POLITECNICO DI TORINO

Master Degree in ICT for Smart Societies

Master Degree Thesis

Data analysis of patients who received anti-SARS-CoV-2 vaccination with Pfizer vaccine in Italy

Supervisors
Prof. Monica Visintin
Prof. Guido Pagana

Candidate
Ten. Daniele Radaelli

A.Y. 2020/2021

Summary

Many scenarios developed following the COVID-19 epidemic caused by SARS-CoV-2 assume that the infection results in an immune response that confers immunity or reduces the severity of a possible reinfection. The presence or absence of protective immunity due to infection or vaccination will affect future transmission and disease severity.
COVID-19 infection leads to the production of IgG and IgM class antibodies after a few days or weeks. However, it remains to be understood why the infection is sometimes not followed by antibody production in some subjects.
IgG antibodies develop a secondary immune response, which occurs on subsequent exposures to the same antigen [1]. In fact, these antibodies represent the "memory" of the immune system, to be able to intervene in the event of a subsequent infection.
Taking as an example what happened in the course of the infections of other coronaviruses, some hypotheses can be made. For example in Severe Acute Respiratory Syndrome (SARS) (caused by the SARS-Cov-1 virus) the antibody response lasted 4-5 months and then slowly decreased over the next 2-3 years. Also with regard to Middle East Respiratory Syndrome (MERS), the antibody duration in subjects recovered from the disease persisted for at least 34 months [2].
This thesis examines the results of research conducted from the beginning to mid-2021, on 1340 individuals among all the medical staff of the Azienda Ospedaliera Ordine Mauriziano di Torino after both vaccination doses.
The purpose is to find, if it exists, a correlation between the persistence of a certain quantity of IgG anti-S antibodies and the body parameters or the type of tasks performed.

Acknowledgements

I would like to thank my supervisor, Professor Monica Visintin for her advice and her availability. I would like to thank Professor Guido Pagana and Doctor Valeria Figini for their support and guide. I would like to thank my family and my girlfriend Martina for always being close to me, even in the hardest moments. Finally, I want to thank my friends and my course mates, especially Jacopo with whom I shared an intense year of study and projects.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 COVID-19 overview

On 31st December 2019, the Chinese authorities reported to the World Health Organization (WHO) the occurrence of several cases of a mysterious pneumonia. The epicenter was located in Wuhan, a metropolis of 11 million inhabitants in the Hubei region [5].
In a few days the number of cases amounted to 41. After a few days, on 7th January, China declared that the agent of the respiratory disease was a new virus, provisionally named 2019-nCoV and later officially classified under the name SARS-CoV-2.
On 11th February 2020, the WHO announced that the respiratory disease caused by the new coronavirus was named COVID-19. In the short term, the city of Wuhan entered lockdown, while in the rest of the world the risk is still underestimated, since coronaviruses (CoVs) are a large family of respiratory viruses that can cause mild to moderate illness.
Other viruses that belong to this category are MERS and SARS. SARS, discovered in 2002 was the first virus called Coronavirus, mainly localized in Asia, caused the deaths of 744 people globally.
Most people have mild symptoms including: fever, dry cough and weakness and recover quickly without hospitalization.
On average, it takes 5-6 days for a person who has contracted the virus to show symptoms; however, the incubation period can last up to 14 days. During this period, the person in question does not know that he is sick and can easily infect all the people around him.
In rarer cases, more severe symptoms occur such as: muscular and articular pain, diarrhea, vomiting, headache, loss of taste and smell, conjunctivitis, rashes all over the body and in even rarer cases it can even lead to death.

The chances of death are generally correlated with the age of the patient, but also with previous pathologies such as: heart disease, diabetes, renal failure, chronic respiratory diseases, tumors and liver diseases.
The new coronavirus cannot be cured with antibiotics, because they have no effect against viruses, only against bacteria. At first, the infection remained almost exclusively confined to China, but from mid-February 2020 it spread rapidly around the world. Global numbers have followed steady growth. Outside China, the number of infected people was very high in Italy, Iran and South Korea, even if for the WHO COVID-19 was not yet a pandemic. At the beginning of March 2020 the virus begins to spread in Italy. Initially, attempts were made to contain the expansion by isolating only some high-risk areas, especially in northern Italy. On 9th March, the Italian government extends the containment measures to all of Italy, effectively becoming in lockdown. Two days later the WHO will declare a global pandemic.
Based on the data provided daily by the European Centre for Disease Prevention and Control (ECDC), on 30th August 2021 the total cases are 212418662, with 4436327 deaths. On that date there were a total of 205 nations and territories with at least one case of positivity. Our World in Data (OWID) [6] site offers the possibility to consult the situation of the epidemic up to that date, as shown in figure 1.

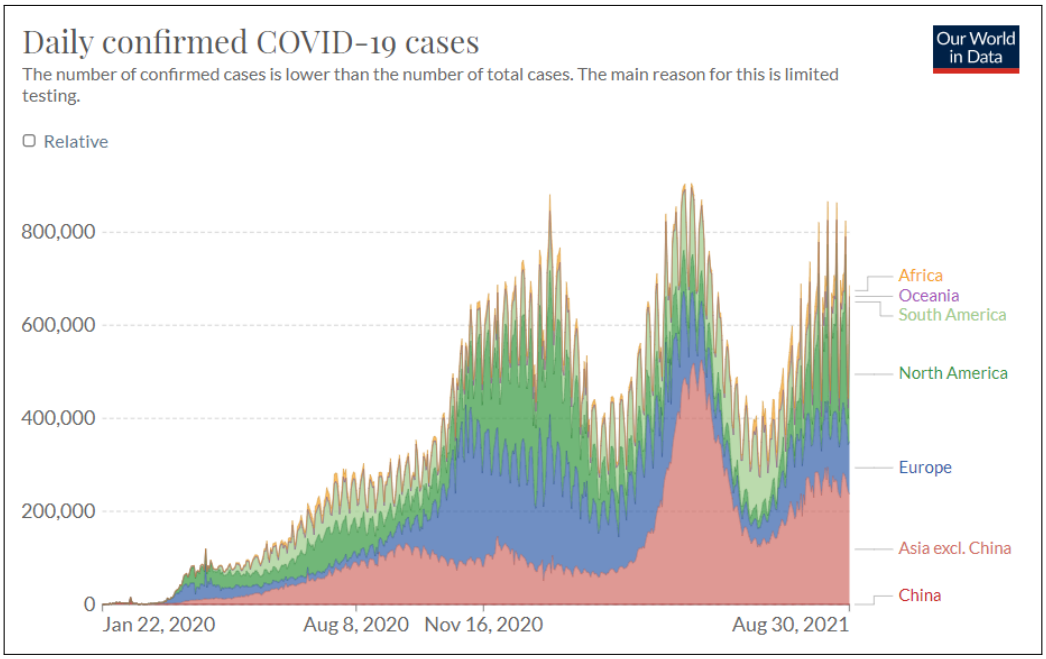


Figure 1: Daily confirmed COVID-19 cases by region until 30th August 2021. Image taken from OWID site.

A little more than two months after the identification of Sars-Cov-2, the first trials of vaccines for COVID-19 began.
However, given the long experimentation phase necessary to produce safe vaccines for humans, it was deemed necessary to carry out research and experimentation on existing drugs, as widely discussed on the Agenzia Regionale Sanitaria (ARS) website [7]. In Italy, the evaluation of all clinical trials on drugs has been entrusted to Agenzia italiana del farmaco (AIFA). AIFA authorizes controlled clinical trials that involve the use of certain treatments on patients suffering from COVID-19. The number of trials is constantly growing and whose updates can be checked on the relative page [8]. At the moment, 31st August 2021, in Italy there are 70 trials in progress on medicines.
In addition, Italy participates in the study promoted by the World Health Organization Public health emergency SOLIDARITY TRIAL [9]: is an international clinical trial to help find an effective treatment for COVID-19. It is one of the largest international studies for COVID-19 treatments, recruiting nearly 12,000 patients in over 30 countries to date. On 15th October 2020, the SOLIDARITY trial published the first results of a study on different types of drugs.
Four treatments gave positive results:

- **remdesivir**: is an antiviral drug, it was developed as a treatment for Ebola virus disease;
- **hydroxychloroquine**: is an antimalarial drug belonging to antirheumatic drugs and used in the treatment of malaria;
- **lopinavir and ritonavir**: is used to treat HIV infection and combines lopinavir with a subtherapeutic dose of ritonavir;
- **interferon**: are proteins produced by the human body of great importance for the immune system.

They had little or no effect on overall mortality, in most of hospitalized patients.
In the UK, another trial called RECOVERY (Randomized Evaluation of COVID-19 Therapy) has given excellent results. The study was carried out on the basis of data collected by the University of Oxford and published on 16th June 2020 [10]. The RECOVERY trial aims to identify effective drugs in the treatment of hospitalized adults with COVID-19.
In particular the only molecule that has been shown to be effective in reducing the number of deaths in the most severe cases of COVID-19 is Dexamethasone, an old anti-inflammatory drug.
Currently the number of studies on COVID-19 is constantly evolving, on ClinicalTrials.gov [11] it is possible to follow the development of all the studies. However, the need for a definitive solution is the only way to permanently eliminate the

COVID-19 epidemic.
Fortunately, on 14th December 2020, the first vaccine against COVID-19, developed by Pfizer BioNTech, was approved in emergencies by the Food and Drug Administration (FDA), it was the first vaccine with mRNA technology.

