeline since April 2020 [2], to create effective vaccines to control the spread of the pandemic virus [3].

China was the first to approve the emergency use of Covid-19 vaccine in June 2020, through a self-developed vaccine, targeting groups with high risks of infection. Russia approved for the first time the emergency use of its Sputnik V vaccine on August 11. In US, on 11 December 2020, the Food and Drug Administration (FDA) authorized the Pfizer-BioNTech vaccine, and the vaccinations started 4 days later. In UK, on 30 December 2020, the Medicines and Healthcare products Regulatory Agency (MHRA) gave the approval to Oxford/AstraZeneca vaccine. Australia launched vaccination program on 22 February 2021.

In European Union, European Medicines Agency (EMA) gave the ahutorization for the first Pfizer-BioNTech vaccine on 21 December 2020. On December 27th, 2020 this hard work finally resulted in the start of the vaccination campaign in Europe with the first vaccines approved from the European Medicines Agency (EMA) for people aged 16 and over, being deployed on December 31, 2020. From them on, four other vaccines were approved: Spikevax Moderna on 6 January 2021, Oxford/AstraZeneca on 29 January, Vaccine Janssen on 11 March 2021 and, finally, Nuvaxovid by Novax on 20 December, 2021 [4].

In Italy, as well as in the rest of the world, in the initial phase of limited availability of vaccines, it was necessary to define priorities, taking into account the international and european recommendations. As stated by IIS (Istituto Superiore di Sanità), three categories have been identified to be vaccinated, with the first dose, as a priority, beginning on December 31, 2020: health and socio-sanitary workers, residents and staff of residential care facility, the elderly over 80. The vaccination then has been extended

to all people over 16 years of age, leaving Italian regions to decide independently how to schedule them. From 28 May 2021 the vaccine was extended up to 12 years old people and from 25 November it was extended to the 5-11 year range. To date, the doses available for the entire vaccinable population are three, at regular intervals, while the fourth dose is addressed, for now, only to the fragile categories.

The herd immunity threshold is defined as the proportion of individuals in a population who, having acquired immunity, can no longer participate in the chain of transmission. Assuming no population immunity and the same probability of contracting and transmitting the virus, the herd immunity threshold for SARS-CoV-2 would stay around 70%. Therefore with an effective vaccination program, herd immunity can be sustained, and it must be efficient to cover those who cannot be directly protected [5]. For this reason, to obtain an extensive and effective vaccination, two of the most important elements are: the availability of the vaccine all over the world and the willingness to be vaccinated by all those who are allowed to.

Vaccine hesitancy (i.e. the delay in acceptance or refusal of vaccines despite availability of vaccination services), that has important negative consequences on rates on vaccines' acceptance rates and coverage [5], has been identified from the World Health Organization as one of the top ten global health threats in 2019 [6].

Investigating the causes for which part of the population has decided not to get vaccinated is therefore a way to understand the reasons and, above all, if there may be a solution, in order to prevent the situation from recurring in the future.

1.1. Vaccinations: society and the individual

Immunization of large populations thourgh vaccination is an important issue with serious implications for public health, especially if the context is expanded worldwide.

Due to adverse events, allergies, or health problems, not all individuals are eligible to get a vaccine. For this reason, in order to achieve the highest possible immunization rate, it is necessary that all those who can get vaccinated actually do so [7].

Although the choice to get vaccinated or not against Covid-19 was considered more a fulfillment of a duty towards the whole world, it must still be contextualized in a moment of strong physical and mental stress. Findings from recent studies [8] have shown that it is important to taking into account decision-making processes during stressful conditions, especially in the older and more physically vulnerable population. During the first months of Covid-19 pandemic, individuals were exposed to severe stress, due to both the rigid lockdown and, above all, health conditions of people around the world.

Vaccination is the only medical intervention that has been able to successfully eradicate a disease up to date [9]. At the same time, several issues have to be examined to highlight the complexities of vaccine development and use. There are rights of individuals in deciding about their own vaccination, while at the same time the rights of the society to be safe and immunized have to be preserved.

AIFA (Italian association for pharmacovigilance) has released the 11th , as well as the most recent, report on the surveillance of anti-Cvoid-19 vaccines, which covers the period from 27 December 2020 to 26 March 2022. It collects the reports of reactions observed after the administration of the vaccine; it does not imply that reactions were actually caused by it. Furthermore, it is important to specify that the benefit/risk ratio of a drug or vaccine is defined as favorable when the expected benefits exceed the known risks of the target population. Some definitions are specified below.

An adverse event is an adverse episode that occurs after the administration of a drug or vaccine, not necessarily caused by it.

An adverse reaction is a noxious, unintended response to a drug or vaccine, for which a cause-and-effect relationship is established.

An undesirable effect is an unintended, not necessarily harmful, effect related to the properties of the drug or vaccine.

The number of reports does not quantify the danger of the vaccine, but serves to monitor its safety. The distribution of reports and doses administered by type of vaccine is shown in Table 1.1.

Table 1.1: Reports, doses administered and related rates for currently authorized Covid-19 vaccines

VACCINE	REPORTS AT 26 MARCH, 2022	DOSES ADMINISTERED AT 26 MARCH 2022	REPORTING RATE (PER 100.000 DOSES ADMINISTERED)
COMIRNATY	89.315	88.552.383	101
SPIKEVAX	19.472	33.592.002	58
VAXZEVRIA	23.826	12.170.299	196
JANSSEN	1.731	1.507.726	115
NUVAXOVID	47	27.578	170
TOTAL	**134.391**	**135.849.988**	**99**

A total of 879 serious reports had a fatal outcome, with a rate of 0,65 events per 100.000 doses administered. Table 1.2 shows the distribution of these cases by type of vaccine.

Table 1.2: Distribution of reports of death by type of vaccine

Vaccine	Fatal cases	Rates per 100.000 doses administered
COMIRNATY	569	0,64
SPIKEVAX	162	0,48
VAXZEVRIA	31	2,06
JANSSEN	117	0,96
TOTAL	**879**	**0,65**

About a year after the start of the vaccination campaign, the Istituto Superiore di Sanità shows how the administration of the vaccine has reduced infections and serious hospitalizations in intensive care, as can be seen through Figure 1.1 (*Istituto Superiore di Sanità*).

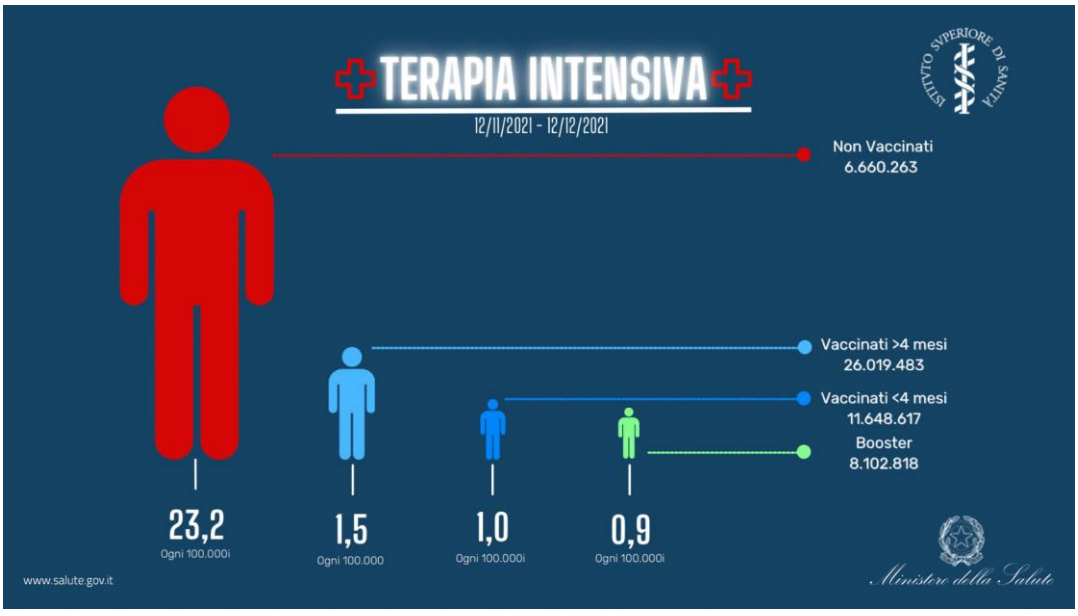


Figure 1.1: Intensive care admissions rate for Covid-19, grouped by population. From left to right are represented the unvaccinated, those vaccinated with one

For society as a whole, the complete elimination of a disease is the only path that can be considered, but what is best for the society might be seen as potentially differente from what will benefit individuals. Concerns about real or perceived adverse effects of the vaccines may lead individuals to disagree with government mandates for population wide vaccination. A greater awareness of the consequences of failure to vaccinate, for example through better health education, might be a way to address the problem [9].

1.2. Personas

The creator of Personas is Alan Cooper, who defines them as a design for specific indivduals with specific needs. Personas are not real people, but they represent them through a precise design. They are hypothetical identities, based on real identities. Through research and analysis, specific profiles are developed, with peculiar characteristics, which can describe the identities of the population being analyzed, contextualizing their objectives and identifying their main needs and requirements within a specific area of interest [10].

The usefulness of the Personas depends on the objective with which they are designed and built. Personas are not a tool that is built once and effective overall, they must be studied for one specific purpose and be adapted only to that one. They have found their main usage in marketing, where Personas are used to develop products and services aimed at particular segments of buyers and users of a product. In latest years, however, they usage is being expanded in the medical and psychological contexes, with the aim of improving therapies and identifying the best ways to convey messages, interventions, etc, supporting the decision-making process of professionals [11].

The more specific the Personas profile, the more useful it will be to achieve the intended goal. Personas are created through data analysis of socioeconomic characteristics, behaviors with respect to the objective, motivations and goals.

To better identify Personas a “Persona card” is built on each of them. The card-based technique is used for a user-centered design, where needs, wishes and limitations of the end user are the heart of the design project [12]. A Persona card, as in the example in Figure 1.2, is a intuitly and compact model constructed to present the Persona, accessing to all the informations concerning it. It provides a realistic representation, showing a name, demographic informations, what she likes and dislikes and, above all, her goals . The card is also provided with a name and a face, in order to reflect a real human being. Both photos and illustrations can be used, but

they are critical as they must describe each Persona, and they are extremely influential, affecting how Personas are perceived. It is important for the face's image to be effective, producing the desired effects in the observer [13]

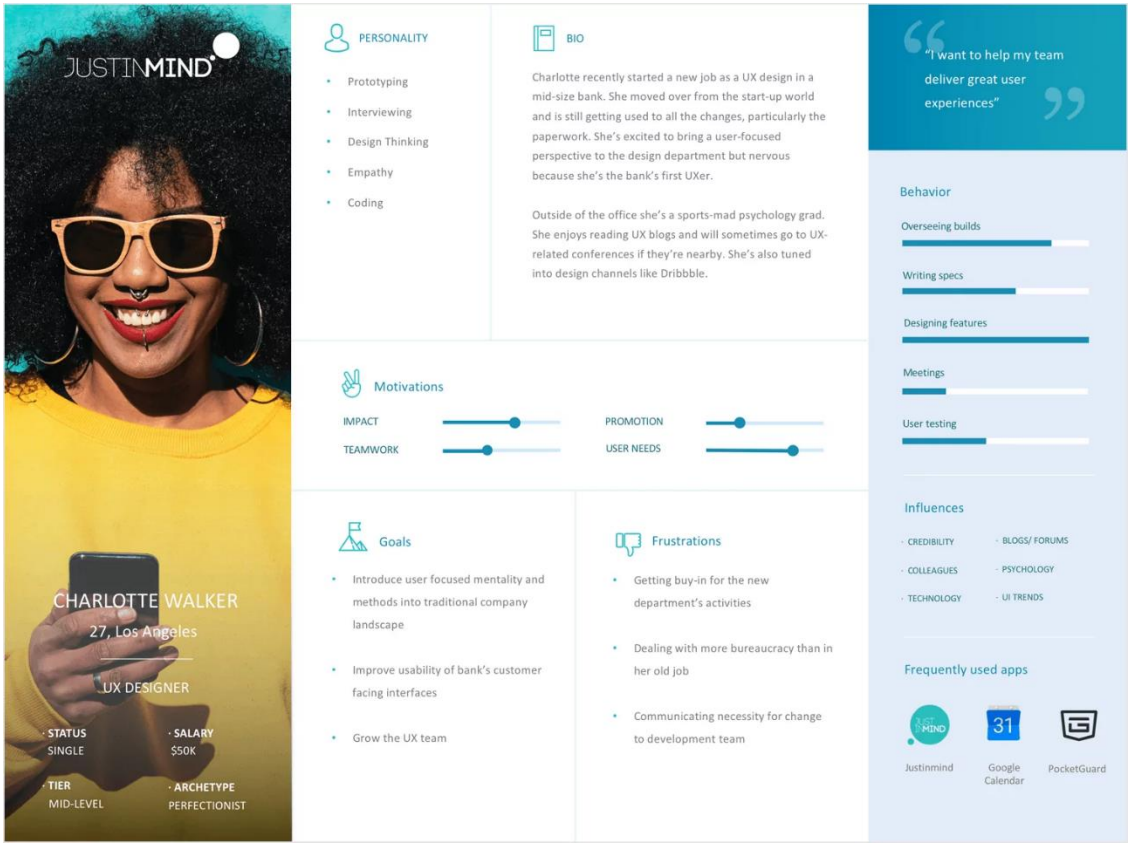


Figure 1.2: Persona card example, showing the photo, identity informations, goals and behaviour. Image taken from https://www.justinmind.com/blog/user-persona-templates/

1.3. Aim of the work

The final goal of this study is the evaluation of multiple statistical analysis and machine learning techniques to develop Personas for the willingness of vaccination against Covid-19.

The Personas have the aim of identifying underlying patterns between the socio-economical, lifestyle, psychological and health-related characteristics of the users, resulting in the acceptance or refutal of the Covid-19 vaccination.

Using the question "Do you intend to get vaccinated against Covid-19, when will you have the opportunity?" as the targert variable it was possible to apply supervised machine learning techniques to develop Personas, allowing the possibility of comparisons between clustering and statistical learning techniques against supervised approaches. The aforementioned question has been used to train predictive models developed through a supervised learning analysis, identifying models capable of creating Personas and assigning them to new respondents.

2 Materials and methods

In this chapter it is going to explain how it has been worked in order to achieve the set goal. Specifically, the data collection, the data preparation and the various machine learning techniques applied are deepened.

In Figure 1 a flowchart is presented to show the development of the whole process.

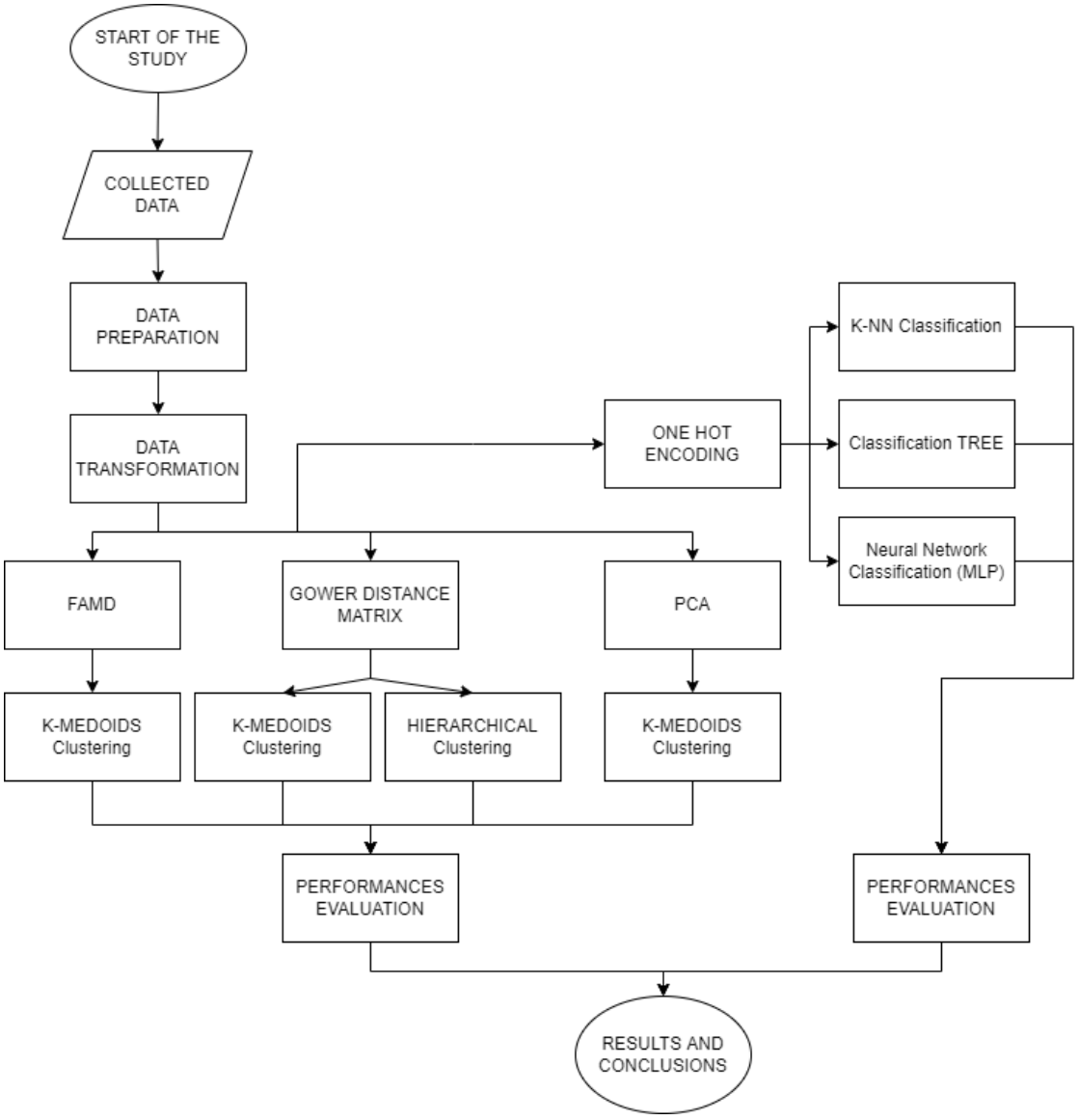


Figure 3.1: Flowchart of the study, emphasizing all analysis performed on dataset.

2.1. Data collection

An online survey on the emotional and social impact of the new COVID-19 was developed by the University of Milan, in collaboration with Centro Clinico Sempione APS. Any information collected by the participants was used completely anonymously. In fact, at the beginning of the compilation it was asked to enter an alphanumeric code, so that each compilation was uniquely identified, but not attributable to the subject. The code, consisting of 4 letters and 3 digits, is composed as follows: the first two letters of the name, the first two letters of the surname, the day of the date of birth, the last digit of the year of birth. Furthermore, given that the survey was completed electronically, data were also anonymized by removing the Internet Protocol (IP).

The online survey was developed and distributed online across Italy using the Qualtrics software (Provo, UT, USA) during January and February, 2021. Each participant who was administered the survey had the opportunity to give their consent to participate in the study, by checking the appropriate box.

The survey is composed of nine different blocks, each one assessing a different kind of information about the participants. Below, each section is described in further detail, with the related questions. Potential answers of the questions will not be reported to make the reading easier.

2.1.1. Sociodemographic factors

This section was an overview of the social, economical and lifestyle characteristics of the person, such as age, gender and profession. Below is the complete list of questions belonging to this section.

- Indicate your age
- What's your bilogical sex?
- What's your marital status?
- What's your ethnicity?
- What is the major educational qualification you have achieved?
- What is your current employment status?
- Do you or did you have a profession in the health sector among the following?
- What region do you live in?

2.1.2. Health

This part investigated the current physical and mental status of the participant, regarding any ongoing pathologies. The questions in this block are:

- Do you suffer from one or more of the following diseases? Please select one or more answers ONLY IF there is actually a medical diagnosis
- How do you consider your physical health to date?
- How do you consider your state of mental health to date?

2.1.3. Psychotherapy

This section investigated whether, and in case of what type, a therapy followed by a psychologist or psychiatrist was ever performed. Below is the list of questions in this block:

- Have you ever been treated by a psychologist and / or a psychiatrist in your life?
    - (if yes) What kind of treatment did you follow?
- What kind of difficulty prompted you to go to a psychologist and / or a psychiatrist? It can score more than one alternative.

2.1.4. Vaccinations

This section concerned the execution of vaccines, other than COVID, between the years 2019 and 2020, in order to investigate whether the intention to get vaccinated or not for Covid-19 was closely linked to vaccinations in general, or to the specific case of the pandemic in progress. Below the questions belonging to this group:

- Did you have one or more of the following vaccinations in 2019?
- Did you have one or more of the following vaccinations in 2020?