1.2 COVID-19 vaccinations

During summer of 2020, the situation seems to be improving across Europe. In Italy, daily infections drop below 200 cases and intensive cares are empty. Soon bars, discos and senior centers reopen, and are also authorized team, contact and individual sports.
After the summer period, the number of infections begins to rise again and some security measures are reintroduced, thus starting the second wave. The infections start to grow again in a worrying way and the pressure on the hospitals returns to be felt.
Figure 2 shows the trend of infections during 2020.

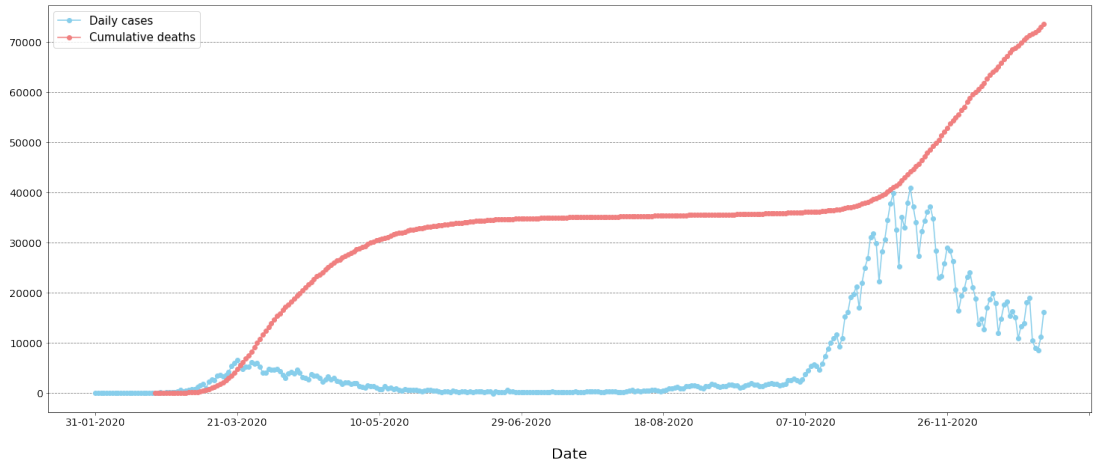


Figure 2: Cumulative curve of confirmed deaths by COVID-19 and daily new cases in Italy during 2020.

In 2020, the restrictions implemented played an important role in containing the virus. After a year characterized by a severe economic crisis that affected all sectors and significant health costs, the research showed that with the passage of time and with the increase in closure stress, the limitations have become less and less effective.
Figure 3 shows the GDP loss of the world's largest countries in the first half of 2020 compared to the previous year. Italy is among the most affected from an economic point of view and for this reason the lockdown is not sustainable for a long time.

Furthermore, the prohibitions are broken with a progressively greater frequency and also the rules are respected in a less scrupulous way. Thus, the need to vaccinate the population as soon as possible becomes essential.

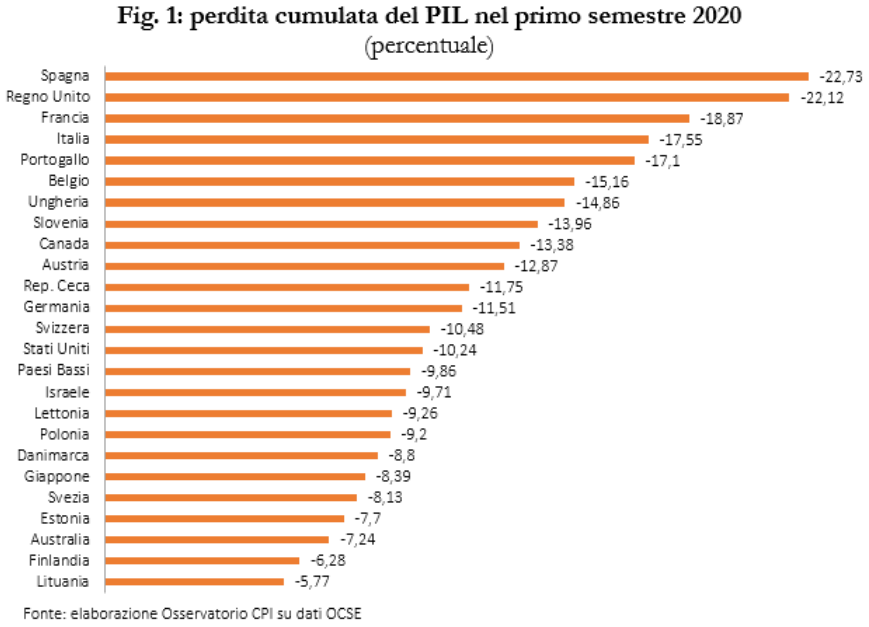


Figure 3: Cumulative loss in GDP in the first half of 2020.
Image taken from [3]

1.2.1 The supply of a COVID-19 vaccine

The discovery and development of a vaccine against COVID-19 soon became the goals of a vast scientific effort worldwide, which also stimulated further investigations to understand where the virus could most infect the world's population. In just under a year, all the research and verification phases of efficacy and safety were concentrated.
As reported on the website of the Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Marco Negri [12], the short lead times were made possible by:

- close collaboration among most of the countries of the world, which have made their knowledge and structures available to contribute to the experimentation;
- speeding up the bureaucratic phases;
- using of new technologies;
- taking advantage of the knowledge acquired by studying the previous coronaviruses: MERS and SARS;
- provisioning of a large number of public funding.

Goals of the vaccination campaign

The main goal of the vaccination campaign is to prevent deaths from COVID-19 and to achieve herd immunity (corresponding to 70% of the population) for SARS-CoV2 as soon as possible.
Being vaccinated can also protect the people around us, because if a person is protected from infection and disease, he/she is less likely to infect someone else.
The campaign of the first COVID-19 vaccine started on 27th December 2020 in demonstration form in Italy and Europe with vaccine day and effectively on 31st December 2020 after the approval by the European Medicines Agency (EMA).
The vaccination campaign is developing in an increasing way, up to the target of 500,000 daily doses, following the strategic plan. Vaccines are offered free of charge to the entire population, according to an order of priority, which takes into account the risk of contracting the virus, age, regressed diseases, types of vaccines and their availability.

Which vaccines are available in Italy?

In mid-2021 there are more than 200 vaccines under development worldwide. As reported on the website of the ministry of health [13], Italy can count on the availability of over 255 million doses and to date there are four anti-COVID vaccines available:

1. **Comirnaty Pfizer vaccine**, approved in Europe on 21st December and on 22nd December 2020 by the AIFA. The American company, in collaboration with Biontech, a German biotechnology company, carried out the studies on 44,000 people. The effectiveness was calculated on over 36,000 people from 16 years of age and was found to be 95%;

2. **Moderna vaccine**, approved in Europe on 6th January and in Italy on 7th January 2021: the US company carried out the studies on a sample of 30,000 people between 18 and 94 years, observing an effectiveness of 94%. Also in patients suffering from chronic diseases such as diabetes, cardiovascular diseases and obesity was measured an effectiveness of 91%;

3. **AstraZeneca vaccine**, developed in collaboration with the University of Oxford, approved by the EMA on 30th January 2021. The study involved 18,000 people. The effectiveness against symptomatic infection is 60%, but increases to 80% in the case of severe forms of virus;

4. **Janssen vaccine by Johnson & Johnson**, it was approved by the EMA on 11th March 2021 and by the AIFA on 12th March 2021. 22,000 patients received the vaccine, developing an overall effectiveness of 66% against infection.

Furthermore, Table 1 indicates the differences between the vaccines administered in Italy.

	Pfizer/Biontech	**Astrazeneca**	**Moderna**	**Janssen**
Nationality	USA/Germany	United Kingdom	USA	Belgium
How effective	95%	62-80%	95%	72-86%
Doses	2	2	2	1
Dosing schedule	21 days	42 days	28 days	-
Type	mRNA	Viral vector	mRNA	Viral vector
Age range	$\geq 16y/o$	$\geq 18y/o$	$\geq 18y/o$	$\geq 18y/o$
Storage	-70 $\pm$ 10°C	-20°C	+2°C to +8°C	+2°C to +8°C

Table 1: COVID-19 vaccines comparison.

It is easy to see that vaccines using messenger RNA have been found to be more effective in covering the contraction of the virus. However, viral vector vaccines such as Astrazeneca and Janssen were found to be better suited for people over 55 [14]. Although the Pfizer vaccine covers more than 60% of the doses used in Italy, it requires lower temperatures and a more complex cold chain; on the contrary the Astrazeneca and Janssen vaccines can be stored at temperatures between 2 and 8 °C for up to 6 months.

How do COVID-19 vaccines work?

Pfizer and **Moderna** vaccines use mRNA or messenger RNA technology. It is a nucleic acid molecule that contains the genetic information to allow our body to produce the protein that forms the spikes of the coronavirus "autonomously". The spike protein, which represents the means by which the virus manages to enter our organism, is therefore the target against which our antibodies are directed.
The messenger RNA is not capable of reproducing itself in host cells but can only induce the synthesis of its spikes. The genetic material is in fact enclosed within a microscopic bubble, made of fats or lipids, between 1 and 100 nanometers large. Once the vaccine solution has been injected, the nanoparticles, with the various copies of mRNA inside, are absorbed by each single cell around the injection site. At this point, the cells begin to produce viral proteins, which then in turn stimulate the immune system.
The injected genetic material has a very short life: in fact, it degrades within 8-10 hours. Therefore, it is absolutely impossible for it to be able to enter the nucleus of our cells, where the DNA is present, altering our genes.
Furthermore, the vaccine does not contain the virus and therefore cannot cause disease.
If a vaccinated person finds himself in contact with the virus, he will be immune

because the antibodies produced thanks to the vaccine will block the entry of Spike proteins into the cells. Vaccination also activates T cells or T-lymphocytes, which come into play as a second line of defense. They instruct the immune system to become capable of responding to subsequent exposure to SARS-CoV-2 as well.
Figure 4 shows how works these types of vaccine.

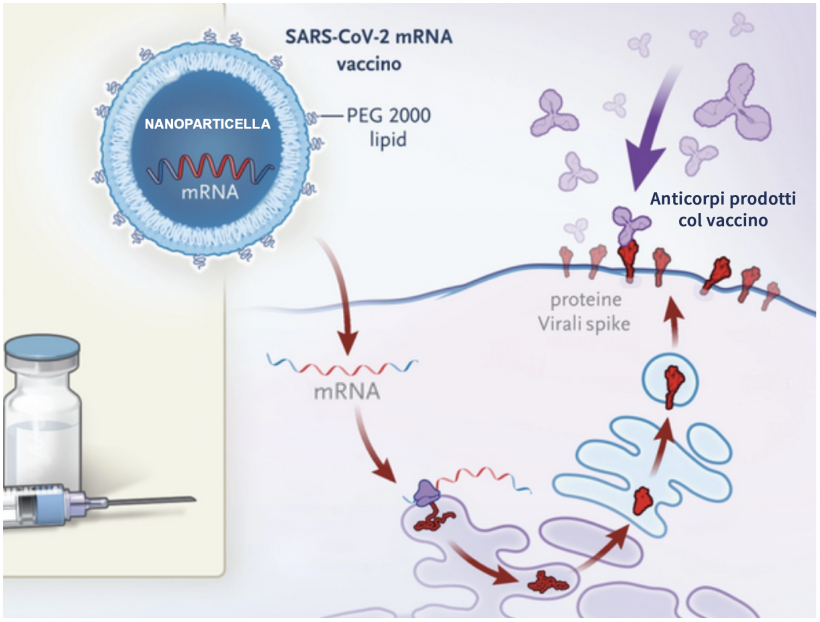


Figure 4: How mRNA vaccine works
Image taken from IRCCS site.

The **Vaxzevria** and **Johnson & Johnson** vaccine, on the other hand, use a different vehicle to introduce the viral genetic information needed to produce spike proteins into the body. As with the Ebola vaccine, the vehicle is an adenovirus, specifically the virus that causes colds in chimpanzees. However, these adenoviruses have been inactivated and that is rendered harmless to humans. Inside the shell of the modified virus, there is material called a "viral vector". Once the viral vector is inside our cells, it gives cells instructions to make a protein that is unique to the virus that causes COVID-19. Using these instructions, our cells make copies of the protein.
As in the previous case, this allow our bodies to build T-lymphocytes and B-lymphocytes that will remember how to fight that virus if we are infected in the future.
Figure 5 shows how works these types of vaccine.

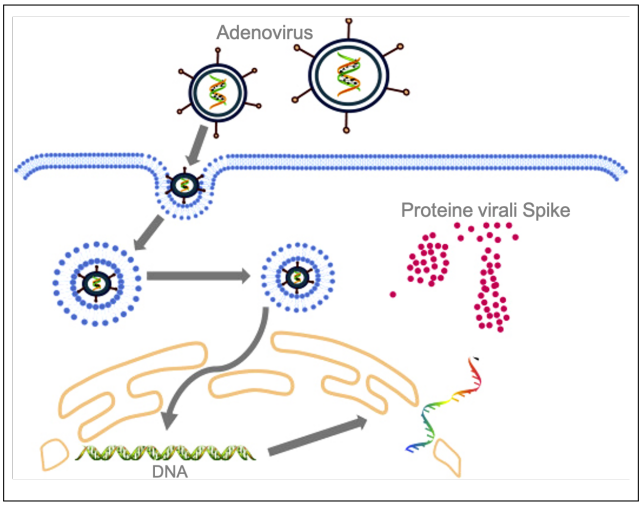


Figure 5: How viral vector vaccine works
Image taken from IRCCS site.

1.2.2 Analysis on national data

In order to better analyze the efficiency of vaccines, national data relating to the Italian territory alone were analyzed. The dataset used was downloaded from Our World in Data's GitHub profile [15] which is updated daily with the data provided by each country. All the graphs shown below are based on data downloaded on 12th June 2021.

Efficiency of vaccines in Italy

It should be kept in mind that the end of October marked the second wave of infections and at the beginning of 2021 the number of daily infections stood over 10,000 cases. At first, the impact of vaccines on the entire population was analyzed. The period taken into consideration starts on 27th December 2020, the day the vaccinations began and ends on 10th June, 2021.
Figure 6 shows the trend of daily vaccinations compared with the number of cases (multiplied by 10, to make the data more readable).
The darker, solid lines represent the moving average calculated over a 7-day period. This allows for greater stability in the display of trends.

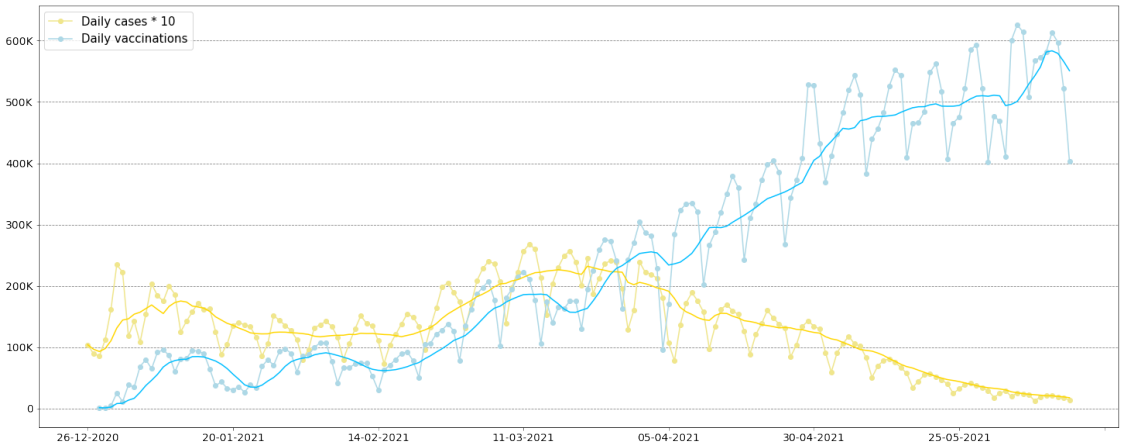


Figure 6: Number of daily vaccinations and COVID infections

In 2021 mid-May, the target of 500,000 vaccinations that the Special Commissioner for the COVID emergency had set was reached. At the same time, the positive cases also drastically decreased.
To get a clearer response, however, it is necessary to compare the cumulative number of vaccinations day after day with the number of cases (multiplied by 1000). Furthermore, according to the first data found from the current campaign, protection from the virus is not immediate after the vaccine is inoculated but develops progressively after at least 7-14 days from the injection.
Figure 7 shows the cumulative trends of the first and second vaccine doses.

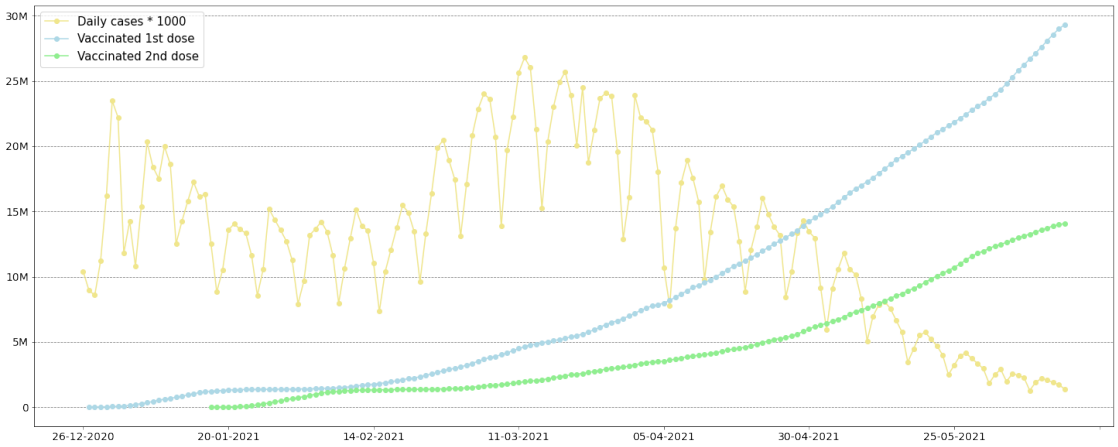


Figure 7: Cumulative vaccine doses administered compared to the number of daily infections

Figure 8 shows the same data, but the number of vaccinated people is delayed

by 14 days, in order to represent a more realistic view of the situation, in which each individual is truly protected.