2.1.5. COVID-19

In this section it was investigated whether there has been positivity to the Covid, in the first person or among acquaintances or family members, and possibly in what form. Having already contracted the virus can affect the choice of vaccination in both directions, for this reason it was considered important to investigate in what form it was contracted. A more severe form can instill fear of the disease and can therefore lead to the possibility of a vaccine. At the same time, having already contracted it can make people feel protected, and therefore not needy of further coverage and protection.

- Have you tested positive for the new coronavirus-19 (COVID-19) following a swab and / or serological test?
- (if yes) In what form would you say you have contracted the new coronavirus-19 (COVID-19)?

- Have your loved ones tested positive for the new coronavirus-19 (COVID-19) following a swab and / or serological test?
- (if yes) In what form would you say that your loved ones have contracted the new coronavirus-19 (COVID-19)?

2.1.6. Perception of risk

The participants' estimated probability of acquiring COVID-19 was estimated through a visual analogue scale ranging from 0 (i.e. "Not likely at all") to 10 (i.e. "Very likely");

- How likely do you think it is for you to contract the new coronavirus / COVID-19?
- How likely do you think it is to contract the new coronavirus / COVID-19 for people similar to you, that is to say your age, your sex, similar health, doing a job similar to yours and with a lifestyle similar to yours his?

Participants' fear of acquiring COVID-19 (related to self, friends and family), trust in government, health institutions and science, and participants' opinions related to COVID-19 were all evaluated through a 5-points Likert scale (from 1 = "Strongly disagree" to 5 = "Strongly agree"), with the following questions.

- My health could be severely damaged if I contract the new coronavirus / COVID-19
- In general, the disease due to the new coronavirus / COVID-19 is more serious than the flu
- If I were to contract another disease other than the new coronavirus / COVID-19, I would not be taken to the hospital as I would be afraid of contracting the new coronavirus / COVID-19 in hospital.
- I am afraid of the possibility of contracting the new coronavirus / COVID-19
- I am afraid of the possibility that my friends may contract the new coronavirus / COVID-19
- I am confident in the measures taken by government institutions (government, local administrations) to deal with the health emergency due to the new coronavirus / COVID-19
- I trust the indications given by the scientific community (WHO, National Institute of Health, etc.) (10)
- I have faith in science and scientific research (12)
- I believe the novel coronavirus / COVID-19 vaccine can help resolve the current health emergency
-
- Do you believe that there are any responsible for the present Coronavirus epidemic (Covid-19)?

2.1.7. Perception of vaccine COVID-19

The next question asks for a personal opinion on the possibility that there are valid alternatives to the vaccine

- Do you think that one or more of the following preventive measures could / could be a valid alternative to the vaccine? You can select more than one alternative

Furthermore the willingness to get the COVID-19 vaccination [14] was assessed through the following question:

- As soon as it is possible for you, do you intend to get vaccinated against the new coronavirus / COVID-19?

Participants could answer “Yes”, “No” or “I do not know” and were then asked to motivate their responses, through the following open question.

- We ask you to write below the motivation for the answer given to the previous question: "As soon as it is possible for you, do you intend to get vaccinated against the new coronavirus / COVID-19?"

Finally, the subject's opinion was investigated about whether there are any “culprits” to blame for the pandemic. Here have been reported the answers to the first question to better understand its goal.

- Which of the following reasons could make you change your mind with respect to the answer given to the previous question? Select only the answer that you think is more important than the others.
    - I would only change my mind if the health authorities gave me directions to that effect.
    - I may change my mind if someone I trust gives me relevant information to that effect.
    - I could change my mind if a religious authority or spiritual guide in which I believe gave me indications in this regard.
    - I might change my mind if I read or see something (on the internet, on TV, or other media) that gives me good reason to do so.
    - I may change my mind if I am aware of cases of adverse reactions.
    - I could change my mind for reasons that do not appear in this list (if you wish you can write below which ones)
    - I wouldn't change my mind for any reason.
- Would you recommend that your loved ones get vaccinated against the new coronavirus / COVID-19?

- If he got the vaccine, would he continue at least for some time to implement the current preventive behaviors (mask, hand washing, avoid crowds, etc ...)?
- As far as you know, in your circle of closest acquaintances (e.g. friends, partners, family) are there people who have different opinions and intentions from yours regarding the vaccine?
- Do you think that in a year the situation will be better than today if the great majority of the population has undergone the vaccine?

2.1.8. GAD-7

Generalised Anxiety Disorder is a syndrome of ongoing anxiety and worry about many events or thoughts that the patien generally recognises as excessive and inappropriate [15]. GAD-7 is the Generalised Anxiety Disorder 7-item, a scale rating from 0 (never) to 3 (almost every day), investigates the frequency, in the last two weeks, of certain states of anxiety, restlessness or fear.

- In the LAST 2 WEEKS, how often have you experienced the following moods?
  - Feeling nervous, anxious and / or tense
  - Not being able to stop worrying or keep worries under control
  - Worrying too much about various things
  - Having trouble relaxing
  - Being so restless that you find it hard to sit still
  - Easily annoyed or irritated
  - Being afraid that something terrible might happen

2.1.9. MHLCS

MHLCS stands for Multidimensional Health Locus of Control Scale, it is a 6-item scale evaluating the Health Locus of Control [16], an area-specific measure of expectancies regarding locus of control developed for prediction of health-related behavior [17], assessed with a scale from 1 (completely disagree) to 5 (completely agree), as follows:

Please read the following statements carefully and respond by marking the corresponding degree of agreement.

- It doesn't matter what I do: if I have to get sick, I'll get sick
- Many things that could affect my health would happen by accident
- Luck plays an important role in determining how soon I could recover from an illness
- Regardless of what I do, I am likely to get sick.
- I will stay healthy if it is meant to be

2.2. Data preparation

The online surveys contains quantitative variables, such as age and psychological scales, nominal categorical variables such as sex and ordinal categorical variables such as level of education, thus requiring the usage of methods that can correctly assess mixed data types during the analysis.

As a requirement for the following analysis, data preparation is aimed to reorganize and re-process the data to make them usable, containing no missing or incorrect values and only relevant features.

Real-world data may be incomplete, noisy, and inconsistent, Through data preparation a new dataset is generated, potentially smaller than the original one, which can significantly improve the efficiency of subsequent analysis [18].

Several techniques can be employed to reach this goal [19]. All the ensuing analysis were performed using the Python language, version 3.10.3, with the help of openly available libraries.

An in-depth reading of each question posed in the survey was performed, defining the type of response associated with each question (quantitative, ordinal categorical or nominal categorical) and the frequency of those datatypes in the dataset.

Data preparation was divided into specific steps:

- Elimination of all records that cannot be used for analysis (that have not given consent)
- Analysis and resolution of missing values
- Elimination of attributes containing non convertible text
- Attributes removal as a result of data reduction techniques
- Attributes transformation

Below is the in-depth description of how the data preparation steps were executed, with the aim of obtaining a dataset suitable for further clustering and classification methods.

2.2.1. Data Cleaning

A record represents a row in the dataset, corresponding to the answers to each question of the survey of a single respondent. This step is the first to be performed as it eliminates records that would make the execution of the study impracticable.

As the subjects had to give their consent to use the answers provided for the study, 16 records were immediately eliminated as the subjects did not give their consent to participate.

Subsequently, the analysis of the missing values was carried out.

Missing values in the dataset can arise from information loss as well as dropouts and nonresponses of the study participants. The presence of missing values leads to unsuitable data and eventually compromises the analysis and the reliability of the study results. It can also produce biased results when deductions about a population are drawn based on such a sample, undermining the reliability of the data. As a part of the pretreatment process, the nature of missing data is invastigated. They can be either ignored in favor of simplicity or replaced with substituted values estimated with a statistical method [20].

Three types of missing values can be identified, differentiated by their generative models[21].

- Missing Completely at Random (MCAR), when the probability of a record having a missing value for an attribute is not realated to either the observed data or the missing data.
- Missing at Random (MAR), when the probability of a record having a missing value could depend on the observed data, but not on the value which is expected to be obtained.
- Missing not at Random (MNAR), thus the probability of a record having a missing value for an attribute could depend on the value of the attribute.

To handle with missing data, the following strategies can be applied.

Listwise deletion, discarding all records for which the values of one or more attributes are missing. This approach can be used on MCAR kind of values, otherwise, if the dataset is not large enough, it may introduce bias in the analysis.

Pairwise deletion, eliminating the information when the missing value is needed to test a particular assumption. Pairwise is used when the type of analysis that has to be applied refers to a pair of variables. Therefore in this case it is decided to eliminate that pair from that specific analysis. Depending on the analysis to be performed, it has to be chosen to keep or omit the values are used or not, preserving more information than the previous approach. MCAR and MAR data could be treated with paiwise deletion, however, when the dataset has many missing values, the analysis will be complex. Substituing missing data through an algorithm, for example with the mean of the attribute calculated for the remaining observations. This technique can only be applied to numerical attributes and to data that are random, thus it is not indicated for MNAR data. Furthermore it does not add new informations to the dataset.

Inspection, carried out in order to obtain recommendations on possible substitute values.

Identification, where a conventional value might be used to encode and identify missing values, making it unnecessary to remove entire records.

Both inspection and identification are suitable primarily for not random missing values, so as to use the missing values as informations to analyse, not to eliminate.In the database under analysis, multiple null values were identified in all observations. Inspection was carried out on each record to assess their suspected generative models.

For each record, null values were found in correspondence of variables that identified keywords derived from open answers. As an example, the "Willingness to being vaccinated" question is taken into account. It is a multiple choice question with options "yes", "no" and "I don't know". With respect to the chosen answer, a following open question ensues, exploring the reasons for the given answer.

The motivations provided about the willingness/unwillingness to get the COVID-19 vaccination passed through a 3-steps categorization process. The first step consisted in an independent categorization made by three experimenters. The second step consisted in comparing those three independent categorizations and keeping only those categories that had been suggested by at least two experimenters. The third step consisted in an external revision of the categories by a fourth experimenter. The final emerged categories were 21. This means that all observations were categorized first by the willing question, and then further broken down into 21 other attributes [14]. Where the category corresponded to the subject's response it was tagged with the value 1, while a null value was generated in the remaining 20 columns. These missing values were thus identified as not missing at random, since their generative model was well known. Observing the distribution, it was seen that the sparse diversification of the answers was detrimental to further analysis. Thus, all the attributes corresponding to reasons for getting or avoiding vaccination were removed as deeemed not useful, while keeping the general distinction between "Yes", "No", and "I don't know" answers.

Other 15 missing values were found inside Optimistic Bias attribute. It represented the diffrence, in value, between two other attributes, deriving from the question: "How likely do you think it is for you to contract the new coronavirus / COVID-19?" and "How likely do you think it is to contract the new coronavirus / COVID-19 for people similar to you, that is to say your age, your sex, similar health, doing a job similar to yours and with a lifestyle similar to yours his?". Inspectioning values, 12 of null values derived from corresponding null values inside the first question, and 3 derived from the second one. All 15 records with missing values in Optimistic Bias were removed from the dataset, as this variable was considered very relevant for the study, as it

quantifies the bias between the perception of the risk of being able to contract the virus in the first person and that other similar people can contract it.

2.2.2. Standardization

The collected quantitative variables presented different units of measurement and scales between them. To make sure that all variables contributed with the same importance to further methods, it was deemed necessary to perform a standardization of the values. Standardization is defined as the transformation of all quantitative variables into a single measurement scale, establishing a single range, with a minimum and maximum value equal for each variable. In this way each value is recalculated to set it in the new scale. Feature Scaling involves re-scaling of values to be able to make further analysis easier to be performed [22].

In this work, the Quantile Tranformer Scaler was applied on all numerical variables, as it can increase the performance when dealing with thousand data points or more [23]. It converts the variable distribution to a normal distribution and scales it accordingly, reducing also the impact of marginal outliers.

Here a description of most important steps:

1) It computes the cumulative distribution function of the variable.

   Quantile Transformer Scaler ranges all data to the desired distribution based on

$$G^{-1}\big(F(X)\big),$$

   where: F is the cumulative distribution function of the variable, G^{-1} is the quantile function of the output distribution G.

2) It maps the obtained values to the desired output ditribution using the associated quantile function:

If X is a random variable with a continuous cumulative distribution function F, then F(X) is uniformly distributed on [0,1]. If U is a random variable with uniform distribution on [0,1], then $G^{-1}(U)$ has distribution G.

2.2.3. Feature selection

Feature selection is the process of reducing the number of input variables in order to both reduce computational cost of prediction and to improve the performance of the models [24].

Changes made are specified below.

The group of features concerning the different types of vaccinations performed in 2019 were merged into a new variable, called "Vaccination_2019", where all those who had performed at least one vaccination were categorized as 'Yes', while those that had not performed at least one vaccination were categorized as 'No'.

The same procedure was applied for the 2020 vaccinations, that were merged in the single binary variable "Vaccination_2020".

It was subsequently noticed that all those who had vaccinated at least once in 2019 were the same as those who had vaccinated at least once in 2020. Thus, the two variables "Vaccination_2019" and "Vaccination_2020" were combined into a single variable containing values as follows: 'No' for those who did not vaccinate neither in 2019 or 2020, and 'Yes' for those who did vaccinate both in 2019 and 2020. The two variables used for combination were thus removed from the dataset.

Answers relating to the *Country of origin* variable, corresponding to each Italian region, belong almost entirely to the Lombardy region, but the presence, although minimal, of the other regions in northern, central and southern Italy was still noted. For this reason it was decided to aggregate the data, grouping the answers into four new values: "north", "south", "center", "islands".

Question related to the fear of contracting covid about themselves, family and friends were merged into the single *Fear of Covid* variable, containing the sum of the values of the individual questions.

The group of *Trust* variables, which identified through a 5-point Likert Scale the trust levels of Government, Health and Science institutions, and institutions in general, were merged into a single *Institutions trust* variable, containing the sum of all previous ones.

Having chosen to use "Willingness" as the target variable, it was necessary to make it binary, as well as all the variables relating to the same concept. In this sense, all the attributes that originally contained 'yes', 'no', 'I don't know' as values, have been transformed into binary attributes. This operation has been completed considering the 'I don't know' values as 'no' values since it is believed that a subject undecided about vaccination, however, shows a hostility or a perplexity that must be investigated, unlike those who answered 'yes' with certainty.

Also for the variable *Vaccine opinion others*, asking if there were acquaintances or friends with different opinions regarding vaccine, was decided to aggregate answers. Indeed, to simplify, it has been transformed into binary attribute, transforming the answers 'I don't know' into 'no' and the answers 'yes, some' and 'yes, many' into 'yes', as it was considered that 'some' and 'many' did not add relevant information to what

is already an affirmative answer. The *Ethnicity* variable was removed as deemed irrelevant, because more than 99% of the records matched the same value.

In addition to the attributes already listed above, through the creation of a heat map, correlations between variables were analyzed (Figure 2.2) [25]. A heat map is a graphical representation of a correlation matrix that translates data values into colors within a matrix. The value of correlation can take any value from -1 to 1, with -1 identifying strong negative correlation and +1 identifying strong positive correlation. Values close to 0 are significant of no correlation between examined variables. This type of data visualization summarizes a vast amount of data within a single snapshot, which helps to quickly communicate relationships between values [26].

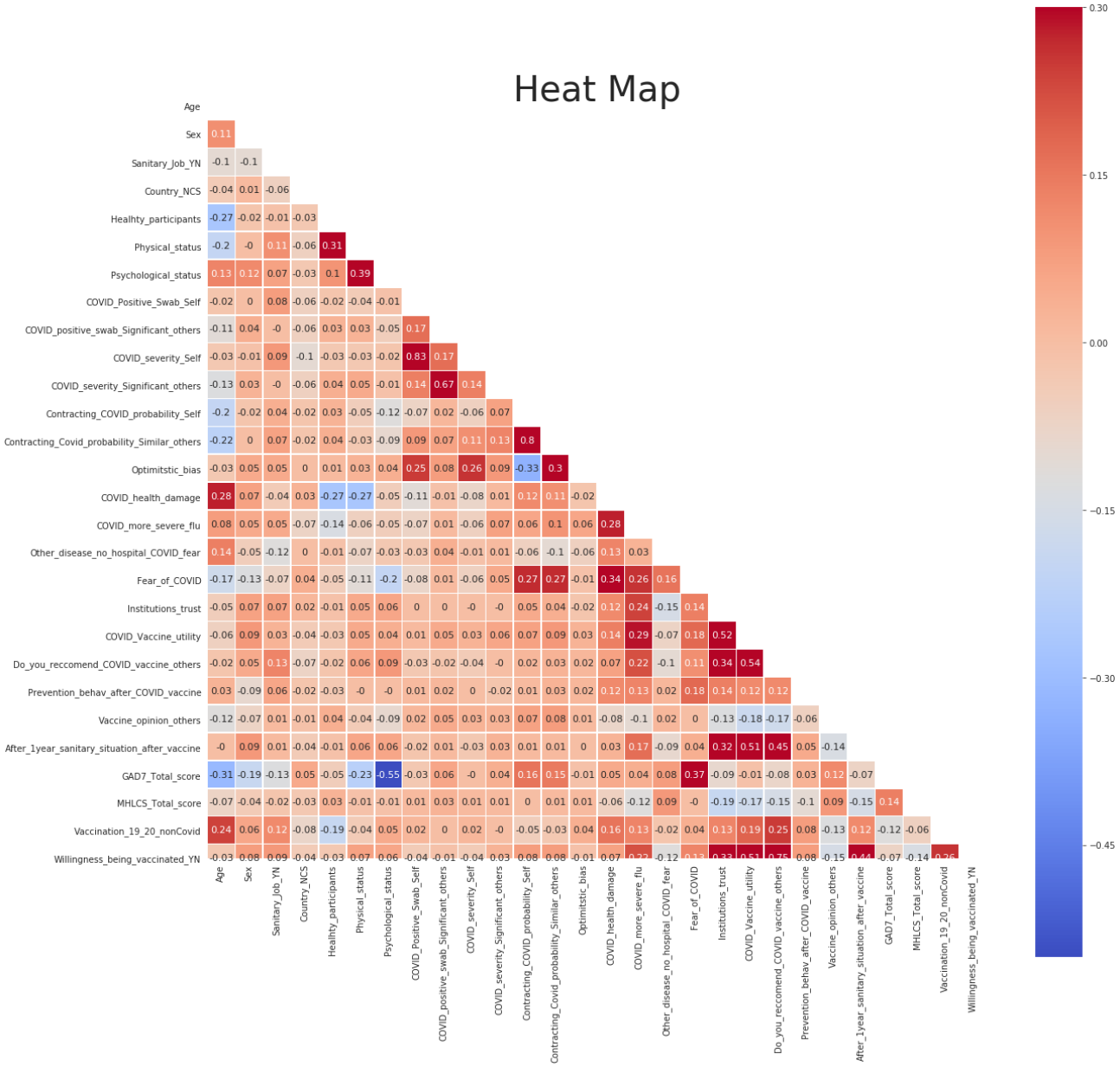


Figure 2.4: Heat map correlation matrix of the analyzed dataset. Only the lower triangle is shown due to the symmetric property of the heat map.

Furthermore, nominal categorical variables were transformed through one hot encoding. In nominal variables the numerical values assigned to each category hold no meaning. In this kind of variables it is not possible to assess that a value is higher than another one, as being widowed is not higher than being married, and it is not possible to perform mathematical operations between values, such as adding “married” and “widowed”. Thus, through one hot encoding, a new binary dummy variable is created for each possible category, or answer. A value of 1 is given to the

dummy variable corresponding to the given answer, while a value of 0 is given to all other dummy variables.

As an example, the attribute *Marital Status* presented 4 possible values: single (1), in a relationship (2), married (3), widowed/other (4). Through one hot encoding, this variable was divided into four dummy variables, one for each potential answer. For each respondent, the value 1 was assigned to the dummy variable corresponding to the given answer, and the value 0 to all other ones. In case of multiple choice questions, more than one dummy variable can have the value 1. This transformation is equal to transforming a single question into a series of yes/no questions (i.e. "Are you single?", "Are you in a relationship?", "Are you married?", "Are you widowed/other?").

From here on, further methods were applied to two cases: the complete database and another one without the W*illingness being vaccinated_YN* column, which was defined as a target variable and accordingly removed..

At the end of these operations, the number of attributes was reduced from 131 to 42.

2.2.4. Principal component Analysis (PCA)

PCA is one of the most widely known technique of dimensionality reduction by means of projection. It defines principal components, uncorrelated linear combinations of variables that explain the highest amount of variance in the dataset. These principal components constitute a new basis for the projection of data. They allow to reduce the number of features in a dataseet while mantaining most of the information. It can also be inferred that the percentage of variance, or information, lost during this process is attributable to noise in the data [27].

Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance, that is the axes that captures most information of the data. The larger the variance along a principal component, the more the information it has. Principal components are constructed so that the first principal component accounts for the largest possible variance in the data set. The following ones are calculated in the same way, with the condition that they are linearly uncorrelated with preovious ones, orthonormal to them, and accouns for the next highest variance amongst remaining principal components.