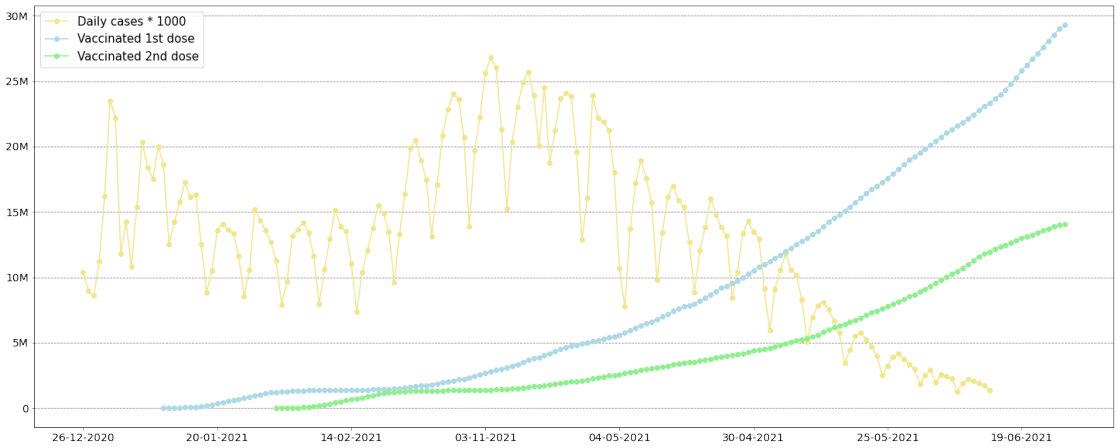


Figure 8: Cumulative vaccine doses administered compared to the number of daily infections shifted by 14 days

Especially in figure 7 it is possible to notice a drastic drop in infections, precisely in correspondence with the achievement of 10 million vaccinated. It is also important to underline that since March, the containment measures throughout the Italian territory have gradually eased. However, the number of infections is not an indicative parameter since it depends a lot on the swabs performed.
One of the indices that best represents the trend of is the number of dead patients per day, as shown in figure 9.

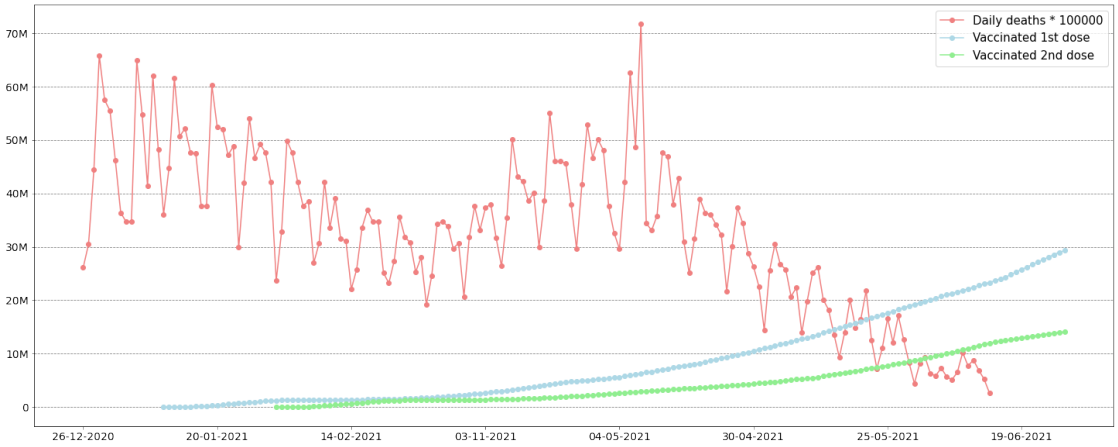


Figure 9: Cumulative vaccine doses administered compared to the number of daily deaths shifted by 14 days

In this case, to make the number of vaccinations comparable with the number of deaths, the latter was suitably multiplied by 100,000. Figures 6,7,8, and 9 show the correlation of the number of vaccinations with some reference parameters. However, it is necessary to underline that although the result is positive, also in 2020 there was a significant decrease in the number of infected people in conjunction with the summer period.

Comparison between 2020 and 2021

As already seen above, the precautionary lockdown measure adopted between March and May 2020 in Italy was almost completely successful in solving the COVID emergency. However, it had only managed to stem the problem, again forcing the entire population to submit to new containment measures.
The vaccination plan aims to put an end to the virus definitively, simultaneously guaranteeing the gradual resumption of all activities.
In this section, various data relating to the virus will be compared in 2020 and 2021, in the period from 1st March to 10th June.
The number of daily cases is not an objective parameter because it also depends on the number of swabs. A clearer index is the "*positive rate*", that represents the number of positive swabs out of the total. Figure 10 compares the values of 2020 and 2021.

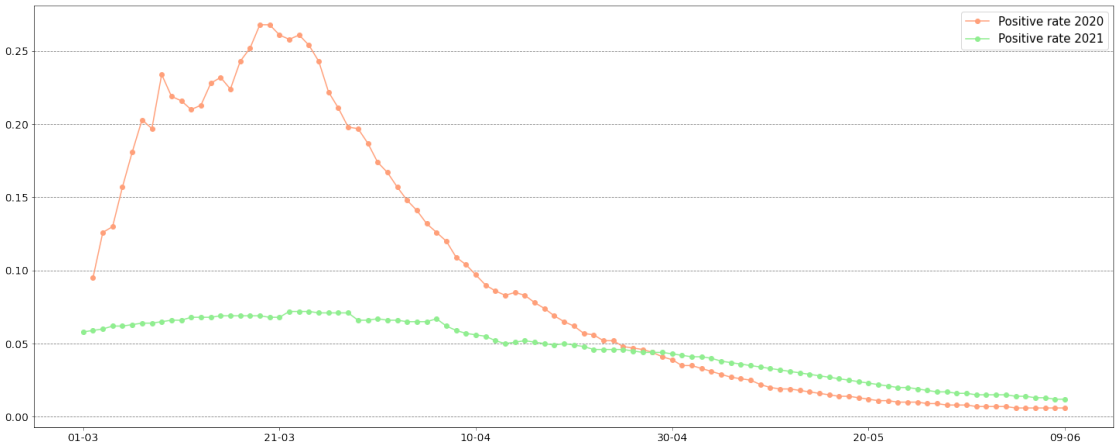


Figure 10: Daily positive rate comparison of 2020 and 2021

As in 2020, 2021 also peaks in mid-March, albeit significantly lower. In both cases there was a significant decrease in the number of infected people, reaching a value below 0.01 in June, i.e. 1%. However, even in this case the data are difficult to compare due to the different criteria for using the tampons. In fact, in 2020 they were only done to people who were already experiencing symptoms

related to COVID. While in 2021 they are carried out to anyone who comes to the hospital and to all people who have come into contact with a positive person. Figures 11, 12 and 13 instead show the data of daily deaths, hospitalized patients and patients in Intensive Care Unit (ICU).
The difference between a hospitalized patient and one in need of intensive care is that the latter is aimed at stabilizing the vital functions of the patients. Which requires an advanced monitoring of the patient 24 hours a day and the use of technologies that primarily support the respiratory and cardio-circulatory functions.

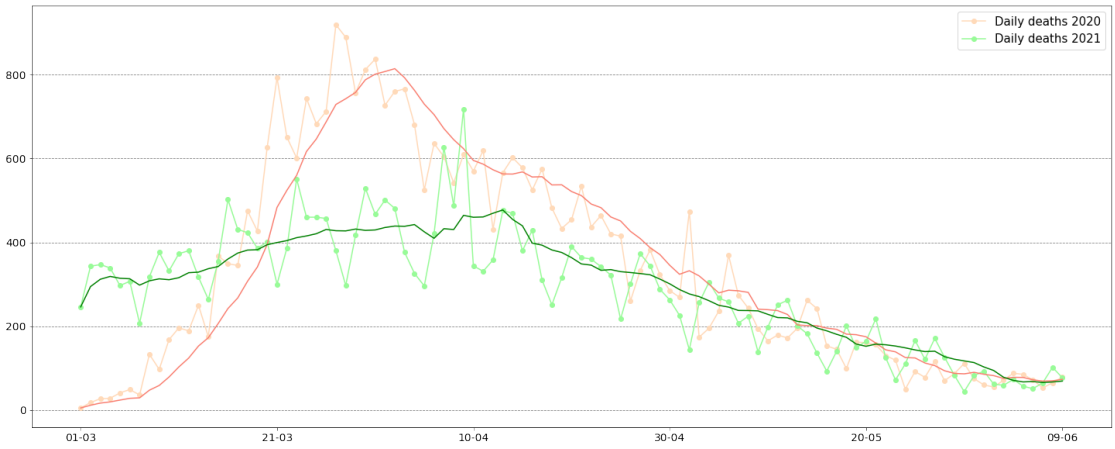


Figure 11: Daily deaths comparison of 2020 and 2021

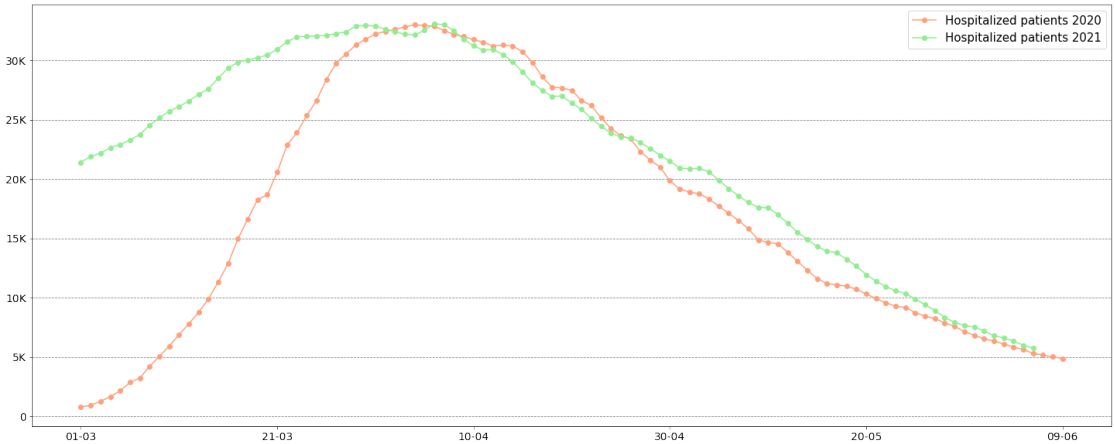


Figure 12: Hospitalized patients comparison of 2020 and 2021

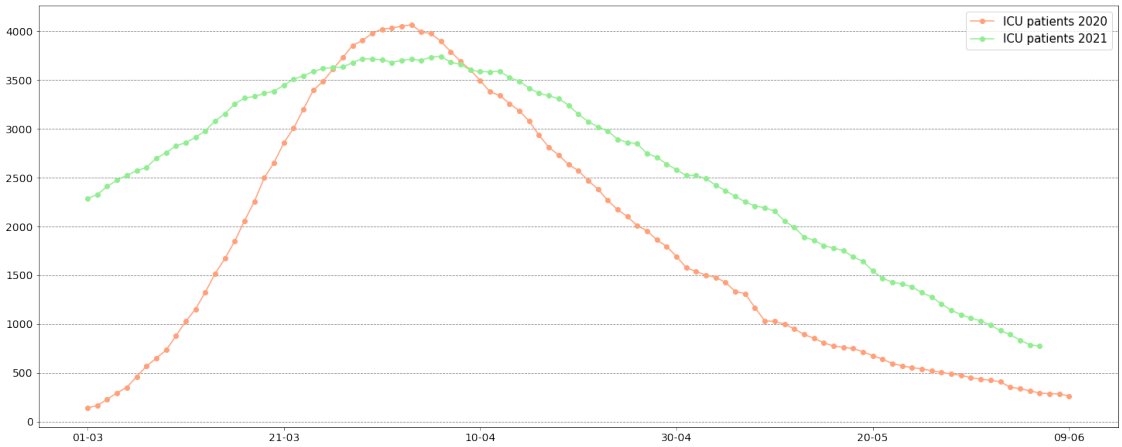


Figure 13: ICU patients comparison of 2020 and 2021

A final graph that allows us to understand how the situation is managed, but above all to put forward hypotheses on the future of the virus, is the reproduction rate (R).
The R number is a way of rating coronavirus or any disease's ability to spread. R is the number of people that one infected person will pass on a virus to, on average.
As can be seen in figure 14 in March 2020 the value far exceeded the value 3, in fact the severity of the virus in Italy was strongly underestimated. The lack of masks and social distancing then led the country to drastic countermeasures. After April, the values of both years have taken on very similar trends.

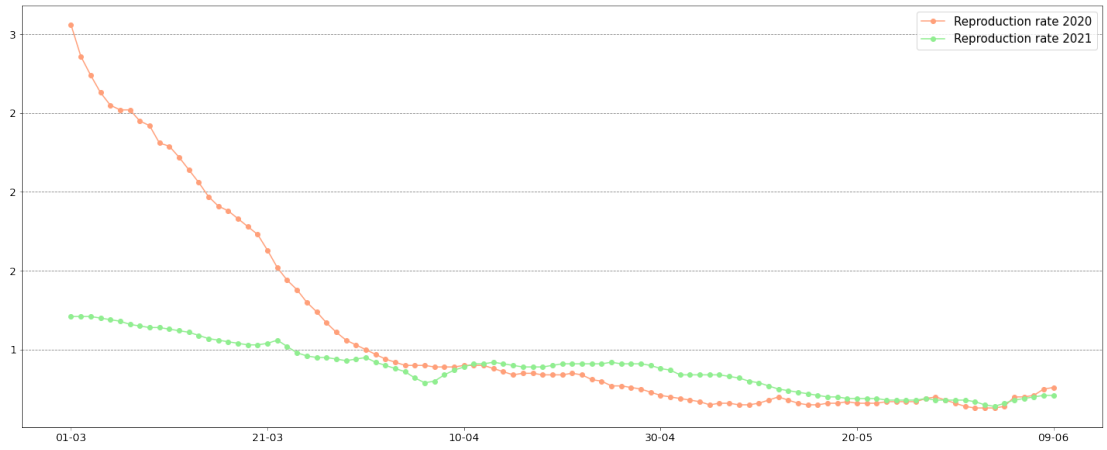


Figure 14: Reproduction rate comparison of 2020 and 2021

It could be briefly concluded that the vaccine is giving in 2021 the same, albeit slightly worse, results than the first lockdown in March 2020.
However, the lockdown would not be sustainable for a long time, because in addition to depriving citizens of living a normal life, it also entails a number of disadvantages.
Many hospital departments have focused more on the emergence of issues related to COVID-19. For this reason, attention may be lacking on many other acute and chronic diseases, especially the rarest ones. This scarcity of interest can cause serious problems or even death. Furthermore, the house confinement can be easily practiced in the first weeks, but a prolonged lack of routine programs like school, working and sport, could increase the occurrence of psychological consequences and distress. Finally, the economic collapse, which occurred in most countries, impacted most of the working classes and the health system.
Instead, the purpose of vaccines is to ensure a gradual resumption of all activities until the herd immunity is reached, which will guarantee indirect protection even for those who do not get vaccinated.

1.2.3 Goal of this thesis

In Italy, as in most other countries, a category very affected by COVID is that of health workers with a positive percentage of about 10% [16]. The Mauriziano hospital of Turin, in May 2020, conducted an analysis through anti-coronavirus serology on healthcare workers, confirming this finding.
The Pfizer mRNA vaccine (Comirnaty) contains a messenger RNA, which as already mentioned induces the synthesis of antigens of the SARS-CoV-2 virus. The S antigens of the virus stimulate the antibody response of the vaccinated person with the production of neutralizing antibodies. The intensity and persistence of the antibody response following the vaccine stimulus has not yet given certain results but could represent an important element in recognizing the level of protection, and could be influenced by general clinical factors (age, physical characteristics or therapies in course).
In addition, the digital COVID certificate came into force in Europe starting from 1st July 2021. It can be issued to all EU citizens and residents and verified throughout the EU, valid for 9 months from the second dose. In Italy, at the end of August 2021 an amendment that extends the validity of the green pass to 12 months was approved. The extension should also be guaranteed to those recovered who have received a single dose of the vaccine. Furthermore, understanding the decrease in the amount of S antigens over a long period would allow us to understand if it will be necessary to repeat the vaccination in certain categories of people, before the epidemic ends.
For this reason, at the end of 2020, the Mauriziano hospital started a second

study concerning the medical staff employed, who carried out both vaccine doses.
The main objectives of the study are:

- the analysis of the IgG anti-RBD Spike antibody positivity rate for SARS-CoV-2 after 1 month (±2 week) from the day of the second vaccination dose (to study the maximum intensity of the response);
- the analysis of the IgG anti-RBD Spike antibody positivity rate for SARS-CoV-2 after 3 months (±2 week) from the day of the second vaccination dose (to analyze the persistence of the response);
- correlation between IgG levels and other information requested in the questionnaire proposed to the medical staff.

The use of machine learning models has already been widely used in modeling and forecasting of epidemiological phenomena.
The prediction of the number of antibodies S that each individual will develop in the two time frames will be carried out through the application of different regression models.
The chosen models will be applied to the dataset cleaned from all outliers, comparing the results.
Furthermore, the aim of this work is focused on discovering which are the most important and incisive features on the development of antibodies.

Chapter 2

Data and methods

Having introduced the goal of this thesis, this chapter describes the method used for data collection. In addition, the regression algorithms used are presented and compared.

2.1 Data collected

The data collection involved 1340 people including health workers and administrators of the AO Mauriziano Order of Turin. They were asked to give a 9 ml blood sample in a serum tube after 1 month and 3 months after the second dose of SARS-CoV-2 vaccine. Following this, the subjects had to fill in a short anonymous questionnaire by accessing a web-based platform for the collection of:

- personal data (gender, date of birth, therapies in progress);
- work activities carried out in the last period (COVID departments, Non-COVID services);
- SARS-CoV-2 infection (asymptomatic, symptomatic) and contacts with positive patients and parents;
- vaccinations date and symptoms.

Before starting the survey, participants had to give their consent for the analysis of their data.
All data relating to the results of blood tests and completed questionnaires were collected in an Excel data sheet. A first dataset was drawn up in February 2021 and subsequently updated with the data of the second blood test in May 2021.