PCA needs no distributional assumptions and can be used on quantitative, ordinal categorical and binary variables, thus requiring one hot encoding of nominal categorical variavbles to be applied.

Starting from dataset X, the covariance matrix V is constructed,

$$V = X^T X \quad (1)$$

Let p_j, j belonging to N, the n principal components, obtained as the linear combination of

$$p_j = X u_j \quad (2)$$

where u_j are the weights to be determined. U denotes the nxn matrix whose columns are the eigenvectors u_j, and P indicates the nxn matrix whose columns are the principal components p_j.

Then it has to be computed the w_j of length 1 that maximizes the variance of p_j. This can be calculated as

$$\underset{||w||=1}{\operatorname{argmax}}(p^T p) = \underset{||w||=1}{\operatorname{argmax}}(w^T X^T X w) \quad (3)$$

where the matrix X is assumed to be mean-centered; the optimal u is the first eigenvector of the corresponding cross-product matrix $X^T X$.

The principal component p_h assumes the form

$$p_h = u_{h1} a_1 + u_{h2} a_2 + \cdots + u_{hn} a_n. \quad (4)$$

The coefficient u_{hj} can be intepreted as the weight of the attribute a_j in determining the component p_h.

From the n attributes of the initial dataset X, the principal component method derives n orthogonal vectors, principal components, which constitute a new basis of the space R^n.

At the end of the procedure the principal components are ranked in non-increasing order with respect to the amount of variance that they are able to explain [19], [27].

In this study PCA was applied twice, one on the complete database, another one on the same database but deprived of the target variable "Willingness being vaccinated_YN". In both cases, a total of 12 variables were calculated, cumulatively explaining al least 75% of the total variance (Figure 2.3).

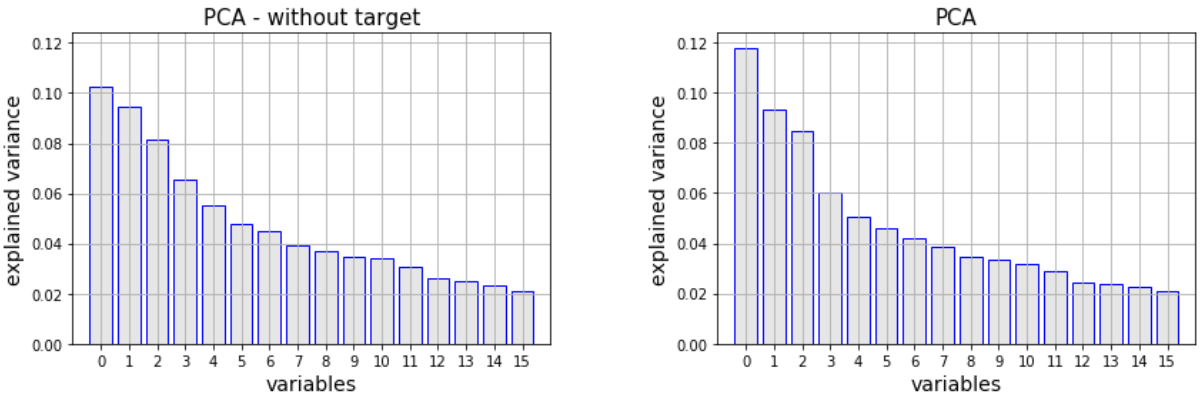


Figure 2.5: Comparison between the computed PCA: on the left, PCA on the dataset

2.2.5. Factor Analysis of Mixed Data (FAMD)

As an alternative to PCA, FAMD is designed to handle dataset with mixed types, both quantitative and categorical. It has the same scope as PCA, but when data contains a combination of variables, FAMD is a valuable alternative approach. It can be seen as combining PCA for continuous variables and Multiple Corrispondence Analysis (MCA) for categorical variables. Multiple factor analysis is extended to include these types of variables in order to balance the influence of the different sets when a global distance between units is computed. [28]. Due to FAMD implementation in Python, one hot encoding is automatically performed on the nominal categorical variables of the input dataframe.

Also in this case, applying the FAMD to the database with and without target, to reach at least 71% of cumulative inertia, 22 columns were obtained in both databases (Figure 2.4).

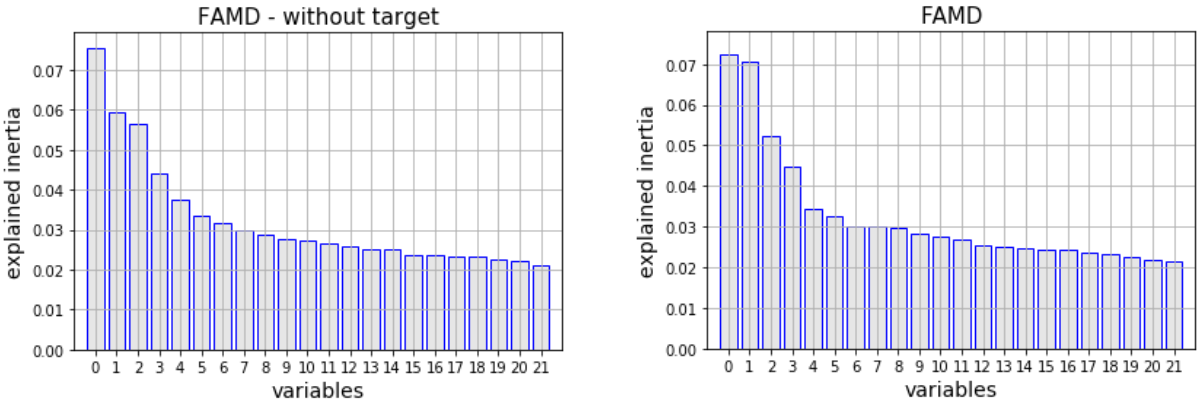


Figure 2.8: Comparison plot of inertia explained by eigenvectors, computed by FAMD, on two different datasets. On the left, dataset without target variable "Willingness being vaccinated Yes/No"; on the right, the dataset containing the target variable.

2.2.6. Gower Distance

Gower Distance is a measure that can be used to calculate distance between two entity whose attribute are a mix of nominal categorical, ordinal categorical and quantitative values [29]. To calculate the distance between observations *i* and *j*, GD is computed as the average of partial dissimilarities (pd) across the *m* features of the observations.

$$GD_{ij} = \frac{1}{m}\sum_{f=1}^{m} pd_{ij}^{(f)} \quad (5)$$

Partial dissimilarities (pd) calculation depends on the type of the feature being compared.

For a numerical feature, the pd of two individuals *i* and *j* is the difference between values in the specific feature (in absolute value), divided by the total range of the feature (R_f).

$$pd_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f} \quad (6)$$

$$R_f = \max f - \min f \quad (7)$$

For a categorical feature, the pd is also known as Dice Distance. Whenever the observations have exactly the same value, Dice distance is equal to 0. When they're not equal, Dice distance is calculated as

$$DiceDistance = \frac{NNEQ}{NTT + NNZ}, \quad (8)$$

N: number of dimensions;

NTT: number of dimensions in which both values are True

NTF: number of dimensions in which the first value is True, second is False;

NFT: number of dimensions in which the first valuea is False, second is True;

NFF: number of dimensions in which both values are False;

NNEQ (number of non-equal dimensions) = NTF + NFT;

NNZ (number of non-zero dimensions) = NTF + NFT + NTT.

Partial dissimilarities have a range from 0 to 1 in both categorical and numerical features. The same range will be obtained after averaging all features to calculate the complete GD. Zero means that the observations completely equal, while one means that they are completely different.

The output of the Gower distance is a symmetric matrix of n rows and n columns, where n is the number of rows in the original dataset, defined as the number of observations. Each cell of this symmetric matrix contains the $GD_{i,j}$ between points *I* and *j*, with the diagonal being composed only of zeros, representing the distance from each point from itself.

2.3. Clustering Method

By defining appropriate metrics and the induced notions of distance and similarity between pairs of observations, the purpose of clustering methods is the identification of homogeneous groups of records. With respect to the specific distance selected, the division should be done in such a way that the observations are as similar as possible to each other within the same cluster.

Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. In this study the first two methods were applied. In partitioning clustering, the algorithm categorizes the data points into k partitions, where each partition represents a cluster. In hierarchical clustering the algorithm builds a hierarchy of clusters starting from the bottom, with each element of the dataset as a cluster, until all elements belong to a single cluster. [30]. The methods used in this work will now be explaied in further detail.

2.3.1. K-Medoids Clustering on PCA and FAMD

K-Medoids algorithm [31] is a partitioning clustering method. It derives from the more common K-means clustering [32], that divides the dataset into *k* partitions, or clusters. It has the requiriment of defining *k*, the number of clusters to be created, a priori before the algorithm begins. The main difference between K-means and K-medoids is in the computation of the central point of each cluster. In K-means, the central point of each cluster, or centroid, is defined as the point in the space that minimizes the distance from all other points in the same cluster. In K-medoids, the central point of each cluster, or medoid, must satistfy the previous requirement of a centroid while also being an actual element of the dataset. Furthermore, K-means aims at minimizing the mean square error, while K-medoids aims at minimizing the sum of dissimilarities

between each point in the dataset and the medoid of the corresponding cluster. In literature, comparisons between K-means and K-medoids clustering show that from an execution time point of view, the K-Medoids algorihm performs reasonably better than the K-Means when data points are increased. [33]. It is also less sensitive to outliers, while also reducing noise as it minimizes the sum of dissimilarities of clustered points [34]. For these reasons, K-Medoids algorithm was the one chosen for this application.

In K-medoids, the Partitioning Around Medoids (PAM) algorithm was chosen due to its requirement of less than ten thousand subjects [35]. It takes the input parameter *k*, the number of clusters to be created, and performs the following steps:

1) Starting k medoids are chosen through the n data points of the dataset. This is done through the k-medoids++ algorithm, an optimization algorithm based on the k-means++ algorithm that gives more spread out starting points when compared to random choice, while also reducing the possibility of encountering a local minimum [36].
2) For all remaining non-medoid data points, the distance from all medoids is computed through the usage of the square euclidean function.
3) Each non-medoid data point is assigned to that cluster that minimizes the medoid distance.
4) The Total Cost (TC), defined as the sum of square euclidean distances of all non-medoid data points from the medoid of their assigned cluster, is calculated and defined as dj.
5) A non-medoid data point *i* is randomly selected.
6) The data point *i* is swapped with a random medoid *j*. The total cost is calculated as in step 4, defined as *di*.
7) If di < dj, then the temporary swap done in step 6 is permanent and forms a new set of k medoids. Otherwise, the temporary swap is undone.
8) The swap phase, defined as the combination of steps from step 4 to step 8, is repeated convergeance is reached or a pre-defined maximum number of iterations is reached.

In Figure 2.11 a schematization of the process described above is shown.

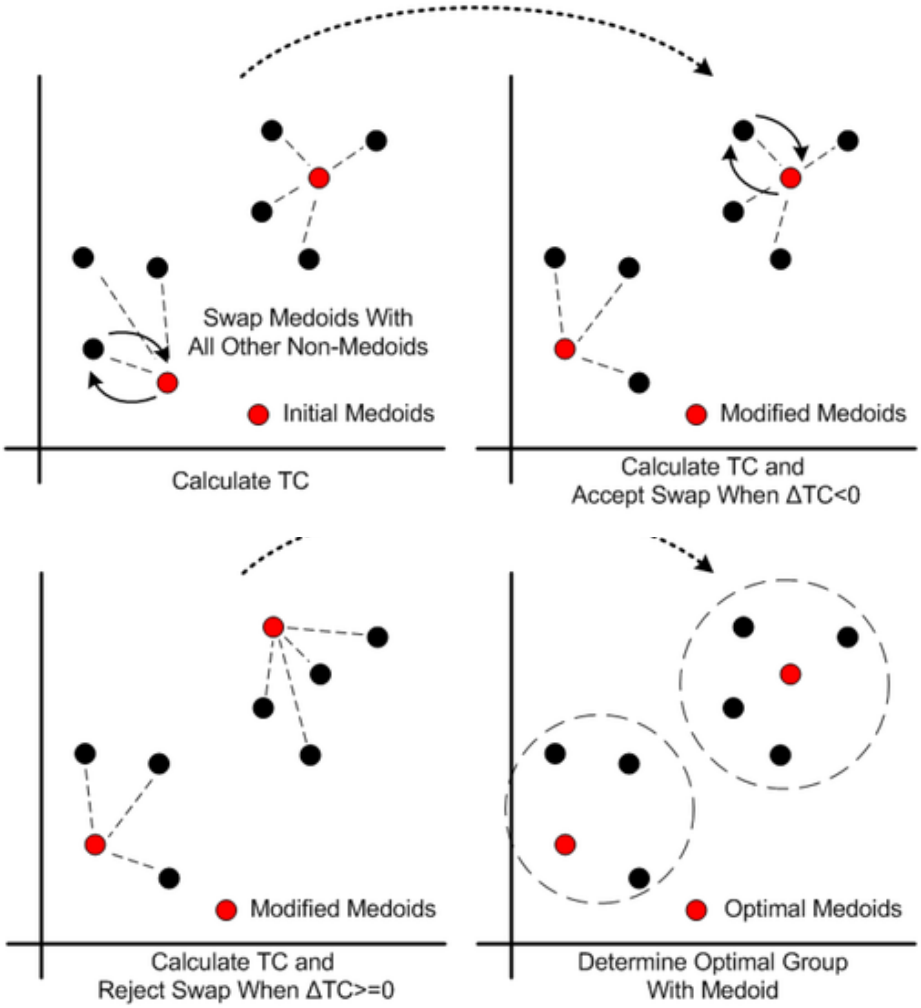


Figure 2.11: Swap phase process in K-Medoids clustering with PAM algorithm. Through this phase the optimal medoids for each cluster are identified.

K-Medoids clustering with PAM algorithm was applied following the preprocessing performed earlier through the usage of PCA and FAMD.

2.3.2. K-Medoids Clustering on Gower Distance Matrix

In addition, the K-Medoids clustering with the previously presented PAM algorithm was also applied following the calculation of the Gower distance.

Since Gower distance is defined as non-Euclidean and non-metric, changes had to be performed on the clustering algorithm to ensure a correct application. Thus, in the previosuly defined algorithm, the distance between data points was assumed as

already precomputed. K-medoids was thus further simplified, not requiring to perform additional calculations to identify the distance between data points.

2.3.3. Hierarchical Clustering

Hierarchical Clustering is a distance-based algorithm, whose output can be seen as a tree structure. It does not require the number of clusters to be determined a priori. In this work an agglomerative clustering algorithm, built with a bottom-up approach, was chosen. In agglomerative clustering, data points are aggregated through multiple iterations on the basis of the distance between them, deriving clusters of increasingly larger cardinalities. The algorithm is completed when a single cluster including all the records has been created [37].

It requires in input a precomputed distance matrix, such as the output of the Gower distance calculation. It then performs the following steps:

1) It starts by defining each data point as a cluster. The number of clusters at the beginning is N, equal to the number of data points;

2) A new cluster is formed by combining the two closest clusters and their respective data points. This results in N-1 clusters.

3) The distance between the newly formed cluster and all other clusters is calculated through a linkage method;

4) Steps 2 and 3 are repeated until the algorithm reaces N=1, or one single cluster contaning all data points [29].

Multiple linkage methods are available in literature to calculate the distance between two clusters composed of multiple data points. However, since the Gower Distance is defined as non-euclidean, commonly used algorithms for linkage such as Ward [38] and UPGMC [39] were not available due to the inherent constraint. Defining two clusters as u and v, and the distance between these two clusters as d(u,v), let x and y be respectively the elements belonging to cluster u and cluster v. To compute clusters distance, the following linkage methods, schematized in Figure 2.6, could be used [40]:

Single: the distance between two clusters is defined as the minimum of the distance between points belonging to each cluster.

$$d(u,v) = \min(d(x_{ui}, y_{vj}), i \in (1, \dots, n_u), j \in (1, \dots, n_v) \quad (9)$$

Complete: the distance between two clusters is defined as the maximum of the distance between points belonging to each cluster.

$$d(u,v) = \max(d(x_{ui}, y_{vj}), i \in (1, \dots, n_u), j \in (1, \dots, n_v) \quad (10)$$

Average: the distance between two clusters is defined as the average of the distance between points belonging to each cluster.

$$d(u,v) = \frac{1}{n_u n_v} \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} d(x_{ui}, y_{vj}), i \in (1, \dots, n_u), j \in (1, \dots, n_v) \quad (11)$$

Centroid: similar to average linkage, it produces a tree in which the distances from the root to every branching point are equal. This approach takes the distance between the centroids of the data points in clusters.

$$d(u,v) = d\left(\left(\frac{1}{n_u}\sum_{i=1}^{n_u}\vec{x}\right), \left(\frac{1}{n_v}\sum_{j=1}^{n_v}\vec{y}\right)\right), i \in (1, \dots, n_u), j \in (1, \dots, n_v) \quad (12)$$

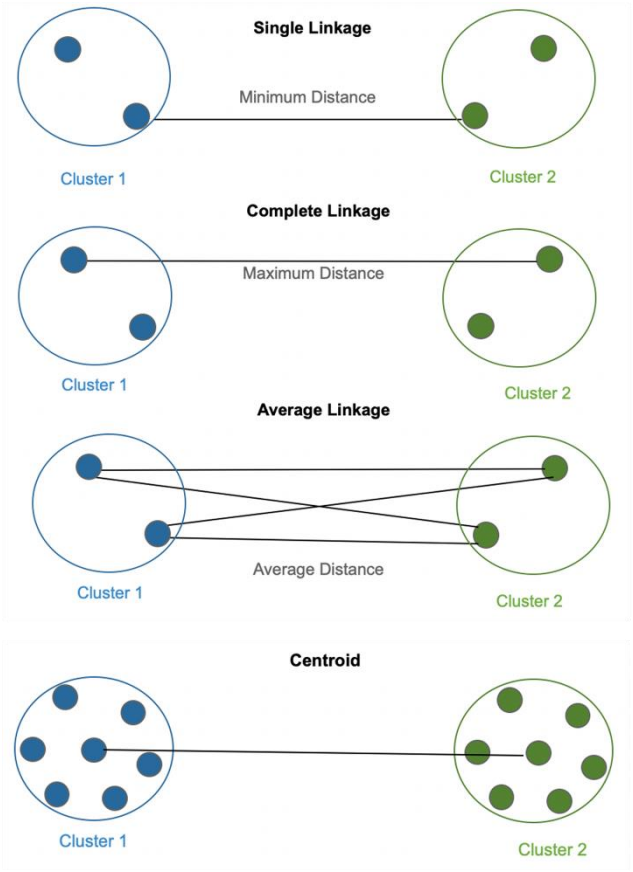


Figure 2.14: List of graphical examples of used linkage methods.

For the purpose of this work, the available linkage methods were tried, and Complete linkage was chosen as the optimal one.

A dendrogram can be used to visualize the result of the algorithm. On the y axis it shows the value of distance corresponding to each merger, and on the x axis the set of observations. It is composed by drawing a U-shaped link between non-singleton cluster and its children. Whenever two clusters are merged, they will be joined in the dendrogram, and the height of this join will be the distance between clusters.

In Figure 2.7 a dendrogram is shown in two phases of the algoritm. On the left, after the first iteration of the agglomerative clustering algorithm, with one new cluster combined from two datapoints. On the right, after the algorithm is completed and the full dendrogram is drawn.

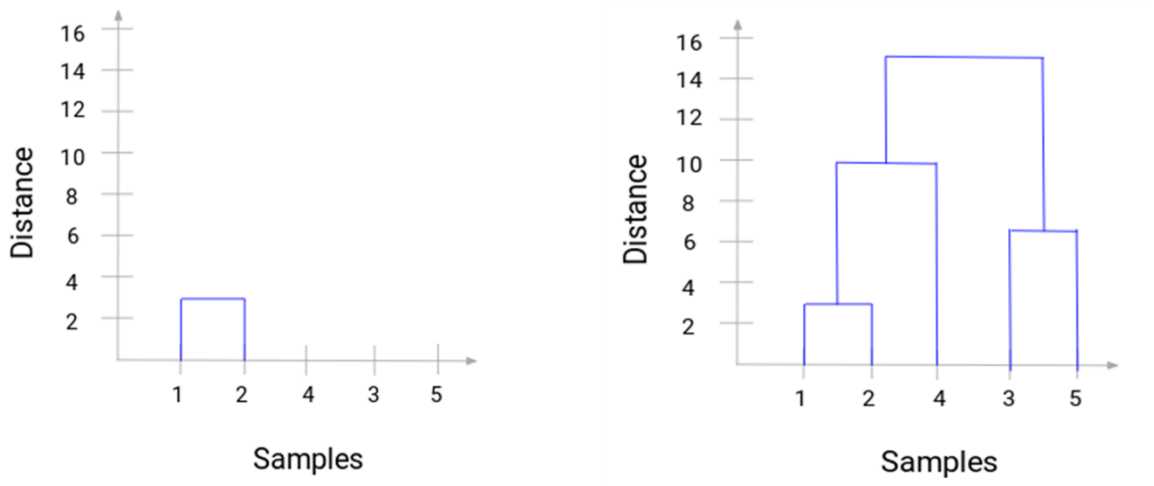


Figure 2.15 Plot of the composition steps of a dendrogram. On the left, the first step in drawing a dendrogram. On the right, the completed dendrogram.

After the dendrogram has been plot a distance threshold is chosen to cut the clusters and obtain the dataset partitioning. Depending on the distance threshold a different number of clusters will be created. As an example, in Figure 7, a distance threshold of 12 would result in two clusters, while a distance threshold of 4 would result in four clusters.

2.3.4. Evaluation of clustering models

In order to define the best clustering method among the ones presented, and to identify the optimal number of clusters to obtain through said method, the average silhouette score was used. [41]. Each point in the dataset is defined by a silhouette score, assessing where that point lies with respect to his own clusters and all the other ones [42]. The silhouette score is calculated for each point x_i, defining a(xi) as the average distance of xi from all other points in the same cluster and b(xi) as the minimum of the distances between point xi and all points in clusters different from its own. It is then possible to define the silhouette score for point xi as [43]:

$$S(xi) = \frac{b(xi) - a(xi)}{\max\left(a(xi), b(xi)\right)}. \qquad (13)$$

The silhouette score S(xi) can range from -1 to +1. A score close to +1 means that the point is appropriately clustered, since the distance of point xi from the closest point of another cluster is higher than the average distance of xi from points of the same cluster. A score close to zero means that point xi is close to the border between two clusters, while values lower than 0 mean that the point is not appropriately clustered, as it is closer a point belonging to a different cluster rather than the average of points in its own.

The silhouette score is calculated for each point in the dataset and then averaged to obtain the average silhouette score, related to a certain method and a certain number of clusters.

In Figure 2.8 an example of average silhouette plot ranging from clusters 2 to 10. The best option is the peak in the graph, or the highest value of the average silhouette score, found for k = Y with Hierarchical clustering.

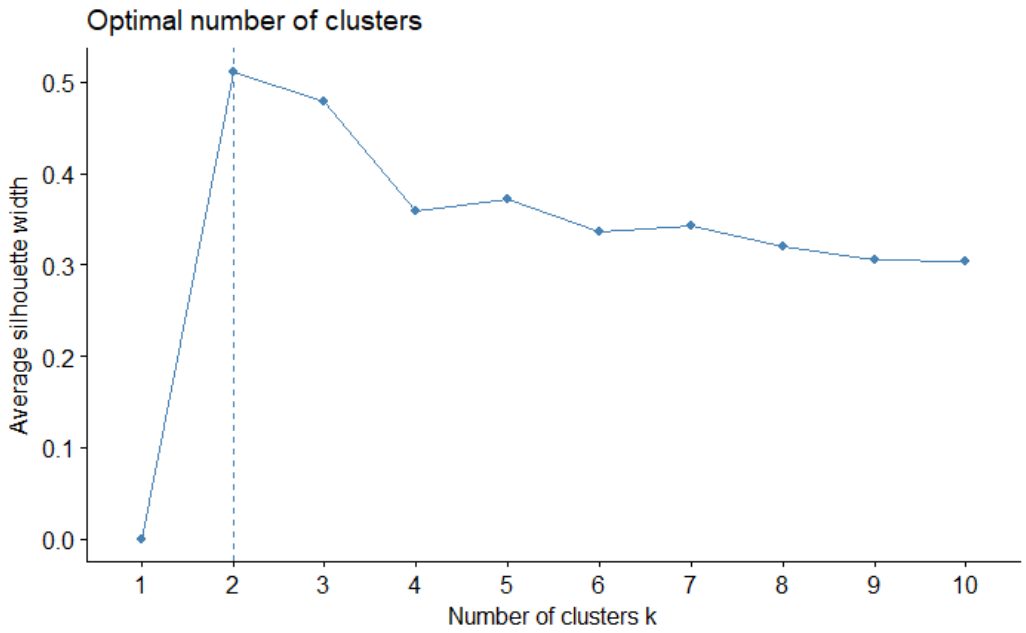


Figure 2.16: Average Silhouette score representation plotted over a varying number of clusters from 1 to 10.

2.3.5. Statistical Analysis

Statistical tests were used to discern which variables were significantly different between obtained clusters. In this study, statistical tests were used to identify the characteristics that differentiate between Personas on the basis of collected data and applied preprocessing and clustering method.

In these tests of significance, P is an indicator of the strenght of evidence against a null hypothesis: P is the probability of the trial outcome under the assumption that the null hypothesis H_0 is true. H_0 refers to there being no difference between, for example, the means or proportions of two or more populations, based on applying tests of significance to the samples taken from them. This probability P is also called p-value.

The chosen threshold to reject the null hypothesis is set at $p < 0.05$, as the literature suggests [44].

To apply the analysis on the resulting clustering, non-parametric tests where preferred because they do not assume the population data belongs to a normal distribution [45], [46]. Three different types of statistical non-parametric tests were chosen depending on the number of clusters to analyse and the type of variable in analysis:

1) Mann-Whitney U test for differences between two groups on a single, ordinal variable with no specific distribution [47], hence used on clustering with only 2 clusters. The assumptions are specified below:
   - The dependent variable, hence the attribute that has to be evaluated as significant or not, should be on an ordinal scale. For this reason, no binary variable can be evaluated with this test
   - The independent variables should be the two independent categorical groups
   - There should be no relationship between the two groups

2) Fisher's exact test, used to compare the distribution of variables in a sample. First, the test involves the creation of a contingency table for each variable. A contingency table is a frequency distribution table, where two or more variables are shown simultaneously. The values at the row and column intersection are frequencies for each unique combination of the two variables. Even though it is has a high computational cost, Fisher's exact test was used as it does not require the minimum number of frequencies to be 5 for each cell in the contingency table, as request by the Chi-squared test. A combination of Python and R programming languages was used to perform Fisher's exact test calculations [48].

3) Kruskal-Wallis test assesses the differences among three or more independently sampled groups on a single, non-normally distributed continuous variable [49]. If the result is a p-value lower than 0.05 for an attribute, then more comparison have been performed. The Mann-Whitney U test has been applied to each combination of two clusters, and the new computed p-value has been compared with a threshold. That is, the so called Bonferroni correction establishes the calculation of a new p threshold to consider the p-value significant [50]:

$$p < \frac{\alpha}{N}$$

Where α = 0.05 and N is the number of combinations performed. As an example, having 4 clusters, the resulting combinations to which apply the Mann-Whitney U test are 6, therefore the new p-value threshold under which consider the attribute as significant would be p = 0.008.
Hence, after calculating the p-value using Kruskal-Wallis test, the goal is to understand between which clusters there is a statistically more significant difference for that specific attribute.

2.4. Classification

Classification is a supervised machine learning technique used to predict group membership for data instances. It is a two steps process of learning and testing: in the first step, a model is trained on a set of past observations whose ground truth label is already known, in order to understand the underlying patterns of the data and generate a set of rules. In the second step, the trained model tries to predict the class of future data that was never seen before. [51].

In Figure 2.9 an overview of the training and testing process for supervised machine learning algorithm is given.

Before training the model, the starting dataset was divided into train and test set. 75% of data was kept for training purposes, while 25% of data was to be used for testing. The train set has been further subdivided, taking 75% of the data for the final train set and 25% for the so-called validation set. Splitting was stratified in both splits, so that in training, validation and testing data the percentages of labels were kept identical. This is fundamental for unbalanced data, as it allows to maintain the proportions of classes and enable the model to generalize on future data. The issue of unbalanced data will be further discussed in the following section [52].

After splitting, the model was trained on the training set. Classification algorithm was given in input the data and the ground truth labels belonging to the subset of the dataset, in order to fit the model. Fitting the model corresponds to adjusting the parameters characterizing it, allowing it to understand the underlying relationship between observations and the target variable.

In the next phase the validation set is used to evaluate the model fit of the previous phase, tuned on the model hyperparameters. Hyperparameters are parameters set before the learning process begins. They are external to the model as their value cannot change during the training phase. Examples of hyperparameters are number of branches in a Decision Tree or number of hidden layers in Neural Networks. In the validation phase the model is assessed using data not contained in the train set. In this work the hyperparameter optimization has been performed through a Grid Search, a grid of parameter values specified in the training phase, finding the optimal combination of values for the model, which results in the best prediction selected.

During the test phase, the model used rules previously generated to classify the observations belonging to the test set. A ground truth was available for such observations, but was not given to the model and was instead used to evaluate its performances and capabilities to generalize on data that was never seen before [53].

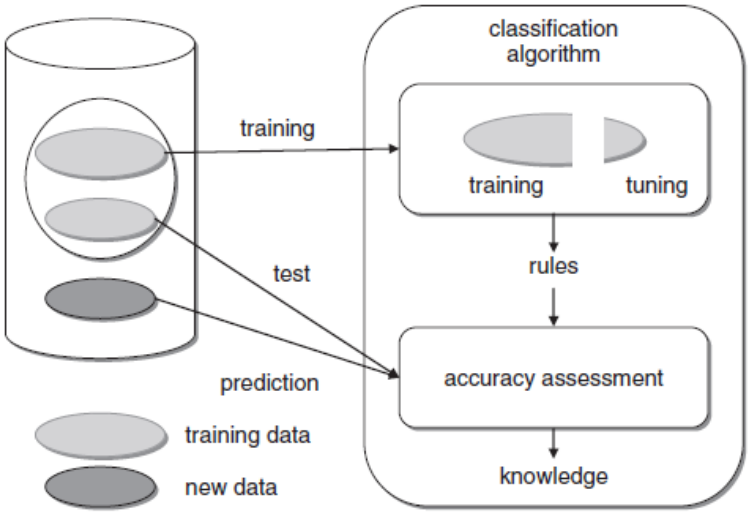


Figure 2.17: Overview of dataset division in supervised machine learning approaches. Data is divided into training and test set. Training set is then subsequently divided into training and tuning to define the optimal rules, or parameters to the method.

2.4.1. Data balancing

As already stated, splitting the data into training and testing set was done in a stratified way, ensuring that percentages of labels were kept identical in both training and testing set. As shown in Figure 2.10, the *Willingness being vaccinated_YN* attribute, defined as the target variable, caused the dataset to be strongly unbalanced, with 899 observations belonging to class 1, i.e. people who answered 'Yes', and 156 belonging to class 0, i.e. people who answered "No/I don't know".

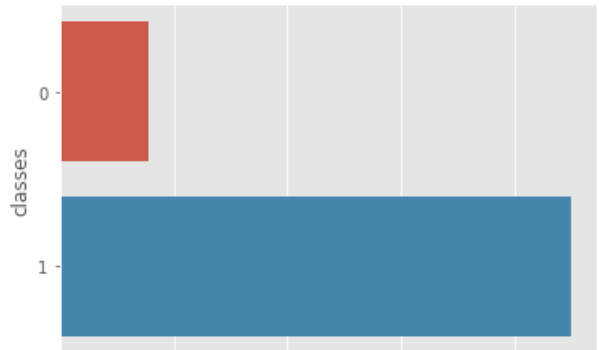


Figure 2.18: Distribution of the target variable "Willingness being vaccinated_YN" in the current dataset.

Exploration and investigation of literature on clinical data revealed that a classifier often shows a strong bias toward the majority class, being subject to error rates [54]. The relationship between set size and improper classification performance for imbalanced data set implies that, on small datasets, the minority class may be poorly represented by an excessively reduced number of examples that might not be sufficient for learning [55].

There known disadvantages suggest the use of sampling to balance the dataset and implement cost-sensitive learning. The disadvantage with undersampling is that it discards potentially useful data. The main disadvantage with oversampling is that, by making exact copies of existing examples, increases the possibilities of overfitting. In fact, with oversampling it is quite common for a learner to generate a classification rule to cover a single, replicated, example. Even though the disadvantages with sampling are to be taken into account, it is still a popular way to deal with imbalanced data rather than a cost-sensitive learning algorithm [56].

With downsampling, part of the observations belonging to class 1 were randomly chosen and eliminated to reach the same number of samples between classes. Thus, 743 elements of class 1 were eliminated reaching a total of 312 records of both classes combined. With oversampling part of the observations belonging to class 0 was replicated through the , Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC) algorithm [57], generating 743 new records until an equal number of observations was reached for each class. SMOTE-NC yields better results for re-sampling in probabilistic estimate techniques [58]. It creates artificial samples to increase the size of minority class, balancing the data by increasing instances. SMOTE-NC is a generalized approach of SMOTE, that works only with all numerical features [59]. A brief explanation of the two algorithms is described below [60].

For each minority sample, SMOTE works in three steps:

I. Its k-nearest neighbours belonging to the same class are found.
II. A number q of those neighbours, defined by the algorithm and depending on the amount of oversampling desired, is randomly selected.
III. Synthetic samples are randomly generated along the lines joining the minority sample and its q selected neighbours.

SMOTE-NC approaches with both numerical and categorical features:

I. Median computation of standard deviations of all numerical features; if nominal features differ between a sample and its potential nearest neighbours, this median is included in the Euclidean distance computation

II. Nearest neighbours computation: it computes the Euclidean distance between the identified k-nearests and the other minority samples. Foer each differing nominal variable between the feature vector and the potential nearest neighbour, it includes the median, computed in step I.
III. Population of the synthetic sample: numerical features of the new synthetic class sample are generated using SMOTE approach, previously described; for a nominal feature, value is given by the value occurring in the majority of the k-nearest neighbours.

The final dataset to give as input to the decision tree algorithm was composed by 1798 observations and 43 attributes.

2.4.2. Decision tree

Decision tree is a flow-chart-like tree structure, it is commonly used due to simple implementation and high explainability of the obtained results, even at the cost of some accuracy when compared to other supervised machine learning techniques such as multi-layer perceptrons [61].

Decision tree can be divided into binary and general trees, based on the number of descendants for each node. For this purpose, decision trees have been implemented as binary, since the considered target variable in the study is a binary variable [61]. The chosen classifier is based on CART, constructing binary trees using the feature and threshold that yield the largest information gain at each node, which will be detailed further in this section.

Classification with decision trees has two main steps. In the first step a tree is created, while in the second one classification rules are obtained from the tree structure.

1) In the initialization phase, each observation is placed in the root node S of the tree; the root is listed in the L group of active nodes.
2) The best attribute in the dataset is selected, using Attribute Selection Measure.
3) Node S is divided into subsets that contains possible values for the attributes.
4) A new decision tree node J is generated, containing the attribute selected.
5) The optimal rule is computed to split the records in J; the rule is applied and the descendant nodes are constructed by subdividing in two groups the observations contained in J. Each time, the conditions to stop the division are verified through stopping criteria, if these are met, node J becomes a leaf, to which a target class is assigned; otherwise the nodes are added to L. Then, the procedure is repeated, as shown in Figure 2.11.

It is necessary to deepen some aspects mentioned above [62].

- Attribute Selection Measures: They are techniques aimed to select the best attribute for nodes and to split its records. Numerical attributes are separated based on a threshold value; binary attributes are trivially divided according to whether they take on one value or the other. The two most known techniques are Information Gain and GINI Index. Information gain computes how a given attribute separates the training samples with respect to their target; the attribute with highest value is the best to be selected for split. It is computed as the difference between entropy before splitting and average entropy after splitting, on given attributes values.

$$IG = Entropy(S) - \sum_{j=1}^{m} \frac{N_c}{N_p} Entropy(C_j) \qquad (14)$$

In previous formula S is the node before the division, C_j is the j-child node, N_p is the total number of samples in S and N_c is the total number of j-child node.

GINI index is a measure of impurity; an attribute with low GINI index should be preferred as compared to the high index. The best split increases the purity of the sets resulting from the split. GINI is defined as

$$GINI(S) = 1 - \sum_{i=1}^{j} p_i^2 \qquad (15)$$

where p_i is relative frequency if class i in S.

- Stopping criteria: They are criteria to apply at each node, deciding whether the development should be continued or not. If not, the node becomes a leaf.
- Pruning criteria: applying pruning criteria means trim off the unnecessary decision nodes, to get the optimal decision tree. A too large tree would increase the risk of overfitting, at the same time a small tree may not capture relevant features.

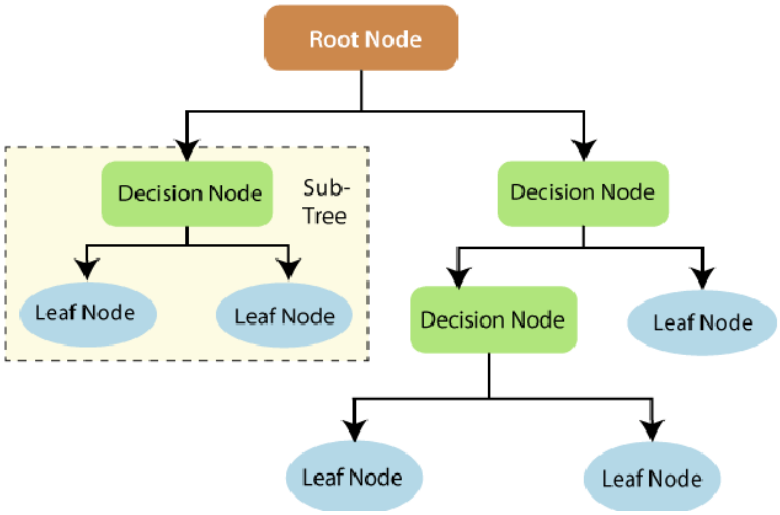


Figure 2.19: Example of splitting nodes procedure in a decision tree algorithm.

2.4.3. K Nearest Neighbor

K Nearest Neighbor (KNN) is a versatile machine learning algorithm for classification. It is also defined as a semi-supervised learning algorithm such that it requires training data and a predefined k value to fine the k nearest data based on distance computation [63].

It assumes that similar things exist in close proximity, thus that similar things are near to each other. KNN captures the idea of similarity calculating the distance between points on a graph [64]. A description of main steps follows.

1) First, it is now necessary to define a k value, representing the minimum number of near neighbors to establish proximity.
2) When a new unlabeled data point is encountered, it has to be calculated the distance with all training data points. Among all possibilities, the Minkowsky distance is the one applied to the study. To explain more in detail the computation, other two distance metrics are introduced, Euclidean and Manhattan ones.

   Euclidean distance measures a straight line between two points,

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}. \qquad (16)$$

   Manhattan distance between two points is the sum of the absolute differences of their Cartesian coordinates, in a 2-dimensional space,

$$d(x,y) = \left(\sum_{i=1}^{m} |x_i - y_i|\right). \qquad (17)$$

Minkowski distance is the generalized form of Euclidean and Manhattan metrics in the space.

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}, \qquad (18)$$

where, with p=1 Manhattan distance is obtained; when p=2 Euclidean distance is obtained.

Some assumptions have to be satisfied to compute this distance:

1. Non-negativity: $d(x,y) >= 0$
2. Identity: $d(x,y) = 0$ if and only if $x == y$
3. Symmetry: $d(x,y) = d(y,x)$
4. Triangle Inequality: $d(x,y) + d(y,z) >= d(x,z)$

3) The k-nearest neighbors are found out, so that the unlabeled point has to be assigned to the label with maximum number of nearest neighbors, as represented in Figure 2.12.

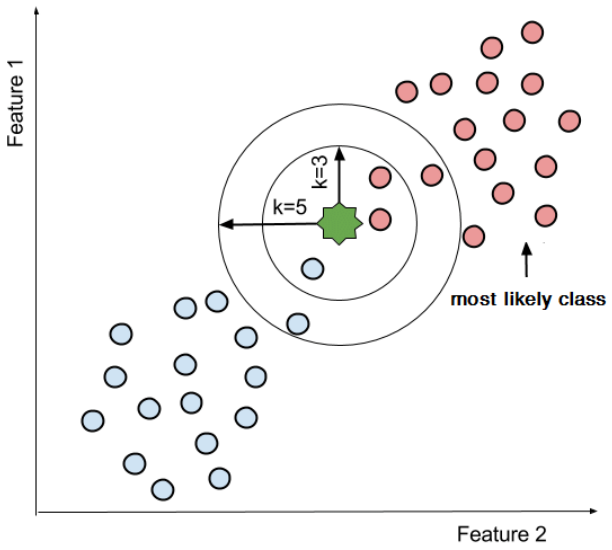


Figure 2.20: Attribution of a data input to a class, based on the k nearest neighbours

2.4.4. Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm available for classification. It is based on the principle of structural risk minimization, that is, minimizing the discrepancy between the target variable y to a given input x and the response provided by the algorithm [65]. It is very helpful in text categorization and it is known for its capability to generalize on small datasets without overfitting [66]. However, it presents the limit of being able to classify only binary variables, with multiclass classification requiring the usage of multiple SVMs to be performed and competitively compared to each other.