2.1.1 Variables

The February 2021 dataset consists of a sample of 1340 units representing all the volunteers who participated in the Mauritian hospital study.
The variables that make it up are 53: the first 9 relate to the personal data of the volunteer, 11 refer to the profession and work activities carried out in the last period, the following 17 clarify the patient's position with respect to the COVID indicating the contacts at risk and possible infection and the last 16 concern vaccinations against the virus and any symptoms and consequences. The list and description of all the variables present in the dataset is shown in appendix A. Of these, 31 were chosen and used later in this study, they have been renamed and are listed in Table 2:

Variable name	Variable description
birth_date	date of birth of the patient
therapies?	some ongoing therapies?
list_of_therapies	list of the ongoing therapies
flu_shot?	had a flu shot in the last year?
flu_shot_date	date of flu shot
covid_assistance	type of assistance since May 2020
covid_contacts?	contacts at risk for COVID on the workplace?
relatives_covid	positive relatives or cohabitants?
symptoms?	symptoms related to COVID since May 2020?
list_of_symptoms	list of COVID similar symptoms
nasal_swab?	has he been swab for COVID?
positive_swab?	result of the swab
serology?	did he have serology for COVID?
positive_serology?	result of the serology
hospitalization?	has he been hospitalized for COVID?
vaccinated?	has he been vaccinated for COVID?
first_dose_vaccine	date of first vaccine dose
last_dose_vaccine	date of last vaccine dose
first_effect_vaccine?	any side effects after the first vaccination dose?
first_side_effect	which side effect after the first?
second_effect_vaccine?	any side effects after the second vaccination dose?
second_side_effect	which side effect after the second?
positive_after_vaccine?	did he develop a COVID infection after vaccination?
blood_sample_date	date of first blood sampling
IgG_anti_S	amount of antibody IgG anti S after a month
IgG_anti_N	amount of antibody IgG anti N after a month
IgM_anti_S	amount of antibody IgM anti S after a month

Table 2: Features chosen from the February questionnaire

The second dataset was received at the end of May 2021, with the data a second questionnaire and the results of the second blood test.
All the volunteers, in the second questionnaire, are also asked the same questions present in the first, relating to the work activities carried out, the contacts they had with sick people and the symptoms referable to COVID, starting from February 2021. At the same time, additional data such as gender, height, weight and smoking habits were also provided by the Mauritian doctors. These were added in the drafting of both the first and second questionnaires.
Since the goal of this paper is the prediction of the S antibodies trend, the only added features are those shown in Table 3:

Variable name	Variable description
gender	sex of the patient
height	height of the patient
weight	weight of the patient
smoker	is the patient a smoker?
blood_sample_date_2	date of second blood sampling
IgG_anti_S_2	amount of antibody IgG anti S after 3 month
IgG_anti_N_2	amount of antibody IgG anti N after 3 month
IgM_anti_S_2	amount of antibody IgM anti S after 3 month

Table 3: Features added from the May questionnaire and provided by Mauritian doctors

2.1.2 Summary statistics variables

As already mentioned, the questionnaire was submitted to all the volunteers of the Mauritian Hospital in Turin, including both doctors and nurses and administrative staff. The age range varies from 20 to over 70, as shown in the bar chart in Figure 15.

The "*age*" variable was calculated with reference to 1st January 2021.
Almost 40% of the participants belong to the age range between 50 and 60 years. In addition, the 5 people over 70 are retired chief medical officers.
Of the 1340 people, not all reported their gender. Figure 16 shows how women are the majority with 68.3% corresponding to 791 people and men are 31.7% i.e. 367 people.

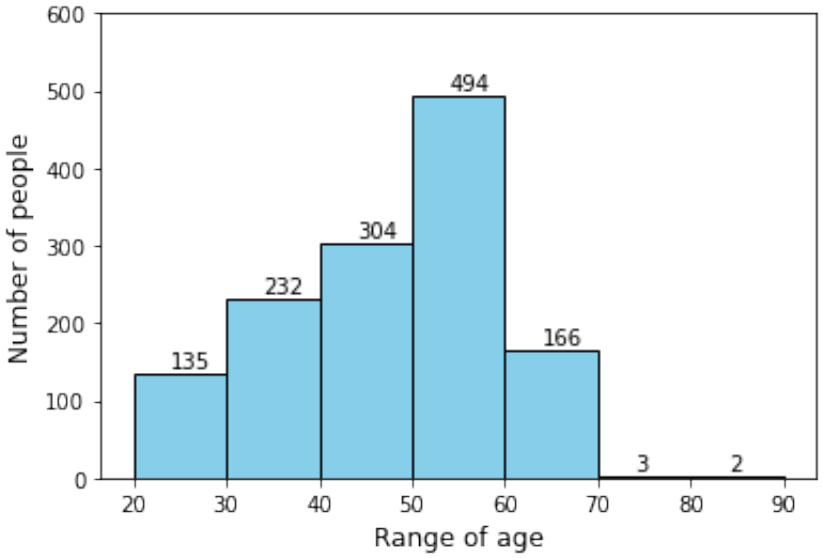


Figure 15: Age distribution histogram

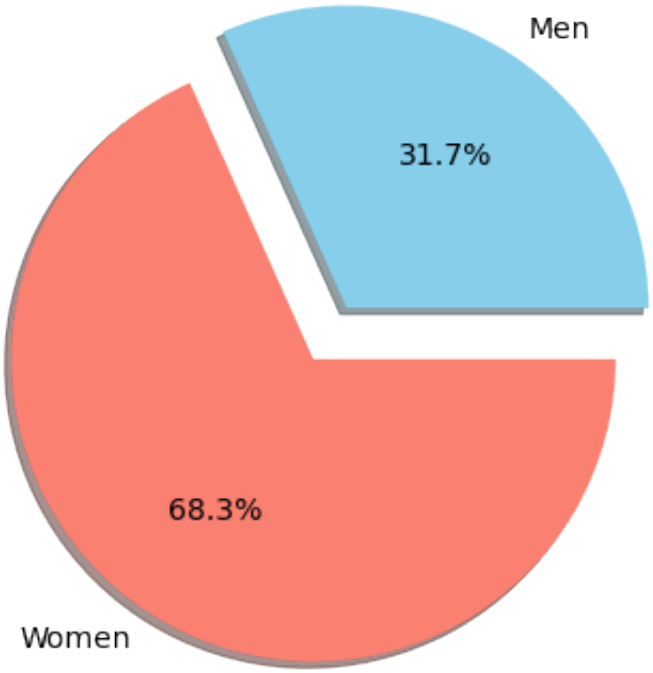


Figure 16: Participation by gender pie chart

The variable *"therapies?"* indicates that the volunteer takes drugs or regularly carries out treatments to deal with some pathology. While the variable *"flu_shot?"* indicates whether they have had the flu vaccination since May 2020. Both of these statistics are represented in an age distribution histogram in Figure 17.
In both graphs, the total of the area subtended by the red squares represents the number of people who filled out the questionnaire, divided into age groups of 10 years. While the blue areas, which are part of the red ones, represent the number of people who, among them, have ongoing therapies and have had the flu shot. Above the blue squares there is the percentage for each of these age groups.

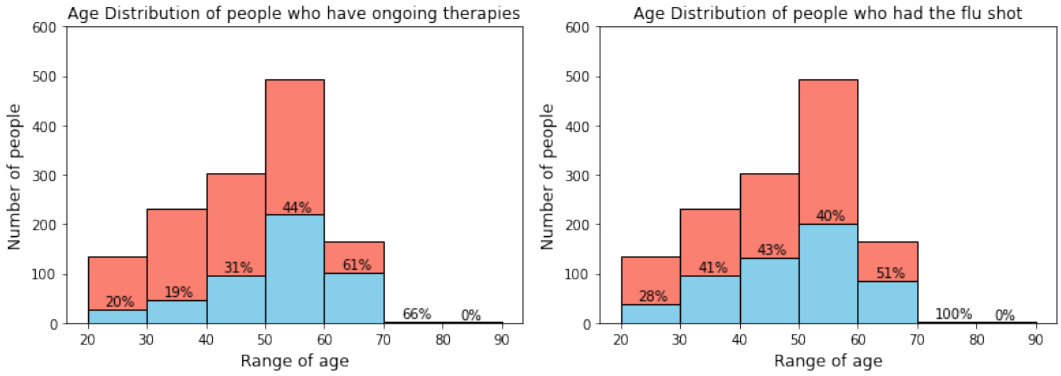


Figure 17: Age distribution histograms based on ongoing therapies and flu vaccinations

2.2 Regression algorithms review

This section of the thesis presents the regression algorithms used in the next chapter: linear regression (Ridge, Lasso and Elastic-net) and Gaussian process regression.

2.2.1 Linear regression models

Linear regression models are the most used machine learning techniques [17], used for finding linear relationship between the target value, also called **regressand** and one or more predictor variables, called **regressors**.
There are two main types of linear regression:

- **Simple linear regression**
  Simple Linear Regression (SLR) can only be used to find a relationship between two variables. Graphically it can be represented as a straight line on a two-dimensional plane.
  An SLR is mathematically representable by a linear function in one variable:

$$f(x) = mx + b \tag{2.1}$$

- **Multiple linear regression**
  Multiple Linear Regression (MLR) is an extension of the SLR, in which more than one regressor (or independent variables) are considered to predict the dependent variable. A general model for multiple linear regression is:

$$Y_i = \omega_0 + \omega_1 X_{i1} + \omega_2 X_{i2} + \ldots \omega_p X_{ip} + \epsilon_i \tag{2.2}$$

  where:
  $i = 1, \ldots, n$
  n is the number of observations;
  p is the number of independent variables chosen;
  ω is the weight associated to each variable

In the specific case of this paper, the need is clearly to solve a multiple regression problem. The linear regression algorithms implemented are described below.
Ordinary Least Squares (OLS) is the most common linear model and its purpose is to find the weights such that the prediction obtained from the regressors comes as close to the data as possible, as described in [18]. The weights are the ones that minimize the quadratic sum of the distances between the observed data and those estimated by the approximation, as explained in function (2.3):

$$\min_{\omega} \parallel X\omega - y \parallel^2 \tag{2.3}$$

where y indicates the observed values and is represented by a column vector $\in \mathbb{R}^n$, where n is the number of observations, X is the input matrix of size $\mathbb{R}^{nxp}$ where p is the number of features passed to the model.
The dataset, using linear regression, must be divided into training and test sets, calculating the scores on both can give us a rough idea about whether the model is good for our data.
Such a model may not always be suitable for what we want to estimate. If the chosen dataset has few features and the test and training results are not good, the OLS will incur under-fitting. Otherwise, with a large number of features and a test score much lower than the training one, the OLS will incur over-fitting.
Ridge and Lasso regression aim to prevent over-fitting by slightly increase the model complexity.

Ridge regression

Ridge regression is a regularized version of linear regression, adding a penalty factor α to the cost function [19].
α penalties minimize the residual sum of squares:

$$\min_{\omega} \parallel X\omega - y \parallel^2 + \alpha \parallel \omega \parallel^2 \tag{2.4}$$

This type of regression analysis method is also called **L2 regularization** because it penalizes the square of the value of the model coefficients.
α varies between 0 and $+\infty$ and penalizes the weight of each feature depending on how much the penalty must be accentuated. None of the coefficients are cancelled.
The α parameter defines a phenomenon called "feature shrinkage": the closer the α value is to 0, the greater the robustness of the coefficient to collinearity.

Lasso regression

Least Absolute Shrinkage and Selection Operator Regression (LASSO) regression concerns the management of features of minor importance. Unlike Ridge regression, which minimizes more or less the weight of some features and reduces their contribution to the model, Lasso regression also tends to reduce the number of independent variables used (feature selection), bringing the others to zero, as described in [20].
Lasso regression minimizes:

$$\min_{\omega} \frac{1}{2n} \parallel X\omega - y \parallel^2 + \alpha \parallel \omega \parallel \tag{2.5}$$

Where α has a fixed value and the learning algorithm is forced to keep the weights $\| \omega \|$ as low as possible. This method is also called **L1 regularization** since it adds the "absolute value of magnitude" of coefficient as penalty term to the loss function.

Lasso generally fits better than Ridge regression when:

- $n > p$
- there are low collinearities among the predictors

Elastic-Net regression

Elastic-Net is a linear regression model that includes L1 and L2 regularization [21] and combines their advantages:

- The L1 part of the penalty generates a sparse model (when the number of samples n is less than the number of features p);
- The L2 part of the penalty:
  - has no limitation on the number of selected variables;
  - promotes grouping effect that is, if a chosen variable is highly correlated to a group, all the variables in the group are automatically added into the model;
  - stabilizes the L1 penalty.

The objective function in this case becomes:

$$\min_{\omega} \frac{1}{2n} \| X\omega - y \|^2 + \alpha\rho \| \omega \| + \frac{\alpha(1-\rho)}{2} \| \omega \|^2 \tag{2.6}$$

Also in the case of Elastic-Net, cross validation is used to set the parameters: α and ρ, also called $l1\_ratio$.
α is the amount of penalization and ρ is the compromise between L1 and L2 penalization both chosen after the cross validation.
Figure 18 shows a comparison among the level curves of the regularization term of the 3 linear models described.

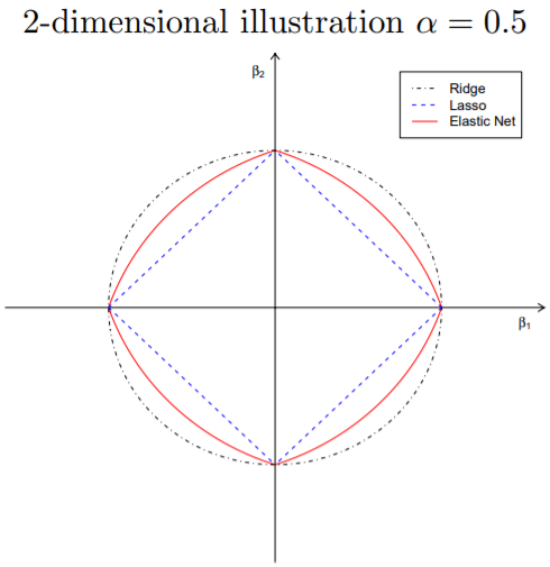


Figure 18: Geometry of Elastic-Net, Ridge and Lasso
Image taken from [4]

2.2.2 Gaussian Process Regression

Gaussian Process Regression (GPR) is a stochastic non-parametric (i.e. not limited by a functional form) Bayesian approache which is computationally simple. [22] They work well with small datasets and the prediction is Gaussian, hence, it allows to take into consideration all the confidence intervals to analyze if they fit the model and if not, refit that region of interest.
Unlike linear models which establish a coefficient for each variable, a Bayesian approach worked out a probability distribution for each variable.
The purpose is therefore to calculate the prior distribution referred to the ω coefficient with reference to the observed data. In this way the distribution to be calculated will be called posterior distribution $p(\omega|y, X)$ and is calculated with the Bayes rule:

$$p(\omega|y, X) = \frac{p(y|X, \omega)p(\omega)}{p(y|X)} \tag{2.7}$$

Where:
$p(y|X, \omega)$ is the likelihood;
$p(\omega)$ is the prior;
$p(y|X)$ is the marginal likelihood.
By weighing all the possible posterior distributions it is possible to calculate the

predictive distribution, which will estimate the prediction at the new points of interest x^*.

$$p(f^*|x^*, y, X) = \int_\omega p(f^*|x^*, \omega)p(\omega|y, X)d\omega \tag{2.8}$$

The first step of the GPR is to define the prior mean $m(x)$ and the covariance function $k(x, x')$:

$$m(x) = \mathbb{E}[f(x)] \tag{2.9}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{2.10}$$

where $\mathbb{E}$ represents the expectation. These functions allow to identify a Gaussian Process as:

$$f(x) \sim GP(m(x), k(x, x')) \tag{2.11}$$

The dataset is divided into test and training subsets where X is the observation matrix and y are the variables.
Assuming now that y is the set of training outputs and f_* a set of test outputs, the prior is defined as:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right) \tag{2.12}$$

Where $K(X, X)$ is a matrix whose i, j^{th} element is equal to $k(x_i, x_j)$, $K(X, X^*)$ is a column vector whose ith element is equal to $k(x_i; x*)$.
The covariance function can also be called **kernel** that is used to compute the GPR's covariance between pairs of datapoints. The implemented kernel functions can be controlled through some hyperparameters in order to better adapt to different functions.
One of the most used kernels for GPR is Radial-basis function (RBF) kernel which is a stationary kernel. It is also known as the "square exponential" kernel. It is parameterized by a length scale parameter l, which can be a scalar or vector with the same number of dimensions as the inputs.
This was chosen given the low complexity of the dataset used and can be represented mathematically with the formula:

$$k_y(x_p, x_q) = \sigma_f^2 exp\left(-\frac{1}{2l^2}(x_p - x_q)^2 \right) + \sigma_n^2 \delta_{pq} \tag{2.13}$$

Chapter 3

Results and critical issues

This chapter proposes the results following an in-depth study on the analyzed data. In the first section some variables will be analyzed in relation to the onset of the COVID or to the antibody values developed by volunteers. This operation can help to get a better idea of how certain variables affect the development of these antibodies.
Finally the results of the supervised learning algorithms described in chapter 2 are reported.
To carry out all these operations, Python was used, a high-level programming language that allows, by importing numerous libraries, to work better on the data available.
Table 4 shows the list of libraries used:

Python library	**Use**
Pandas	Data analysis and manipulation
Datetime	Dates and times manipulation
Numpy	Numerical computing
Seaborn	Data visualization
Matplotlib	Data visualization
Sklearn	Predictive data analysis

Table 4: Python libraries imported

The algorithms used were those proposed by the open source scikit-learn platform [23]. Before applying the chosen models, the datasets have been cleaned up, deleting all invalid entries such as missing or invalid data and shuffled.
Subsequently, the datasets were divided into training for 70% and test sets for 30%.
After this procedure the data were preprocessed, categorical variables were

transformed to 0 or 1 and Numerical variables were normalized using "Min-Max normalization" [24]. It is to be preferred over standardization in cases where the distribution of the data is not known or when it is known that the distribution is not Gaussian as in this case.
The formula is as follows:

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

This was necessary to have all the data on a scale from 0 to 1.
Results are presented in the same way for both the first and the second dataset. For each algorithm a scatterplot is proposed which represents the comparison between predicted and real data. The same graph also shows two lines, one dashed which represents the line of perfect prediction and the second which is the regression line based on the points shown. In addition, for each model there is a bar graph showing the weight "ω" associated with each variable in the regression equation, representing the importance of each of them.
Finally, to univocally compare the performance of each algorithm, the following evaluation parameters will be used: Root Mean Square Error (RMSE), R-squared and variance, defined as follows.

- **root mean square error (RMSE)**: As the name implies is the square root of the mean of the square of all errors. It is a great metric for measuring the accuracy of a model. It is scale dependent, which means that it measures the prediction error for a particular variable and not between variables. The formula for calculating RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\bar{y_i} - y_i)^2}{N}} \tag{3.2}$$

  where:
  $\bar{y_i}$ = Predicted value for the i^{th} observation;
  y_i = Observed value for the i^{th} observation;
  N = Total number of observations.