SVM is a two dimensional description of the optimal surface evolved from the linearly separable case [67]. Two sets of points belonging to binary target classes are said to be separable if there exists a hyperplane capable of separating them in the space R^n, reducing to a line in the two-dimensional case. A hyperplane in a n-dimensional Euclidean space is a flat, n-1 dimensional subset that divides space into two separated parts. The best separation has the largest distance to the nearest training-data point of any class.

Given a set of pair points $P = \{(x_1 y_1), (x_2 y_2), \dots, (x_m y_m)\}$, with x as input vector and y labels vector, it has to be constructed a classifier function f that maps the x into labels y. The goal is to find f which correctly classifies new points, so that $f(x) = y$. As literature confirms [68], the main choice to be done is the kernel-parameter, that defines the structure of the high dimensional feature space, where a separating hyperplane has to be found.

There are two issues to explore:

1. There is not a unique solution to separate points, it is therefore a ill-posed problem.
2. Data might not be linearly separable

The algorithm faces up the ill-posed problem finding the hyperplane that reaches the maximum possible margin of separation between classes.

To solve the non-linearly separable data problem, SVM maps the training data into a higher dimensional feature space, mapping a non-linear classifier function, applying a kernel function. Among most popular kernels there are:

Linear kernel: $$f(x) = \langle w, x_i \rangle + b, \quad (19)$$

where $\langle w, x_i \rangle$ is the dot-product of the weight vector w and the input sample, and b is the linear coefficient estimated from the training data.

Polynomial kernel: $$f(x,z) = (\langle x,z\rangle)^d \quad (20)$$

Gaussian RBF kernel: $$f(x,z) = exp(-\gamma * (-\left||x-z|\right|^2)) \quad (21)$$

Sigmoid kernel: $$f(x,z) = tanh(\alpha * \langle x,z\rangle - C), \quad (22)$$

where α is a weight vector and C is an offset value.

In the Figure 2.13 is shown a 2-dimensional representation of possible kernel applied to an example. Trying to visualize the separation has to be considered a simplification of the concept.

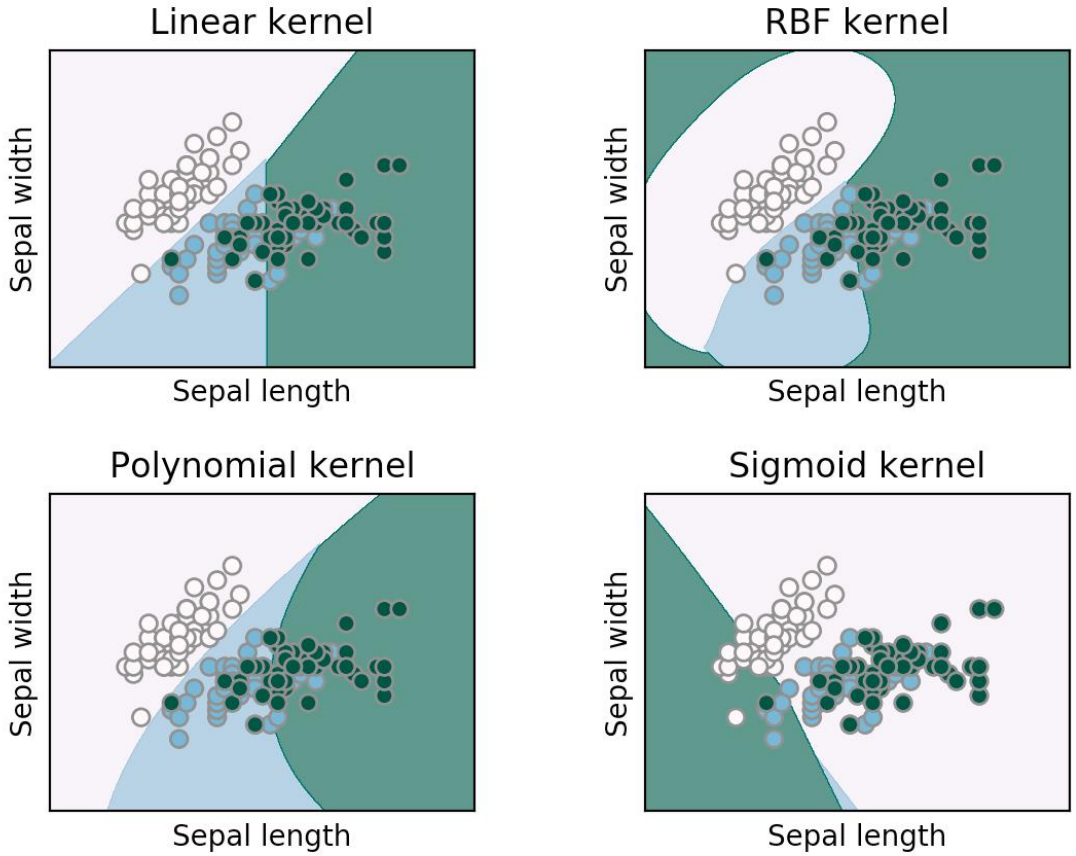


Figure 2.21: Kernel schemtization separation

2.4.5. Neural Network: Multi Layer Perceptron

Multilayer Perceptron is a feedforward artificial neural network [69]. A neural network is a series of algorithm aimed at recognizing underlying relationships in a set of data through a process that mimics the way the human brain operates. It is a model based on biological neural networks, in other words, it is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information [70], [71].

MLP consists of neurons arranged in layers. In Figure 2.14 an example of a MLP model is presented. This model consists of an input layer, an output layer, and two hidden layers. At least three layers are required to define a MLP model, one input, one output and one or more hidden layers. The input layer consists of a set of neurons equal, in number, to the features of the dataset in input. The output layer consists of one or more neurons, the values of which are the probabilities of the class predicted by the neural network. The hidden layers consist of a set of neurons that are connected to neurons in the previous and following layers. If each neuron is connected to each other neuron in previous and following layers, the model is defined as fully connected [72]. MLP is known to provide high accuracy models when compared other methods such as decision trees [73]. However, it is a black-box method, thus giving no explainability on how the result was obtained.

The main steps are outlined below.

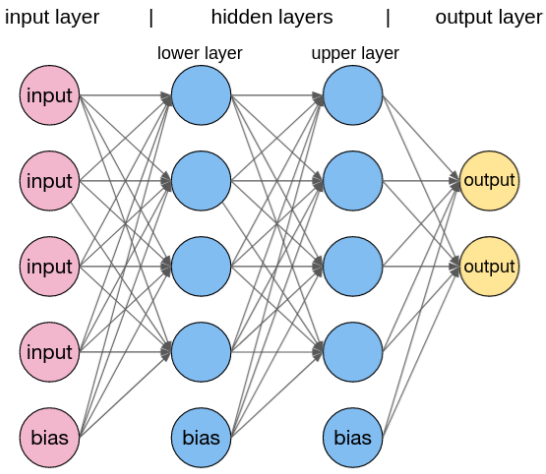


Figure 2.22: Example of a Neural Network representation with 2 hidden layers

1) MLP computes a single output from multiple real-valued inputs by forming a linear combination according to its input weight, and then putting the output through an activation function, that can be written as

$$y = f(x) = B\varphi(Ax + a) + b \quad (23)$$

Where: x is a vector of inputs and y a vector of outputs, A and a are respectively the matrix of weights and the bias vector of the first layer. B and b are the weight matrix and the bias vector of the secondo layer. Function φ denotes the non-linearity.

2) As a feedforward network, MLP works in a constant back and forth through two steps, a forward and a backward pass.
   In the forward pass the algorithm moves from the input layer through hidden layers and the output layer, measuring the output against the ground truth labels.
3) Whereupon, in the backward pass, through the activation function, the algorithm is back-propagated through the layers.

The whole process is iterated until the weights have converged.

2.4.6. Evaluation of classification model

To evaluate the classification model, scores have been computed and evaluated.

2.4.6.1. F1 Score

First of all is important to take into consideration that the results have to be as much similar between train and test set as possible. Thus, to evaluate the correct alignment between train and test, F1 score was computed. F1 takes into account not only the number of prediction error that the model makes, but also the type of error committed. F1 is defined as the harmonic mean of precision and recall, computed as follows:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (24)$$

Where:

$$Precision = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Positives}, \quad (25)$$

$$Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Negatives}. \quad (26)$$

A high F1 score will be obtained if both precision and recall are high; on the contrary it will be low if both are low.

2.4.6.2. ROC curve and AUC

A further metric that has been computed is the ROC curve (Receiver Operating chara cteristic Curve), a probability curve showing the performance of a classification mode l [74]. It plots two parameters: True Positive Rate and False Positive Rate.

True Positive Rate is a synonym of recall and is therefor defined as follows:

$$TPR = \frac{TP}{TP + FN}; \tag{27}$$

False Positive Rate is defined as:

$$FPR = \frac{FP}{FP + TN}. \tag{28}$$

The ROC curve plots TPR versus FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both FP and TP. To evaluate the curve, the Area Under the roc Curve (AUC) has to be computed. AUC measures the area underneath the entire ROC curve, from (0,0) to (1,1) coordinates on the cartesian plan, better shown in Figure 2.15. It ranges in value from 0 to 1. A value of 1 concerns an ideal condition, where all observations have been classified correctly, as true negatives or true positives, thus eliminating from the formula the FN and FP. The worst situation occurs when AUC is approximately 0.5 (bisector), hence the model has no discrimination capacity to distinguish between positive class and negative class. Instead, the closer the area is to 1, the more acceptable the value. As a generic standard rule, an AUC above 0.85 means high classification accuracy, one between 0.75 and 0.85 moderate accuracy, and one less than 0.75 low accuracy [75].

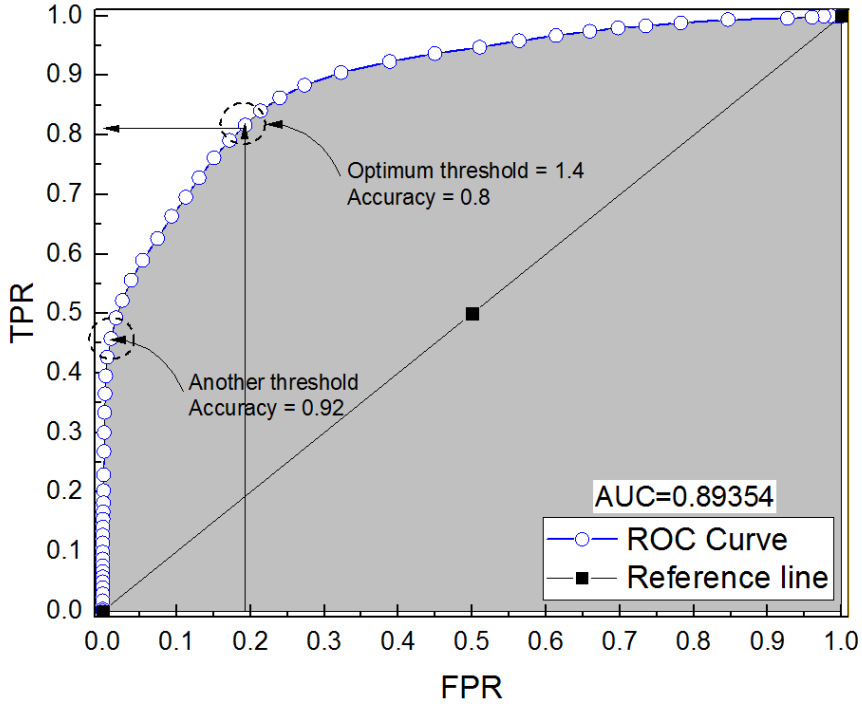


Figure 2.23: ROC curve example with AUC represented in grey.

In the Figure 2.16 shown below is reported an example of the curve, comparing two different algorithm performances. As explained, the orange one is better performing than the green

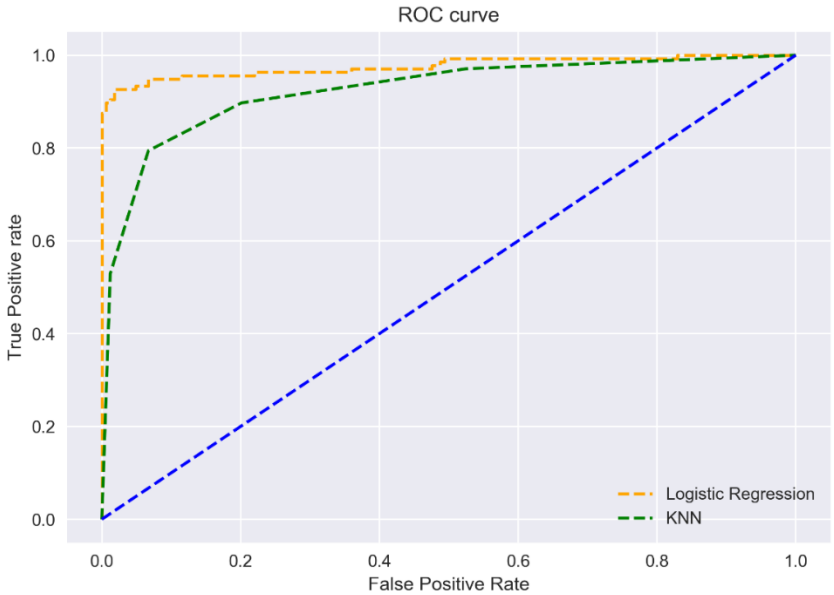


Figure 2.24: Comparison of two different ROC curves.

3 Results

In this chapter the results of the applied method described in the previous chapter will be presented. The analysed population will be described, and the results obtained by clustering and supervised machine learning methods will be presented separately. Finally, Persona tables will be presented for the methods that gave the best results.

3.1 Population

The population in exam consisted of 1089 respondents that answered the online survey between the end of 2020 and the beginning of 2021. Following the preprocessing methods previously explained, 16 (1.4%) respondents were omitted from further analysis due to not giving consent to the online survey, while 18 (1.6%) respondents were removed during missing data analysis, resulting in 1055 completed surveys.

The target variable "Willingness being vaccinated" is distributed with a 85% (899) of respondent who answered "yes", versus 15% (156) that answered "no".

.3.1.1. Survey structure

Respondents were required to answer to a total of 55 questions, resulting in a dataset of 138 columns. Of these columns, 7 where not included in further analysis as they were deemed as not giving useful information::

- start and end times,
- progressive number,
- test duration,
- administration channel,
- unique identification code,
- consent for participation in the study.

Thus, the remaining 131 columns corresponded to the answers to each questions after the preprocessing steps previously presented. The final dataset obtained after data cleaning was thus composed by 1055 records and 43 attributes.

3.1.2. Sex and Age

The population that participated to the survey was composed of 67% (711) women and 33% (344) men, intended as biological gender. Among women, the median age is 42 years, where 25th percentile is at 30 years and 75th is at 54 years. For men, median age is 46 years old, with a 25th percentile at 33 years, while the 75th percentile is at 59.

No data points have been considered as outliers, since none of them where located out of the defined thresholds. Specifically, the upper threshold is defined as

$$q_3 + w * (q_3 - q_1)$$

and the lower threshold is defined as

$$q_3 - w * (q_3 - q_1)$$

Where q1 and q3 are respectively the 25th and 75th percentiles. To identify outliers, a box-and-whisker plot has been created (Figure 3.1), in which median, lower and upper quartiles are represented on the axis. The whiskers are the two marks corrsponding to the minimum and maximum values of the attribute falling inside the two thresholds.

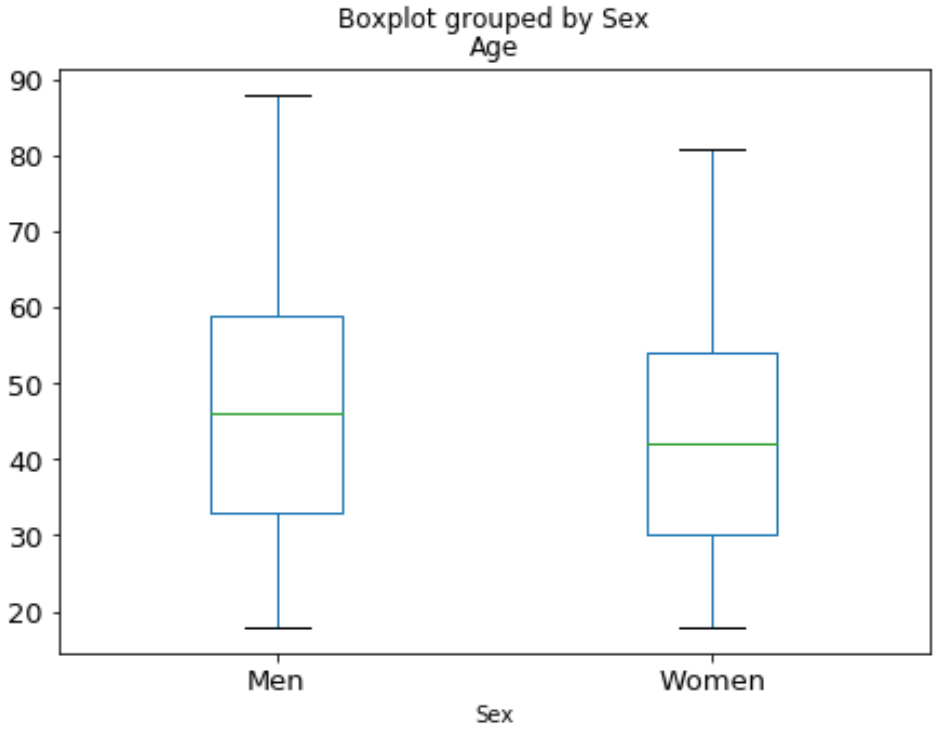


Figure 3.1: box-whisker plot of distribution of the Age, grouped by Sex, in the entire dataset. Men on the left, women on the right.

3.1.3. Education level and Profession

The education level of the participants (Figure 3.2) is broken down as follows: almost 0% (2) of the population have just the Elementary school certificate, 5% (49) have the

Middle school certificate and the 32% (339) have High school diploma. The remaining 63% are divided into 41% (433) with a Degree and 22% (232) with a PhD diploma.

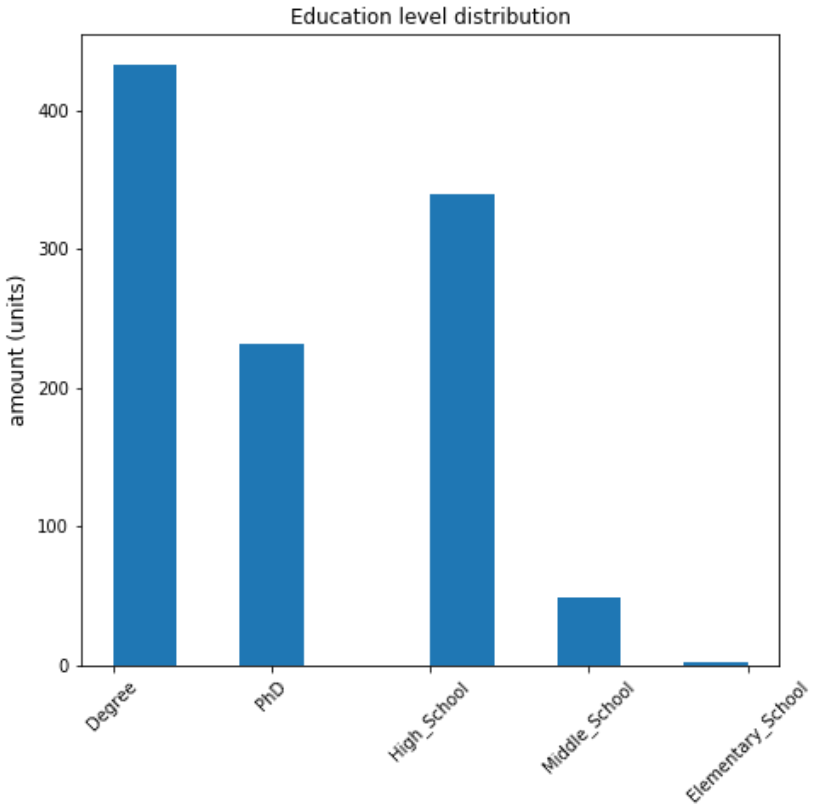


Figure 3.2: Histogram representing the education level distribution on whole dataset

About the professions' distribution (Figure 3.3), a total of 71% (746) have a job, versus 6% (64) that are unemployed. The day workers are the 2% (22), the student are 10% (104) and the pensioner the remaining 11% (119).

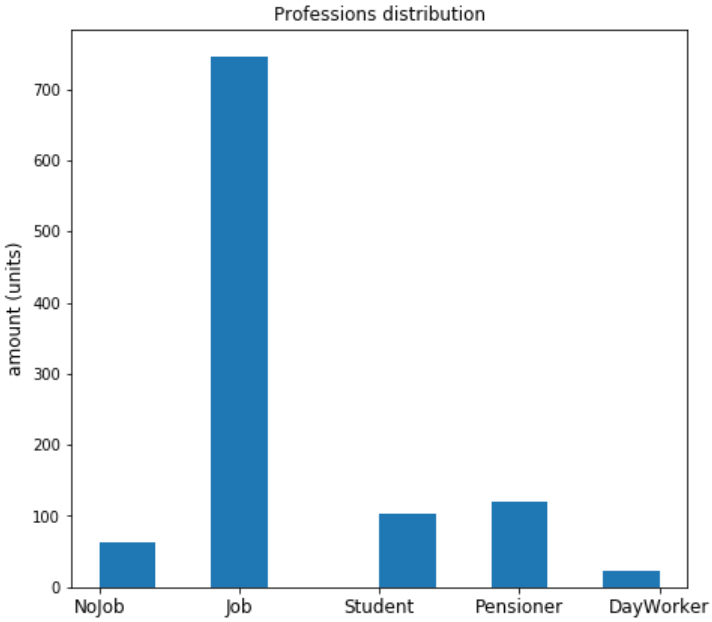


Figure 3.3: Histogram representing the occupation level distribution on the whole dataset

3.2. Clustering evaluation

As previously explained, the algorithm was applied in parallel to two databases: one complete, and one without what was considered the target variable, 'Willingness being vaccinated'.

For each database, the average silhouette method, used to evaluate the optimal number of clusters, was calculated for a range of clusters from 2 to 10. The optimal number of cluster was identified as the point with the highest value in the silhouette plot. In Figure 3.4, the average silhouette plot for the dataset without the target variable "Willingness being vaccinated" is shown. As can be seen, the highest value of the average silhouette was identified for a number of clusters equal to 5, using the PCA preprocessing method combined with K-Medoids clustering.

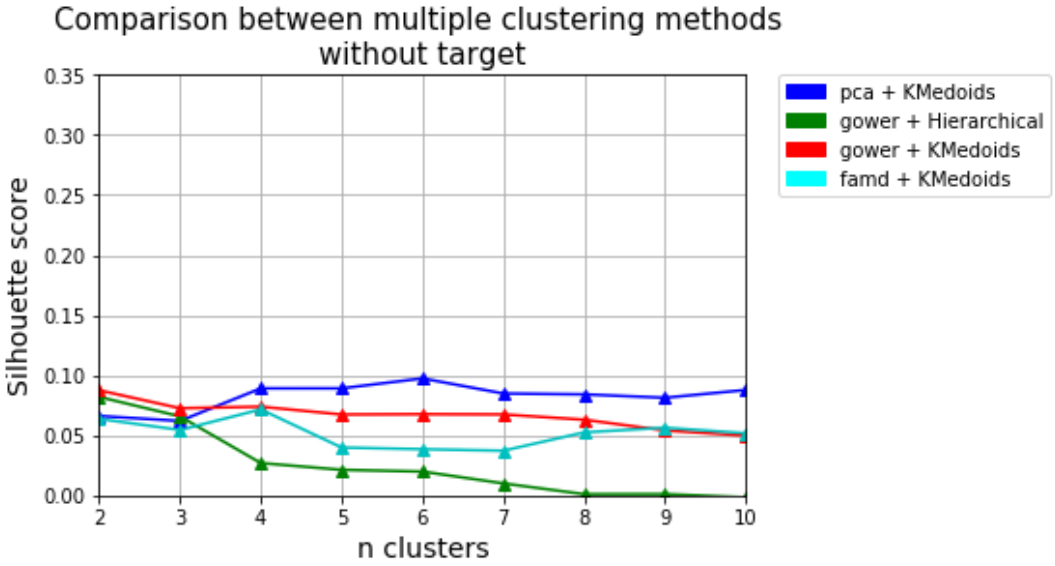


Figure 3.4: Average silhouette scores comparison, plotted over a range of 2 to 10 clusters, in the dataset without target variable

In Figure 3.5, the average silhouette plot for the dataset with the target variable is shown. In this plot the highest average silhouette value was achieved using Gower Distance matrix in Hierarchical Clustering, with a number of clusters equal to 2..

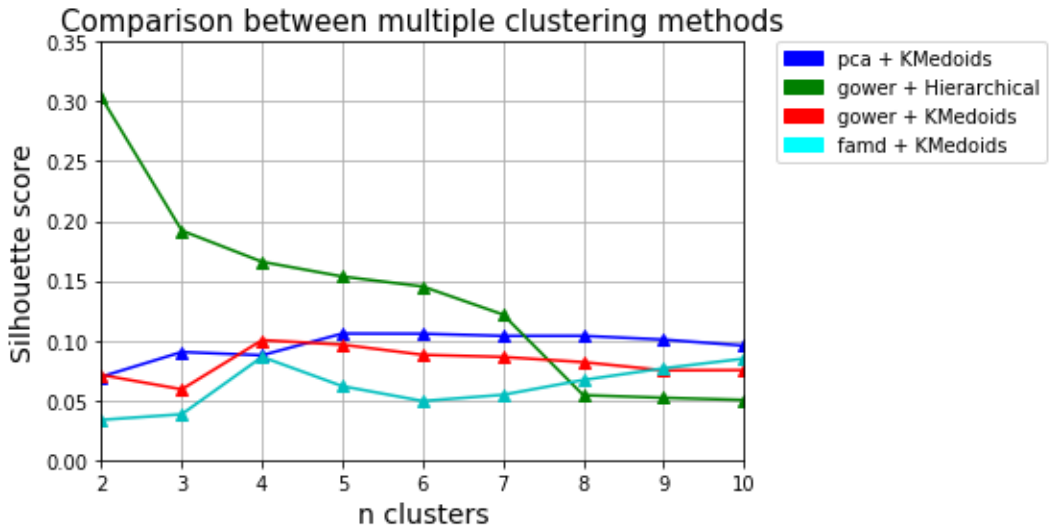


Figure 3.5: Average silhouette scores comparison, plotted over a range of 2 to 10 clusters, in the dataset with target variable

As shown in the plots of the average silhouette values between the two different datasets, the best performance is given by the Hierarchical Clustering, applied on Gower distance matrix, on the complete dataset.

3.2.1. Clustering dataset with target variable

Figure 3.6 shows the results of the Hierarchical Clustering through the dendrogram. The dendrogram allows to identify the distances between each group of clusters. By drawing a horizontal line cutting the branches at a set distance, the number of clusters that are obtained is given by the intersection between the red line and the dendrogram. Following the connecting lines from the intersection point to each element of the dataset it is possible to identify the points belonging to each cluster.

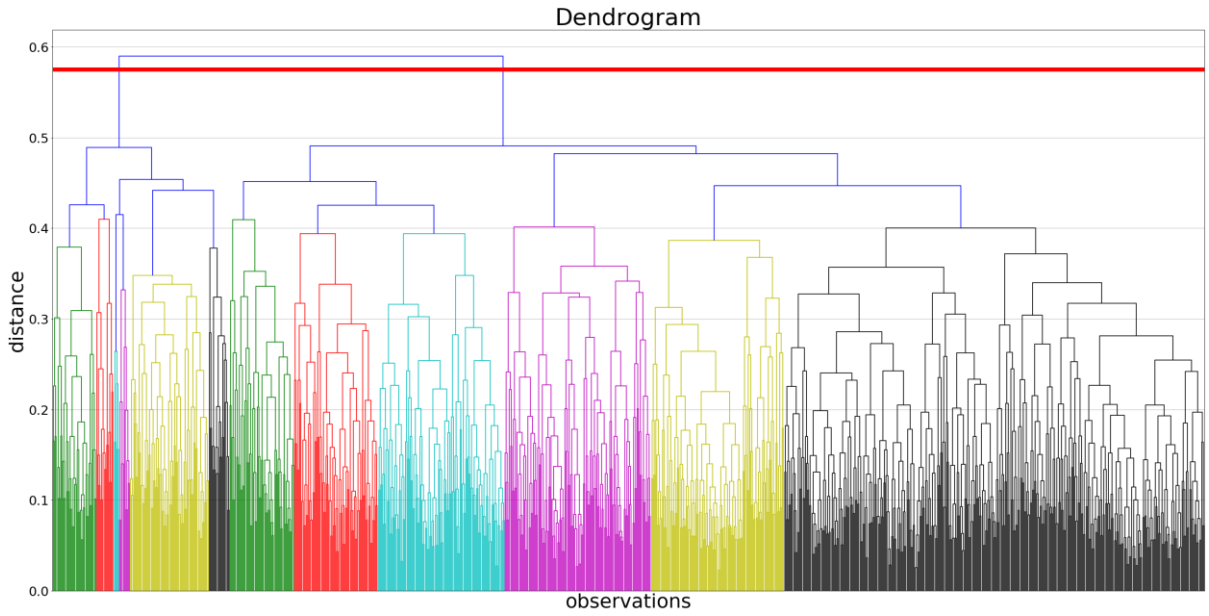


Figure 3.6: Dendrogram showing hierarchical clustering results on complete dataset

Two clusters, named Cluster 1 and Cluster 2, were thus identified as the optimal division. The Age of the population belonging to cluster 1 is distributed as follows: 76% (123) are women with a median age of 47 years, 25th percentile of 32 and 75th percentile of 54. Men are the 24% (39), with a median age of 52 and 25th and 75th percentile of, respectively, 40.5 and 55.5.

In cluster 2 there are 66% (588) of women, with 42 years as median age, and 30, 54 as 25th and 75th percentile. The remaining 34% (305) of men have a median age of 44 years, with 32, 59 as 25th and 75th percentile.

In Table 3.1 results of the comparative statistical analysis have been reported. For numerical variables the median, together with 25th and 75th percentiles, are reported. The corresponding p-values were computed through the usage of Mann Whitney U test. For nominal categorical and binary variables, the mode value and relative percentage are reported. In this case, p-values were computed through Fisher's exact test. Only attributes with resulting p-values lower than 0.05, thus expressing significant difference between the two clusters, are reported in Table 3.1. Out of 43 analysed attributes, only 21 (49%) were found as statistically significantly different.

Table 3.1: Cluster's characteristics: the value with the largest percentage is reported for each categorical attribute; the mean, 25th and 75th percentile are reported for numerical attributes.

	CLUSTER 1 (162)	**CLUSTER 2** (893)	**p value**
Sex	76% women	66% women	0.014
Education	44% high school	42% degree	<0.001
Job	73% job	70% job	0.009
Sanitary job	76% no	66% no	<0.001
Physical status	good	good	0.010
Psychological status	good	good	0.015
Covid severity self	not contracted	not contracted	0.038
Contracting covid probability self	5 (3.5; 6)	5 (4; 7)	0.004
Optimistic bias	0 (0;1)	0 (0;1)	0.032
Covid health damage	3 (3; 4)	4 (3; 4)	0.017
Covid more severe flu	4 (4; 4)	4 (4; 5)	<0.001
Other disease no hospital due to Covid fear	3 (2; 4)	2 (2; 3)	<0.001
Fear of Covid	11 (9; 13)	12 (10; 13)	<0.001
Institutions trust	10 (8; 11)	12 (10; 13)	<0.001
Do you recommend Covid vaccine others	80% no	96% yes	<0.001
Prevention behaviour after Covid vaccine	95% yes	98% yes	0.016
Vaccine opinion others	87% yes	66% yes	<0.001
After 1 year, better sanitary situation after vaccine	67% no	87% yes	<0.001
Vaccination 19/20 nonCovid	90% no	55% no	<0.001
MHLCS total score	14 (11; 18)	13 (10; 16)	<0.001
Willingness being vaccinated	92% no	99% yes	<0.001

In Table 3.2, *t*he level of education of the subjects is presented in greater detail. 66% (585) of the population of cluster 2 has a level of education corresponding to a degree or PhD, compared to 49% (80) of Cluster 1 with the same level of education. In Cluster 1 the mode is high school diploma, with 44% (71), while in Cluster 2 the mode is Degree, with 42% (373). Cluster 1 has thus a statistically different lower level of education when compared to Cluster 2. All participants had at least a Middle school diploma, with nobody in both cluster having only an Elementary school diploma.

Table 3.2: Education level distribution

	CLUSTER 1	**CLUSTER 2**
PhD	12%	24%
Degree	37%	42%
High school	44%	30%
Middle school	7%	4%
Elementary school	0%	0%

The distribution of workers, unemployed, pensioners and students is shown in Table 3.3.

Table 3.3: Occupancy distribution

	CLUSTER 1	**CLUSTER 2**
Dayworker	2%	2%
Job	73.5%	70.5%
No job	11%	5%
Pensioner	6%	12.5%
Student	7.5%	10%

The previous table shows a clear disparity between the unemployed in cluster 1, 11%, and in cluster 2, only 5%. On the contrary, in cluster 1 there are 6% of pensioners, unlike 12.5% in cluster 2. In both clusters, more than 70% of respondents have a job.

As regards the state of physical and psychological health, it does not appear to be significantly different. In both clusters the "good" response for the physical state prevails; with regard to the psychological state of health, both clusters include subjects who have a response for the most part "sufficient" and "good".

"Covid severity self" corresponds to the question "If you have contracted the covid, in what form would you say you have contracted it?", Where the answer 0 in both clusters means that most of the subjects have not contracted Covid at the date of compilation of the survey.

The members of cluster 1 prove to have a higher Fear of Covid and a lower Institution trust than the ones in cluster 2.“Vaccine opinion others” is related to the question “As far as you know, in your closest acquaintances (e.g. friends, partners, family) are there people who have different opinions and intentions from yours regarding the vaccine?”, and in cluster 1, 87% affirm yes, while just 66% of cluster 1 do the same.

Those who answered no to " After 1 year sanitary situation after vaccine " are 67% for cluster 1, stating that one year after vaccination, there will be no health improvement. On the contrary, 87% of the population of the cluster 2 answered yes.

Almost all of cluster 1 declares that they have not carried out vaccinations for diseases other than Covid in the previous two years, unlike cluster 2 where half claim to have carried them out.

The MHLC scale, that measures physical and mental health functioning as previously explained in chapter 2, reports not very different values between clusters, but moderately higher in cluster 1.

As hypothesized, the major discriminant that led to the division into the two clusters is the will to want to vaccinate against Covid-19 or not, as can be seen from the percentages tending to 100%, opposite between the two clusters. Likewise, many of the subjects belonging to cluster 1 would not recommend vaccination to other people. On the contrary, 96% of the subjects of cluster 2 would recommend it.

3.2.2. Clustering dataset without target variable

The better clustering result obtained in the dataset without target variable is given by the K-medoids algorithm applied on the dataset reduced through the PCA. With an average silhouette score equal to 0.10, 6 cluster have been obtained.

The 21 (49%) attributes shown in Table 3.4 presented a significant p value, lower than 0.05, as explained in Chapter 2. Being 6 the clusters founded by the algorithm, for numerical attributes Kruskal-Wallis test has been performed, and median, 25th percentile and 75th percentile have been reported. For categorical attributes, Fisher exact test has been computed, and the value that have been reported in the table for each one, corresponds to the median related to the cluster. Moreover, next to each value are reported the clusters, in brackets, that have statistical significance with that value. The attribute related to the biological sex of the population turned out to be statistically not significant, but it is however shown in Table 3.4 since it is a relevant characteristic for the development of the identity of Personas.

	CLUSTER 1 (114)	**CLUSTER 2 (313)**	**CLUSTER 3 (168)**	**CLUSTER 4 (124)**	**CLUSTER 5 (200)**	**CLUSTER 6 (136)**	**p value**
Age	28 (25,30) [C2,C3,C4, C5,C6]	47(39,55) [C1,C4,C5, C6]	49 (38, 55) [C1,C4,C6]	24 (22, 27) [C1,C2,C5, C6]	42 (35, 50) [C1,C2,C4, C6]	65 (60.7, 70.3) [C1,C2,C3, C4,C5]	<0.001
Sex	Women (76%)	women (64%)	women (73%)	women (71%)	women (69%)	women (54%)	<0.001
Physical status	4 (4, 5) [C2,C3,C4, C5,C6]	4 (4, 4) [C1,C6]	4 (4, 4) [C1]	4 (4, 4) [C1,C6]	4 (4, 4) [C1,C6]	4 (4, 4) [C1,C2,C4, C5]	<0.001
Psychological status	4 (4, 4) [C4]	4 (3, 4) [C4]	4 (3, 4) [C4]	3 (3, 4) [C1,C2,C3, C5,C6]	4 (4, 4) [C4]	4 (4, 4) [C4]	<0.001
Covid severity significant, others	2 (0, 3) [C2,C6]	0 (0, 3) [C1,C4,C5]	2 (0, 3)	2 (0, 3) [C2,C6]	2 (0, 4) [C2,C6]	0 (0, 3) [C1,C4,C5]	<0.001
Contracting covid probability, self	5 (4, 7) [C6]	5 (4, 6) [C6]	5 (4, 7) [C6]	5 (4, 7) [C6]	5 (4, 7) [C6]	4 (3, 5) [C1,C2,C3, C4,C5]	<0.001
Covid health damage	2 (3, 4) [C2,C3,C5, C6]	3 (3, 4) [C1,C6]	3 (3, 4) [C1,C5,C6]	3 (2, 4) [C5,C6]	4 (3, 4) [C1,C3,C4, C6]	4 (3.7, 4) [C1,C2,C3, C4,C5]	<0.001
Other disease no hospital	2 (1, 3) [C2,C3,C6]	2 (2, 3) [C1]	3 (2, 3.2) [C1,C4,C5]	2 (1, 3) [C3]	2 (2, 3) [C3,C6]	2 (2, 3) [C1,C5]	<0.001

Fear of covid	11 (9, 13) [C4]	12 (10, 13) [C4]	11 (9, 13) [C4]	12 (11, 13) [C1,C2,C3, C6]	12 (10, 13) [C6]	12 (9, 12) [C4,C5]	<0.001
Instituti on trust	12 (10, 12.7) [C3,C4]	12 (10, 13) [C3,C4,C5]	10 (8, 11) [C1,C2,C4, C5,C6]	12 (11, 13) [C1,C2,C3]	12 (10.7, 13) [C2,C3]	12 (10, 13) [C3]	<0.001
Gad7 total score	5 (3, 7.7) [C4,C6]	5 (3, 8) [C4,C6]	6 (3, 9) [C4,C6]]	9 (6, 14) [C1,C2,C3, C5,C6]	5 (3, 7) [C4,C6]	4 (2, 6) [C1,C2,C3, C4,C5]	<0.001
Mhlcs total score	14 (11, 17) [C5,C6]	13 (10,16) [C3]	15 (11, 19) [C2,C5,C6]	13 (10.7, 17) [????	12 (10, 15) [C1,C3]	12 (9.7, 15) [C1,C3]	<0.001
Marital status	engaged [C2, C3, C4,C5,C6]	married [C1, C3,C4,C5,C 6]	married [C1, C2,C4,C6]	single [C1,C2,C3, C5,C6]	married [C1,C2,C4, C6]	married [C1,C2,C3, C4,C5]	<0.001
Educatio n level	degree [C2,C3,C4, C5,C6]	degree [C1,C3,C4, C5,C6]	high school [C1,C2,C4, C5]	degree [C1,C2,C3, C5,C6]	PhD [C1,C2,C3, C4,C6]	high school [C1,C2,C4, C5]	<0.001
Job	job [C2,C3,C4, C5,C6]	[C1,C4,C6]	[C1,C4,C6]	student[C1, C2,C3,C5,C 6]	[C1,C3,C4, C6]	pensioner [C1,C2,C3, C4,C5]	<0.001
Sanitary job	yes [C2,C3,C4, C6]	no [C1,C5,C6]	no [C1,C5]	no [C1,C5]	yes [C2,C3,C4, C6]	no [C1,C2,C5]	<0.001
Do you recomme nd	yes [C3]	yes [C3,C5,C6]	no [C1, C2,C4,C5]	yes [C5,C6]	yes [C2,C3,C4]	yes [C2,C3,C4]	<0.001
Vaccine opinion others	5 [C6]	5 [C3,C6]	5 [C2,C5,C6]	5 [C6]	5 [C3,C6]	1 [C1,C2,C3, C4,C5]	<0.001

After 1 year	yes [C3]	yes [C3]	no [C1,C2,C4, C5,C6]	yes [C3]	yes [C3]	yes [C3]	<0.001
Vaccination 19/20 noncovid	yes [C5,C6]	yes [C5,C6]	yes [C5,C6]	yes [C5,C6]	no [C1,C2,C3, C4,C6]	no [C1,C2,C3, C4,C5]	<0.001
Willingness	yes [C3]	yes [C3,C5,C6]	no [C1,C2,C4, C5,C6]	yes [C3]	yes [C2,C3]	yes [C2,C3]	<0.001

Table 3.4: Cluster's characteristics: the median is reported for each categorical attribute; the mean, 25th percentile and 75th percentile are reported for numerical attributes. For all values are reported the list of clusters statistically s

As can be seen from the previous Table 3.4, among 6 clusters, only cluster 3 is distinguished from the others by the willingness of being vaccinated. In fact, with a percentage of 54% (91), it is the only one to have a population that for the most part has decided not to get vaccinated. Furthermore, since there are 6 clusters created, out of a total of 1055 records, the number of values assumed by the attributes is frequently a few units for each cluster. This can lead to the creation of bias due to an insufficient number of observations for such a large clustering, which had in fact detected an average silhouette value of 0.10, albeit the highest.

3.3. Classification evaluation

In deciding whether and how to apply a balancing of the dataset, for further study and further confirmation, the algorithms were applied for the first time to the unbalanced dataset. Figure 3.7 shows the corresponding ROC curves.

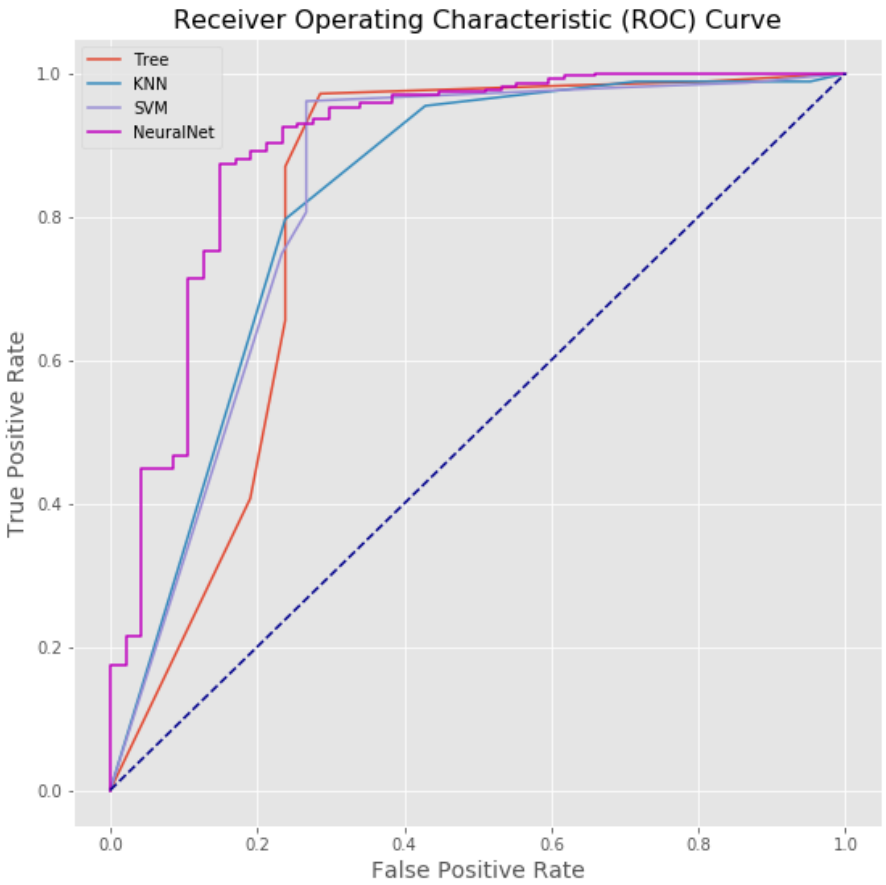


Figura 3.7: ROC curves of predictions in the unbalanced dataset; the bisector represents the reference line, the 4 curves refer to the 4 supervised model, Decision Tree, K Nearest Neighbours, Support Vector Machine and Multi layer Perceptron Neural Network.

Furthermore, Table 3.5 shows the results of the F scores and AUCs, which demonstrate a clear difference in the predictions of the target. In fact, where the target variable has far fewer records, i.e. target 0, the prediction is much less accurate.

Table 3.5: F1 scores of predictions in the unbalanced dataset

	F1 SCORE			
	Classification Tree	**KNN**	**SVM**	**MLP**
0	0.73	0.45	0.76	0.61
1	0.97	0.95	0.96	0.94
AUC	0.81	0.82	0.89	0.83

Table 3.6: Confusion matrices for each applied algorithm for classification. In green there are true negatives on the top left and true positives on the lower right; in red there are false positives on the top right and false negatives on the lower left.

Confusion Matrices of classification without data balancing							
Decision tree		K-NN		SVM		MLP	
15	6	7	14	22	8	15	15
5	172	3	174	6	149	4	151

As can be seen from Table 3.6, confusion matrices show the unbalanced prediction made by each model. To compare the results, Specificity and Precision have been computed. Specificity is the percentage of true negative values over all negatives including those not predicted correctly.

$$Sp = \frac{TN}{TN + FP} \tag{29}$$

Precision is the percentage of true positives over all positives.

$$Pr = \frac{TP}{TP + FP} \tag{29}$$

For all four models values of specificity not higher than 71% (classification tree) have been obtained. On the contrary, precision goes from 93% (classification tree) to 98% (K-NN).

Considering an imbalanced dataset to be ineffective, to apply the classification algorithms chosen it has been proceeded with the balancing of the dataset using SMOTE-NC, as described previously. At the end of the creation of each model, the

evaluation scores were calculated, and the ROC curve was created and the respective AUC have been computed.

Figure 3.8 shows the comparison graph of all the ROC curves associated with each model.

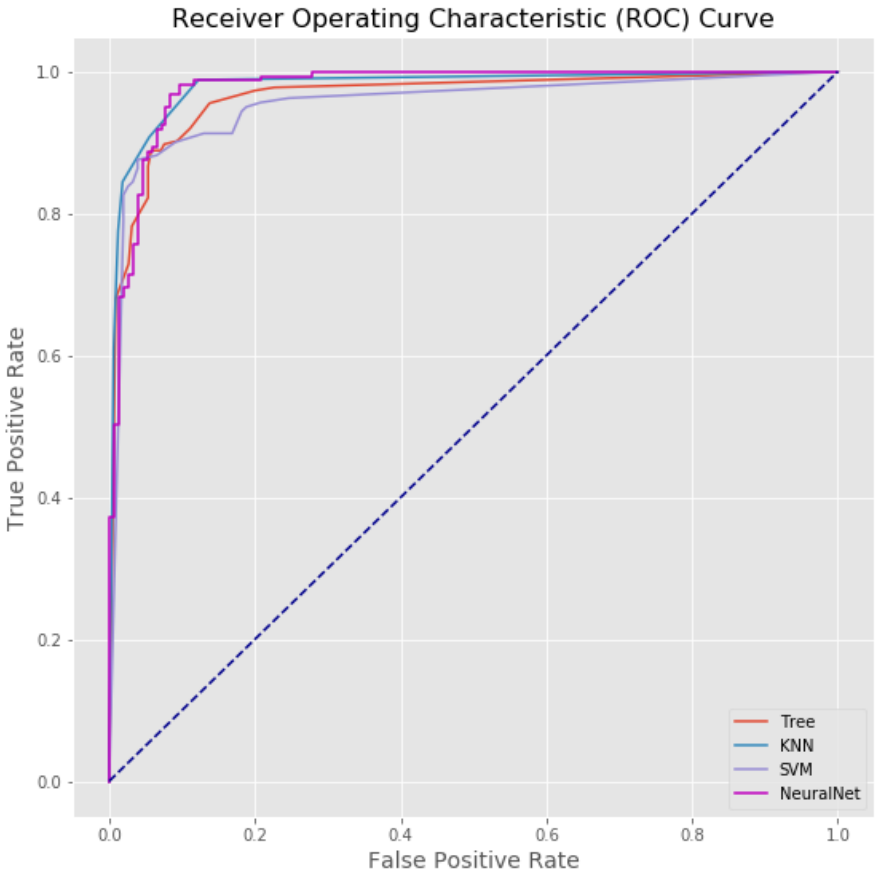


Figura 3.8: ROC curves of predictions in complete dataset

Furthermore, in Table 3.7 below, the F scores of each model are shown, comparing the score detected in the prediction of target 0, by the one detected in the prediction of target 1. Also, AUC values are reported.

Table 3.7: F1 scores and AUC of predictions in complete dataset

F1 SCORE

	Classification Tree	**KNN**	**SVM**	**MLP**
0	0.90	0.91	0.91	0.93
1	0.91	0.91	0.92	0.93
AUC	0.96	0.98	0.95	0.97

Table 3.8: Confusion matrices for each applied algorithm, on the balanced dataset. In green there are true negatives on the top left and true positives on the lower right; in red there are false positives on the top right and false negatives on the lower left.

Confusion Matrices of classification with data balancing

Decisione tree		K-NN		SVM		MLP	
152	11	160	3	156	7	158	5
21	153	27	147	15	159	15	159

On Table 3.8 confusion matrices results show true negatives values and true positives values better balanced, obtaining a specificity and a precision over 90% in all models.

Evaluation results show good values for all the Classification models, and MLP has been chosen due to best compromise between F1 scores and AUC value.

3.3.1. Classification dataset

Statistical analysis have been computed on the MLP predicted clusters, using test set. It is a fictitious dataset, as it is the result of the oversampling applied to the original dataset, as explained previously. The total records contained in the dataset are 450. Table 3.9 shows the 22 (51%) most significant attributes, valued according to the p value.

Table 3.9: Clusters characteristics, Classification model

	CLUSTER 1 (225)	**CLUSTER 2** (225)	**p value**
Age	45 (35; 53)	39 (29; 53)	0.005
Marital status	married (86%)	married (60%)	<0.001
Education	degree (61%)	degree (39%)	<0.001
Job	job (89%)	job (67%)	<0.001
Country	north (96%)	north (85%)	<0.001
Vaccine change idea	adverse reaction (55%)	health authority (68%)	<0.001
Sanitary job	no (90%)	no (70%)	<0.001
Physical status	4 (4; 4)	4 (4; 4)	0.047
Psychological status	4 (3; 4)	4 (3; 4)	0.033
Covid severity significant, others	1 (0; 2.5)	0 (2; 3)	0.017
Contracting Covid probability, self	5 (3; 5.5)	5 (4; 7)	<0.001
Covid health damage	3 (2.5; 4)	4 (3; 4)	<0.001
Covid more severe than flu	4 (3.5; 4)	4 (4; 5)	<0.001
Other disease no hospital for Covid fear	3 (2; 3.5)	2 (2; 3)	<0.001
Fear of Covid	11 (9; 12)	12 (10; 13)	<0.001
Institutions trust	10 (7; 11)	11.5 (10; 13)	<0.001
Do you recommend covid vaccine others?	no (75%)	yes (94%)	<0.001
Vaccine opinion others	yes (97%)	yes (65%)	<0.001
After 1 year, better sanitary situation after vaccine	no (68%)	yes (87%)	<0.001
Vaccination 19/20 nonCovid	no (99%)	no (54%)	<0.001
MHLCS total score	15 (12; 19)	13 (10; 17)	<0.001
Willingness being vaccinated	no (91.5%)	yes (91.5%)	<0.001

The population in cluster 1 is composed of 85.5% (188) of women, having a median age of 44, and 34 and 51 as 25th and 75th percentiles. 16.5% (37) are men of 51 as median age, and 43, 55 years as percentiles.

Cluster 2 have a population of 65% (146) of women with a median of 36 years, and 25th and 75th percentile respectively of 28 and 50 years. Men, 35% (79), have a median of 43 years, the 25th percentile of 31 and the 75th of 57 years.

In cluster 1, 86% (193) of the individuals are married, , 8% (19) are divorced, 5% (11) are engaged, 1% (2) are single, and none have divorced. In cluster 2, 60% (134) are married, 17.5% (41) are single, 16.5% (37) is engaged, 4% (9) are divorced, and 2% (4) are widow.

Table 3.10: Education level distribution among cluster 1, on the left, and cluster 2, on the right; it has been reported percentage and amount in units.

	CLUSTER 1	**CLUSTER 2**
PhD	4.5% (10)	27% (61)
Degree	61% (138)	39% (87)
High school	34% (76)	30% (68)
Middle school	0.5% (1)	4% (9)
Elementary school	0% (0)	0% (0)

The most evident difference in the education level, shown in Table 3.10, is given by the PhD level, which in cluster 1 is owned by 4.5% (10), while in cluster 2 it has been achieved for 27% (61) of the population. Therefore, the 65.5% (148) of people belonging to cluster 1 who continued their studies after high school, 93% (138) stopped at the bachelor's or master's degree. On the other hand, in cluster 2, of the 66% (148) of the population that went to university, 41% (61) obtained a PhD.

In Table 3.11 is reported the distribution of occupation of the population.

Table 3.11: Occupation level distribution among cluster 1, on the left, and cluster 2 on the right; it has been reported percentage and amount in units.

	CLUSTER 1	**CLUSTER 2**
Dayworker	7.5% (17)	2.5% (6)
Job	89% (200)	67.5% (152)
No job	1% (2)	7.5% (17)
Pensioner	2% (5)	8.5% (19)
Student	0.5% (1)	14% (31)

Cluster 1 results made up of 89% of a population with permanent or fixed term contract against cluster 2 with 67.5%. The dayworker are 7.5% in cluster 1 and just 2.5% in cluster 2. In cluster 1 almost no one is unemployed, precisely 1% (2), and in cluster 2 they are slightly more, 7.5% (17). Also pensioner and student are more prevalent in cluster 2 than cluster 1. In particular, students are 14% (31) in cluster 2 and just 1 in cluster 1.

“Vaccine change idea” derives from the question “Which of the following reasons could make you change your mind with respect to the answer given to the previous question?”. The previous question mentioned is the one related to the willingness of being vaccinated. Analysing answers of cluster 1, some deductions can be made. More than half of the population, 55% (124), answered “adverse reactions”. That is, for those who replied “no” to the willingness of being vaccinated, it means that they could change their idea, answering “yes” if adverse reactions decreased or, more plausibly, disappeared. At the same time, with regard to the cluster 1, individuals who replied “yes” to the willingness of being vaccinated, they could change their idea in “no” if they become aware of new adverse reactions.

In cluster 2 the most selected answer is “health authority”, which means that 68% (154) of the population would change their idea about the willingness of being vaccinated only if the health authorities gave new specific indications.

No significant differences can be appreciated between the clusters in the physical and psychological state declared by the participants, good in both cases.

"Contracting covid probability, self" refers to the individual’s perceived probability of contracting Covid-19. The population of cluster 2 has the 25th and 75th percentile higher than cluster 1. They both have median 5, but cluster 1 has 25th and 75th percentiles of 3 and 5.5. Cluster 2, on the other hand, has the 25th percentile equal to 4

and the 75th equal to 7. Cluster 2 has therefore a higher perception of the probability to contract the virus than the cluster 1.

Both "Fear of Covid" and "Institution trust" are slightly lower in cluster 1 than cluster 2. More precisely, it is observed that Fear of Covid in cluster 1 has a median equal to 11 against the median of cluster 2 equal to 12. Moreover the 25th percentile is 9 in cluster 1 and 10 in cluster 2; 75th in cluster 1 is 12, in cluster 2 is 13. A bigger difference is found in the Institution trust, with 25th percentile of 7 in cluster 1, and 10 in cluster 2. Same thing for the 75th percentile, which is equal to 11 in cluster 1, and equal to 13 in cluster 2. It follows that the population of cluster 2 is more afraid of the virus, but places more trust in institutions.

Cluster 1 states that 97% (218) of acquaintances have different opinions from their own regarding the vaccine, compared to 65% (146) of cluster 2.

Cluster 1 includes a population of which 68% (153) have no confidence in the improving of health situation after one year from vaccinations, unlike cluster 2, which 87% (195) believe will be better.

The MHLCS total score is higher in cluster 1 than in cluster 2. In fact the median is equal to 15 in cluster 1, with 25th and 75th percentiles of 12 and 19. Cluster 2 has median equal to 13, 25th percentile equal to 10 and 75th equal to 17.

The 99% (224) of cluster 1 did not have vaccinations between 2019 and 2020, 91.5% (206) respond that they will not vaccinate against Covid-19, and 75% (168) will not recommend it to others. On the contrary, 46% (104) of cluster 2 had vaccinations in the previous two years, 91.5% (206) would like to get vaccinated against Covid-19 and almost the entire population, 94% (211) will recommend it to others.

4 Conclusions and discussion

In this chapter results obtained in the study will be analysed and discussed. The three different clusterings obtained will be compared, studying their characteristics and main statistical differences.

4.1. Difference between clustering characteristics

From the statistical analyses, three clusterings were obtained and, for simplicity, they will be called Clustering A, B and C. Clustering A is related to the complete dataset, on which Hierarchical Clustering has been applied, and from which 2 clusters have been obtained. Clustering B was instead done on the dataset eliminating the target variable "Willingness being vaccinated", on which PCA and K-medoids clustering were applied, and from which 6 clusters were obtained. Finally, clustering C derives from Classification, a supervised machine learning technique, which created 2 clusters. Since the latter is a prediction algorithm using a target variable, the statistical analysis was performed on the subdivision of the dataset obtained from the classification model, as explained in section 2.4..

Clustering A and B identified a total of 21 (49%) statistically relevant variables, while clustering C identified 22 (51%).

With respect to socio-economic variables (age, sex, civil status, education and job), Clustering A found statistical differences between Sex, Education level and type of Job, but not Age. Clustering B found statistical differences only in Age and Sex, while clustering C found differences between Age, Marital status, Education level and type of Job. Clustering A and B identified a total of 21 (49%) statistically relevant variables, while clustering C identified 22 (51%). Table 4.1 shows the significant attributes for the various clustering including Age, Sex, Education level, Occupation level and civil status.

Table 4.1: Difference between the 3 clustering in Age, Sex, Education, Job

	Age	Sex	Education	Job	Marital status
A - C1	/	76% F	44% high school	73% job	/
A - C2	/	66% F	42% degree	70% job	/
B - C1	28 (25,30)	76% F	/	/	/
B - C2	47(39,55)	64% F	/	/	/
B - C3	49 (38, 55)	73% F	/	/	/
B - C4	24 (22, 27)	71% F	/	/	/
B - C5	42 (35, 50)	69% F	/	/	/
B - C6	65 (60.7, 70.3)	54% F	/	/	/
C - C1	45 (35, 53)	/	61% degree	89% job	86% married
C - C2	39 (29, 53)	/	39% degree	67% job	60% married

4.2. Willingness to get vaccinated

In clustering A and B, despite the statistically significant differences calculated, there are no characteristics that are enough relevant as to be able to delineate a specific identity between the attributes. Furthermore, through the analysis of the dataset, eliminating the "Willingness being vaccinated" variable, it was noticed how the Average Silhouette values decrease sharply. This means that by eliminating the specific question “As soon as it is possible for you, do you intend to get vaccinated against the new coronavirus / COVID-19?", the clustering algorithms were unable to determine specific characteristics that effectively distinguished the two populations. In Table 4.2 is shown how the population that does not intend to get vaccinated is distributed among the various clusters.

Table 4.2: Difference between the 3 clustering in Willingness being vaccinated attribute

	Willingness being vaccinated
A - C1	92% no
A - C2	99% yes
B - C1	94% yes
B - C2	88% yes
B - C3	54% no
B - C4	92% yes
B - C5	96% yes
B - C6	98.5% yes
C - C1	91.5% no
C - C2	91.5% yes

In clustering C, on the other hand, there is a marked difference in specific attributes, differently from clustering A and B.

First of all, as in cluster A, it is possible to distinguish cluster 1 as the population that at 91.5% does not intend to be vaccinated against COVID-19, and cluster 2 which, with the same percentage of 91.5%, intends to be vaccinated.

The age difference is well highlighted, in fact the population of cluster 2 is younger. It has been calculated the 25th percentile equal to 29 years, and a median of 39, compared to the 25th percentile of cluster 1 which equals to 35, and has a median of 45 years.

At the occupational level, among those belonging to cluster 2, there are 30% of individuals who have a job in health care, against 10% of cluster 1. Furthermore, cluster 2 is more aware of the risk of contagion, as they have expressed a highest probability score of contracting Covid-19, with the 75th percentile equal to 'very high'.

Trust in institutions has a median equal to 'disagree' for the cluster 1 population. The median corresponds to 'agree' for individuals in cluster 2.

Another marked difference is the previous vaccination history. Indeed, 99% of cluster 1 did not carry out vaccinations between 2019 and 2020, against 46% of cluster 2 who instead carried out other vaccinations.

However, clustering C was formed on a numerically different dataset from the starting one. In fact, as described in section 2.4.1., before training the classification methods, it has been performed the data balancing. Then, through the SMOTE-NC oversampling algorithm, some records belonging to target 0 were replicated, so as to obtain a dataset of 1798 records, divided exactly in half between target 0 and target 1.

After oversampling, the subdivision of the dataset into train, validation, test set was applied. At this point the chosen model, that is the neural network MLP, was applied to the test set, containing 450 total records. The model thus performed the prediction, obtaining an exactly 50% subdivision between target 0 and 1. The prediction was exact for 206 out of 225 records in both cluster 1 and cluster 2. This equally distributed division is probably derived from the balancing performed, which, however, could not be avoid. In fact, by running the prediction algorithms on the original dataset, the results gave a constant overfitting for target 1, that is the most populated one, and a constant underfitting for target 0.

4.3. Comparison between clustering and classification methods

Both clustering and classification are two methods of pattern identification used in machine learning. The main difference lies in the fact that the classification uses predefined classes to which the datapoints are assigned, while clustering identifies similarities between the data and groups them according to specific characteristics in common, forming 'clusters' [76].

Clustering is therefore an unsupervised learning technique, hence from a set of input data, not labelled, information is extracted without knowing the output.

On the other hand, the classification belongs to supervised learning techniques, which means that the algorithm receives labelled data as input, the output of which is known. The binary classification is the one that was applied to this study, having used a precisely binary target variable. It has been chosen to use the "Willingness being vaccinated" attribute as a discriminant, to assess whether the information contained within the dataset was sufficient to allow a prediction through supervised learning.

The objective of both methods was to assess whether they were able to distinguish the population that wants to get vaccinated from the one that does not want, starting from the initial dataset.

Regarding clustering methods, the most promising result is that obtained by Hierarchical clustering on the dataset containing all the selected attributes. The dataset was divided into two clusters, with a marked division between the population with the intention of vaccinating and the population that does not want to be vaccinated.

Regarding the classification methods, valid results were obtained from each model applied to the balanced dataset. Among all, the MLP neural network was chosen, because the one with the highest values in F1 score and AUC. It can therefore be concluded that supervised learning was able to predict the target of the population, starting from the processed dataset.

4.4. Limitations and future developments

In order to improve vaccination compliance, the Personas and, more specifically, the Persona Cards, can be a support to help the population make more informed choices. The development of a number of Personas would have the purpose of identifying specific characteristics among those who are less inclined to vaccinate against Covid-19, so as to be able to study the causes and possible solutions.

In his work it has not been possible to generate Persona Cards, since the resulting clusters where not able to highlight sufficient characteristics to construct effective identities.

The survey was distributed at the beginning of the vaccination campaign, when a large part of the population still did not actually have access to the vaccine. Many people therefore had not yet been directly involved in it and this may have generated bias in the answers given to the survey.

The distribution of the survey was found to be at 85% in Northern Italy, almost exclusively limited to Lombardy. This element meant that the observations collected did not represent a geographically homogeneous sample. Furthermore, the dissemination channel for the questionnaire was online, a choice mainly dictated by the state of emergency and by the stringent Covid-rules. This may have excluded from the compilation people with limited access to the internet or digital tools such as smartphones, tablets or computers, and elderly people who are less familiar with these tools.

For a more in-depth and effective development of the study, it is plausible to expand the population and, having, to date, reached the third dose of vaccine, distributing the survey being able to distinguish who has actually been vaccinated and who has not.

Bibliography

[1] T. P. Velavan and C. G. Meyer, 'The COVID-19 epidemic', *Trop. Med. Int. Health*, vol. 25, no. 3, pp. 278–280, Mar. 2020, doi: 10.1111/tmi.13383.

[2] T. Le *et al.*, 'The COVID-19 vaccine development landscape', *Nat. Rev. Drug Discov.*, Apr. 2020, doi: 10.1038/d41573-020-00073-5.

[3] D.-G. Ahn *et al.*, 'Current Status of Epidemiology, Diagnosis, Therapeutics, and Vaccines for Novel Coronavirus Disease 2019 (COVID-19)', vol. 30, no. 3, pp. 313–324, Mar. 2020, doi: 10.4014/jmb.2003.03011.

[4] 'Rapporti ISS COVID-19 in English', *ISS*. https://iss.it/rapporti-iss-covid-19-in-english (accessed Jul. 04, 2022).

[5] S. B. Omer, I. Yildirim, and H. P. Forman, 'Herd Immunity and Implications for SARS-CoV-2 Control', *JAMA*, vol. 324, no. 20, pp. 2095–2096, Nov. 2020, doi: 10.1001/jama.2020.20892.

[6] 'Ten health issues WHO will tackle this year'. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 (accessed Jul. 04, 2022).

[7] L. K. Gallos, F. Liljeros, P. Argyrakis, A. Bunde, and S. Havlin, 'Improving immunization strategies', *Phys. Rev. E*, vol. 75, no. 4, p. 045104, Apr. 2007, doi: 10.1103/PhysRevE.75.045104.

[8] V. Tarantino, I. Tasca, N. Giannetto, G. R. Mangano, P. Turriziani, and M. Oliveri, 'Impact of Perceived Stress and Immune Status on Decision-Making Abilities during COVID-19 Pandemic Lockdown', *Behav. Sci.*, vol. 11, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/bs11120167.

[9] J. B. Ulmer and M. A. Liu, 'Ethical issues for vaccines and immunization', *Nat. Rev. Immunol.*, vol. 2, no. 4, Art. no. 4, Apr. 2002, doi: 10.1038/nri780.

[10] Alan Cooper and Paul Saffo, *The Inmates Are Running the Asylum.* Macmillan Publishing Co., Inc., USA., 1999.

[11] L. Jalali and R. Jain, 'Building health persona from personal data streams', in *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia - PDM '13*, Barcelona, Spain, 2013, pp. 19–26. doi: 10.1145/2509352.2509400.

[12] E. Tauro, A. Gorini, C. Caglio, P. Gabanelli, and E. G. Caiani, 'COVID-19 and mental disorders in healthcare Personnel: A novel framework to develop Personas from an online survey', *J. Biomed. Inform.*, vol. 126, p. 103993, Feb. 2022, doi: 10.1016/j.jbi.2022.103993.

[13] 'ENTHUSIASTIC ENDORSEMENTS FROM BOTH FOUNDERS OF THE NIELSEN NORMAN GROUP!', in *The Persona Lifecycle*, Elsevier, 2006, p. i. doi: 10.1016/B978-0-12-566251-2.50022-8.

[14] M. Giuliani, A. Ichino, A. Bonomi, R. Martoni, S. Cammino, and A. Gorini, 'Who Is Willing to Get Vaccinated? A Study into the Psychological, Socio-Demographic, and Cultural Determinants of COVID-19 Vaccination Intentions', *Vaccines*, vol. 9, no. 8, Art. no. 8, Aug. 2021, doi: 10.3390/vaccines9080810.

[15] C. Gale and O. Davidson, 'Generalised anxiety disorder', *BMJ*, vol. 334, no. 7593, pp. 579–581, Mar. 2007, doi: 10.1136/bmj.39133.559282.BE.

[16] 'Multidimensional Health Locus of Control Scale - an overview | ScienceDirect Topics'. https://www.sciencedirect.com/topics/medicine-and-dentistry/multidimensional-health-locus-of-control-scale (accessed Jun. 08, 2022).

[17] B. Wallston, K. Wallston, G. Kaplan, and S. Maides, 'Development and Validation of the Health Locus of Control (HLC) Scale', *J. Consult. Clin. Psychol.*, vol. 44, pp. 580–5, Sep. 1976, doi: 10.1037//0022-006X.44.4.580.

[18] S. Zhang, C. Zhang, and Q. Yang, 'Data preparation for data mining', *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, May 2003, doi: 10.1080/713827180.

[19] C. Vercellis, 'Business Intelligence: Data Mining and Optimization for Decision Making', p. 30.

[20] S. K. Kwak and J. H. Kim, 'Statistical data preparation: management of missing values and outliers', *Korean J. Anesthesiol.*, vol. 70, no. 4, pp. 407–411, Aug. 2017, doi: 10.4097/kjae.2017.70.4.407.

[21] Jiří Kaiser, 'Dealing with Missing Values in Data'. Kaiser, Jiří. 'Dealing with Missing Values in Data.' Journal of Systems Integration (1804-2724) 5.1 (2014).

[22] D. Tran Thanh, N. Thai-Nghe, N. Hai, and L. Sang, 'Deep Learning with Data Transformation and Factor Analysis for Student Performance Prediction', *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, pp. 711–721, Sep. 2020, doi: 10.14569/IJACSA.2020.0110886.

[23] N. Abdennour, T. Ouni, and N. B. Amor, 'The importance of signal pre-processing for machine learning: The influence of Data scaling in a driver identity classification', in *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*, Nov. 2021, pp. 1–6. doi: 10.1109/AICCSA53542.2021.9686756.

[24] A. Jović, K. Brkić, and N. Bogunović, 'A review of feature selection methods with applications', in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2015, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

[25] L. Wilkinson and M. Friendly, 'The History of the Cluster Heat Map', *Am. Stat.*, vol. 63, no. 2, pp. 179–184, May 2009, doi: 10.1198/tas.2009.0033.

[26] K. Much, C. Hill, N. Bader, and C. Hill, 'Building Heat Maps for Data Cleaning and Beyond', p. 12.

[27] I. T. Jolliffe and J. Cadima, 'Principal component analysis: a review and recent developments', *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[28] M. Bécue-Bertaut and J. Pagès, 'Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data', *Comput. Stat. Data Anal.*, vol. 52, no. 6, pp. 3255–3268, Feb. 2008, doi: 10.1016/j.csda.2007.09.023.

[29] G. Tuerhong and S. B. Kim, 'Gower distance-based multivariate control charts for a mixture of continuous and categorical variables', *Expert Syst. Appl.*, vol. 41, no. 4, Part 2, pp. 1701–1707, Mar. 2014, doi: 10.1016/j.eswa.2013.08.068.

[30] A. K. Mann and N. Kaur, 'Review Paper on Clustering Techniques', *Glob. J. Comput. Sci. Technol.*, May 2013, Accessed: Jun. 09, 2022. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/353

[31] W. Budiaji and F. Leisch, 'Simple K-Medoids Partitioning Algorithm for Mixed Variable Data', *Algorithms*, vol. 12, no. 9, Art. no. 9, Sep. 2019, doi: 10.3390/a12090177.

[32] T. Kodinariya and P. Makwana, 'Review on Determining of Cluster in K-means Clustering', *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, pp. 90–95, Jan. 2013.

[33] Velmurugan, 'Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points', *J. Comput. Sci.*, vol. 6, no. 3, pp. 363–368, Mar. 2010, doi: 10.3844/jcssp.2010.363.368.

[34] P. Arora, Deepali, and S. Varshney, 'Analysis of K-Means and K-Medoids Algorithm For Big Data', *Procedia Comput. Sci.*, vol. 78, pp. 507–512, Jan. 2016, doi: 10.1016/j.procs.2016.02.095.

[35] S. Pandya and S. Saket, 'An overview of partitioning algorithms in clustering techniques', *Int. J. Electr. Comput. Eng.*, vol. 5, Sep. 2020.

[36] D. Arthur and S. Vassilvitskii, 'k-means++: The Advantages of Careful Seeding', p. 11.

[37] F. Murtagh and P. Contreras, 'Algorithms for hierarchical clustering: an overview, II', *WIREs Data Min. Knowl. Discov.*, vol. 7, no. 6, p. e1219, 2017, doi: 10.1002/widm.1219.

[38] F. Murtagh and P. Legendre, 'Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm', *J. Classif.*, vol. 31, no. 3, pp. 274–295, Oct. 2014, doi: 10.1007/s00357-014-9161-z.

[39] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, 'Efficient agglomerative hierarchical clustering', *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2785–2797, Apr. 2015, doi: 10.1016/j.eswa.2014.09.054.

[40] A. Jarman, *Hierarchical Cluster Analysis: Comparison of Single linkage,Complete linkage, Average linkage and Centroid Linkage Method*. 2020. doi: 10.13140/RG.2.2.11388.90240.

[41] M. Shutaywi and N. N. Kachouie, 'Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering', *Entropy*, vol. 23, no. 6, Art. no. 6, Jun. 2021, doi: 10.3390/e23060759.

[42] P. J. Rousseeuw, 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[43] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, 'The Clustering Validity with Silhouette and Sum of Squared Errors', in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 2015, pp. 44–51. doi: 10.12792/iciae2015.012.

[44] A. Hart, 'Mann-Whitney test is not just a test of medians: differences in spread can be important', *BMJ*, vol. 323, no. 7309, pp. 391–393, Aug. 2001, doi: 10.1136/bmj.323.7309.391.

[45] J. Derrac, S. García, D. Molina, and F. Herrera, 'A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms', *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 3–18, Mar. 2011, doi: 10.1016/j.swevo.2011.02.002.

[46] 'Nonparametric Test - an overview | ScienceDirect Topics'. https://www.sciencedirect.com/topics/medicine-and-dentistry/nonparametric-test (accessed Jun. 10, 2022).

[47] E. U. Oti, M. O. Olusola, and P. A. Esemokumo, 'Statistical Analysis of the Median Test and the Mann-Whitney U Test', vol. 7, no. 9, p. 8, 2021.

[48] H.-Y. Kim, 'Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test', *Restor. Dent. Endod.*, vol. 42, no. 2, pp. 152–155, Mar. 2017, doi: 10.5395/rde.2017.42.2.152.

[49] E. Ostertagová, O. Ostertag, and J. Kováč, 'Methodology and Application of the Kruskal-Wallis Test', *Appl. Mech. Mater.*, vol. 611, pp. 115–120, 2014, doi: 10.4028/www.scientific.net/AMM.611.115.

[50] 'Bonferroni Correction - an overview | ScienceDirect Topics'. https://www.sciencedirect.com/topics/computer-science/bonferroni-correction (accessed Jun. 26, 2022).

[51] G. Kesavaraj and S. Sukumaran, 'A study on classification techniques in data mining', in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Jul. 2013, pp. 1–7. doi: 10.1109/ICCCNT.2013.6726842.

[52] 'A Method for Optimal Division of Data Sets for Use in Neural Networks | SpringerLink'. https://link.springer.com/chapter/10.1007/11554028_1 (accessed Jul. 04, 2022).

[53] S. Neelamegam and D. E. Ramaraj, 'Classification algorithm in Data mining: An Overview', vol. 4, no. 8, p. 7, 2013.

[54] D. L. Olson, 'Data Set Balancing', in *Data Mining and Knowledge Management*, Berlin, Heidelberg, 2005, pp. 71–80. doi: 10.1007/978-3-540-30537-8_8.

[55] N. Poolsawad, C. Kambhampati, and J. G. F. Cleland, 'Balancing Class for Performance of Classification with a Clinical Dataset', *Lect. Notes Eng. Comput. Sci.*, vol. 1, pp. 237–242, Jul. 2014.

[56] V. Ganganwar, 'An overview of classification algorithms for imbalanced datasets'. 2012.

[57] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 'SMOTE: Synthetic Minority Over-sampling Technique', *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[58] A. J. Mohammed, 'Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method', *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3161–3172, Jun. 2020, doi: 10.30534/ijatcse/2020/104932020.

[59] B. Gulowaty and P. Ksieniewicz, 'SMOTE Algorithm Variations in Balancing Data Streams', in *Intelligent Data Engineering and Automated Learning – IDEAL 2019*, Cham, 2019, pp. 305–312. doi: 10.1007/978-3-030-33617-2_31.

[60] L. Demidova and I. Klyueva, 'SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem', in *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, Jun. 2017, pp. 1–4. doi: 10.1109/MECO.2017.7977136.

[61] A. Priyam, Abhijeet, R. Gupta, A. Rathee, and S. Srivastava, 'Comparative Analysis of Decision Tree Classification Algorithms'.

[62] S. Tangirala, 'Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*', *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, 2020, doi: 10.14569/IJACSA.2020.0110277.

[63] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, 'An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm', in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 2015, pp. 280–285. doi: 10.12792/iciae2015.051.

[64] A. Kataria and M. D. Singh, 'A Review of Data Classification Using K-Nearest Neighbour Algorithm'.

[65] V. Vapnik, 'Principles of Risk Minimization for Learning Theory', p. 8.

[66] M. C. Green, A. Khalifa, M. Charity, D. Bhaumik, and J. Togelius, 'Predicting Personas Using Mechanic Frequencies and Game State Traces'. arXiv, Mar. 24, 2022. Accessed: Jun. 04, 2022. [Online]. Available: http://arxiv.org/abs/2203.13351

[67] L. Su and Y. Huang, 'Support Vector Machine (SVM) Classification: Comparison of Linkage Techniques Using a Clustering-Based Method for Training Data Selection', *GIScience Remote Sens.*, vol. 46, no. 4, pp. 411–423, Oct. 2009, doi: 10.2747/1548-1603.46.4.411.

[68] S. Jain, S. Shukla, and R. Wadhvani, 'Dynamic selection of normalization techniques using data complexity measures', *Expert Syst. Appl.*, vol. 106, pp. 252–262, Sep. 2018, doi: 10.1016/j.eswa.2018.04.008.

[69] A. Pinkus, 'Approximation theory of the MLP model in neural networks', *Acta Numer.*, vol. 8, pp. 143–195, Jan. 1999, doi: 10.1017/S0962492900002919.

[70] Z. Car, S. Baressi Šegota, N. Anđelić, I. Lorencin, and V. Mrzljak, 'Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron', *Comput. Math. Methods Med.*, vol. 2020, p. e5714714, May 2020, doi: 10.1155/2020/5714714.

[71] P. J. G. Lisboa, 'A review of evidence of health benefit from artificial neural networks in medical intervention', *Neural Netw. Off. J. Int. Neural Netw. Soc.*, vol. 15, no. 1, pp. 11–39, Jan. 2002, doi: 10.1016/s0893-6080(01)00111-3.

[72] F. Murtagh, 'Multilayer perceptrons for classification and regression', *Neurocomputing*, vol. 2, no. 5, pp. 183–197, Jul. 1991, doi: 10.1016/0925-2312(91)90023-5.

[73] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, 'Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance', *Neural Netw.*, vol. 21, no. 2–3, pp. 427–436, Mar. 2008, doi: 10.1016/j.neunet.2007.12.031.

[74] S. Narkhede, 'Understanding AUC - ROC Curve', p. 6.

[75] A. J. Bowers and X. Zhou, 'Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes', *J. Educ. Stud. Placed Risk JESPAR*, vol. 24, no. 1, pp. 20–46, Jan. 2019, doi: 10.1080/10824669.2018.1523734.

[76] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, 'A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science', in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds. Cham: Springer International Publishing, 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2_1.

List of Figures

List if Tables

Acknowledgments

Come ho già scritto nelle due precedenti tesi – cosa che specifico per puro autocompiacimento - i ringraziamenti su carta non saranno mai sufficienti a raggiungere lo scopo di ringraziare davvero, e una lista di nomi non è quello che voglio lasciare alle persone che sono state con me in questo infinito, travagliato, entusiasmante, sofferto, esilarante, soddisfacente percorso.

Sono più che certa che ognuno* di voi sappia quanto ha significato per me essermi stato* accanto, e scriverlo qui non renderebbe la cosa più importante di quanto già non sia. Quindi sappiate che siete nel mio cuore, ben radicati* nel profondo.

Un grazie esplicito va ad Emanuele Tauro e al prof. Caiani per avermi dato la possibilità di lavorare su un tema così delicato e così attuale.

L'altro grazie, di cui sento il dovere quanto meno di tenerne traccia, va alla mia famiglia, che per una serie di ragioni, alcune delle quali tuttora a me sconosciute – chiedetele a loro – mi ha sempre supportato nelle scelte intraprese. Senza di loro oggi non sarei quella che sono e che ho fortemente voluto essere.

Un sincero, enorme, emozionante ed urlato grazie lo lascio a me stessa, perché facile non è stato, ma lo rifarei mille volte.

Di seguito troverete scritto un mio personale commento, con nessun intento particolare se non quello di lasciare traccia di un pensiero sull'*oggi* e sul *noi*.

"Questa pandemia non è stata, e non è, una guerra. Di questa pandemia se ne è parlato, a lungo, non si è ancora smesso. Il virus che ha contagiato gran parte del mondo è stato affrontato, studiato. È stato tema di ricerca continuo e incessante. Il virus è stato, quasi del tutto, fermato. É stata data la possibilità a *noi*, individui abitanti della nostra terra, di scegliere per impedire che continuasse ad uccidere. Ci è stato dato il potere di agire.

La guerra è un'altra cosa. Implica armi, violenza e odio sotto gli occhi di tutti*. Bombe, fucilazioni, carri armati che avanzano, e nessuno che si metta davanti a fermarli. La guerra è fame e miseria. La guerra è taciuta, deliberatamente ignorata. È lontana, estranea. I notiziari parlano *degli altri*, che lottano per sopravvivere, o che soccombono sotto il rumore assordante delle bombe e della vigliaccheria. La guerra è fatta di popoli che non vengono ascoltati, di anime di cui non si parla mai abbastanza.

La guerra è quella in Messico, in Nigeria, in Siria, in Iraq, nello Yemen, in Etiopia. In Birmania, in Afghanistan, in Pakistan, in Sudan. Nella Repubblica Democratica del Congo, in Somalia, in Mozambico, in Israele e Palestina. La guerra è quella in Ucraina.

A *noi* è stata data la possibilità di salvarci, dal virus e da noi stessi. A *loro*, al massimo, sono state date coperte, cibo in scatola, e armi per combattere. È stato ed è *nostro* dovere salvarci, perché c'è chi non può scegliere di farlo, ma, semplicemente, soccombe."

You are not only responsible for what you say, but also for what you do not say.

Martin Luther King Jr.