- **R-squared (R^2)**: It is a statistical measure that represents how well a regression model can adapt. The maximum value is 1, the closer the R-square value is to 1, the better the model fits.
  R^2 is calculated with the formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.3}$$

  where:
  SS_{res} = Mean square error;
  SS_{tot} = Total sum of squares.

- **Variance (σ^2)**: It is defined as the sum of the squared distances of each term in the distribution from the mean μ, divided by the number of terms in the distribution N [25]. These measures are useful for making comparisons between data sets that go beyond simple visual impressions.
  The variance is calculated with the formula:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(\bar{y_i} - \mu)^2}{N} \tag{3.4}$$

  where:
  $\bar{y_i}$ = Predicted value for the i^{th} observation;
  μ = Mean of predicted values.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\bar{y_i} \tag{3.5}$$

3.1 Data analysis

In the period between May and December 2020, 1273 out of a total of 1340 volunteers carried out at least one nasopharyngeal swab for SARS-CoV-2, of these only 216 were positive.
An interesting fact is that 177 people experienced COVID-like symptoms, but only 118 were truly positive, i.e. 67%, looking at the results of the swabs.
Figure 19 shows the different percentages based on the age group.

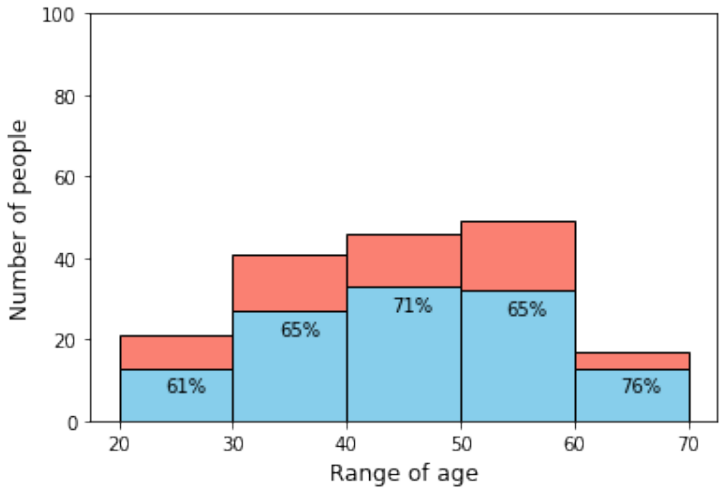


Figure 19: Histogram of subjects with COVID who got a positive swab test

In addition to declaring the hospital ward they work in, each of the participants also had to report what kind of care they have been taking part in since May 2020. There were 4 choices available, with the possibility of inserting multiple answers:

1. High intensity COVID departments
2. Low intensity COVID departments
3. Services for COVID patients
4. 'Non-COVID' departments and services

In the first sub-figure of figure 20 the positivity of the first 3 types of services have been considered, in the second sub-figure only personnel who have never worked in COVID departments are taken into consideration.
It is evident that those who have worked in environments with infected patients have contracted the virus more easily.

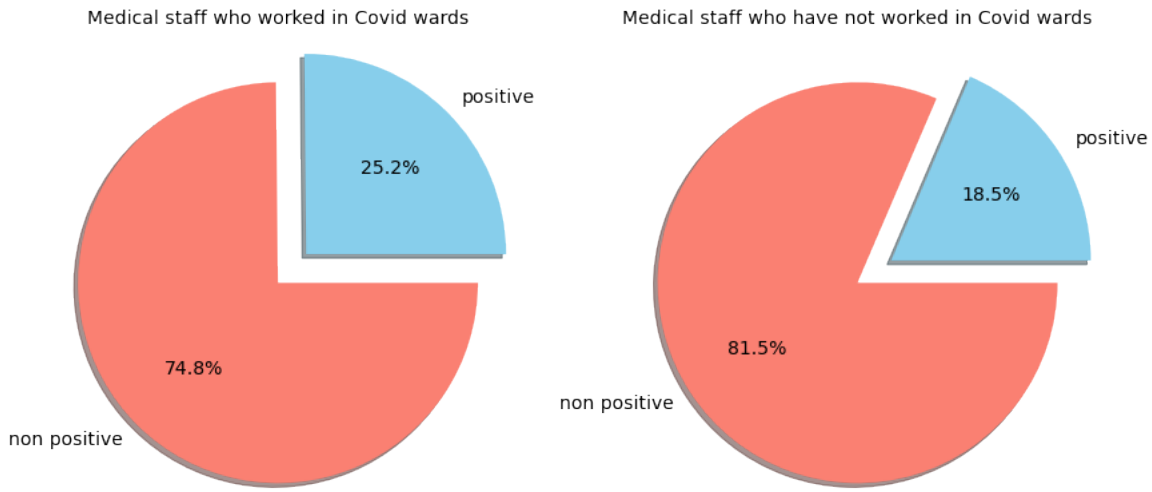


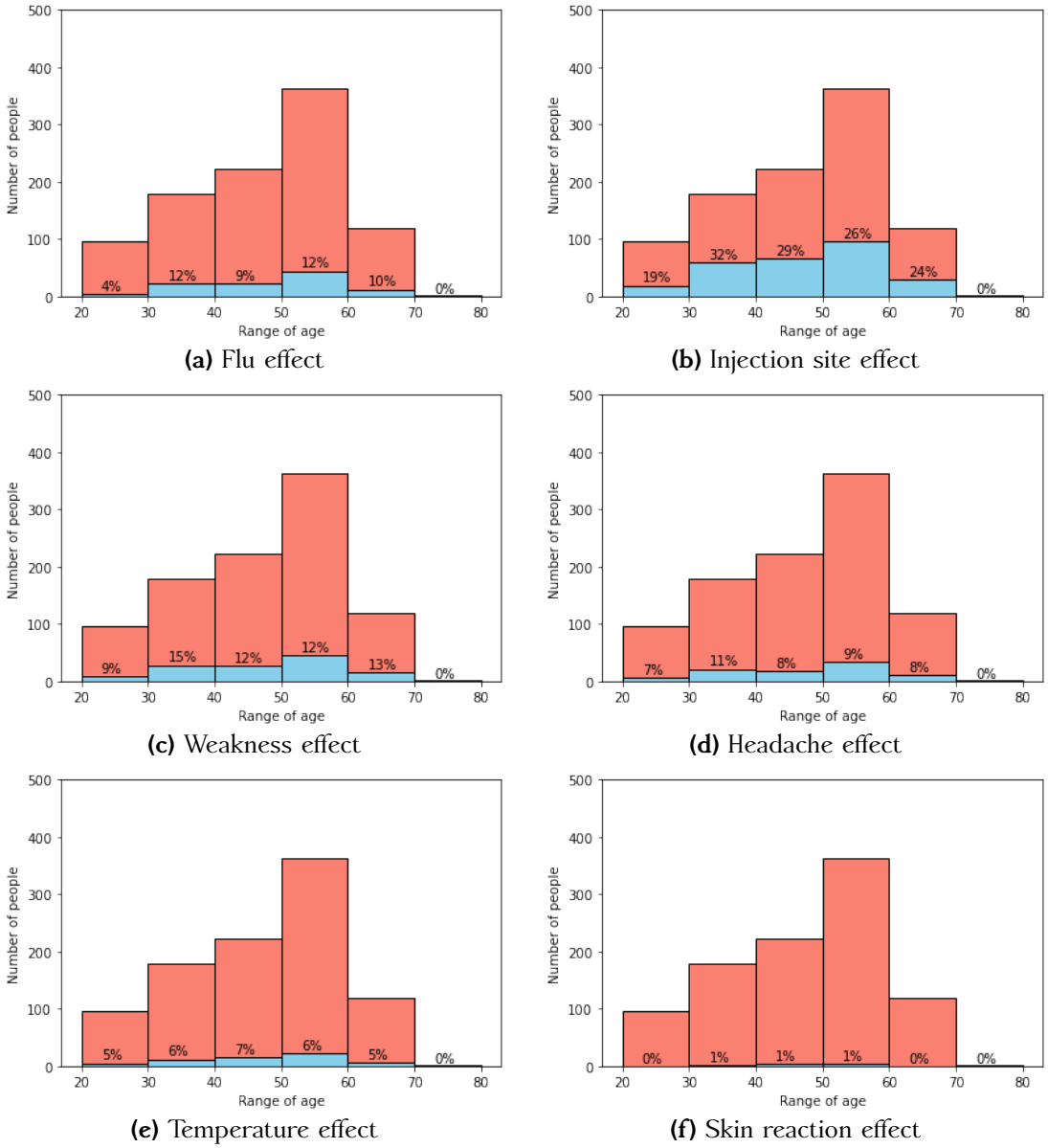
Figure 20: Percentage of positives in COVID and non-COVID departments

3.1.1 Effect of vaccinations

Many people have reported symptoms in the days following vaccinations, organized by age group and symptom type in bar diagrams below.
Comparing the data after the first and second vaccination it is evident that many more people have had symptoms after the latter, this is particularly visible by comparing the pairs of figures: 21a and 22a; 21b and 22b; 21c and 22c; 21d and 22d. In the first questionnaire, they were asked to report if they had had side effects attributable to one or more of the following 6 categories:

- Flu;
- injection site;
- weakness;
- headache;
- temperature;
- skin reactions.

The results are shown in figure 21:



(a) Flu effect

(b) Injection site effect

(c) Weakness effect

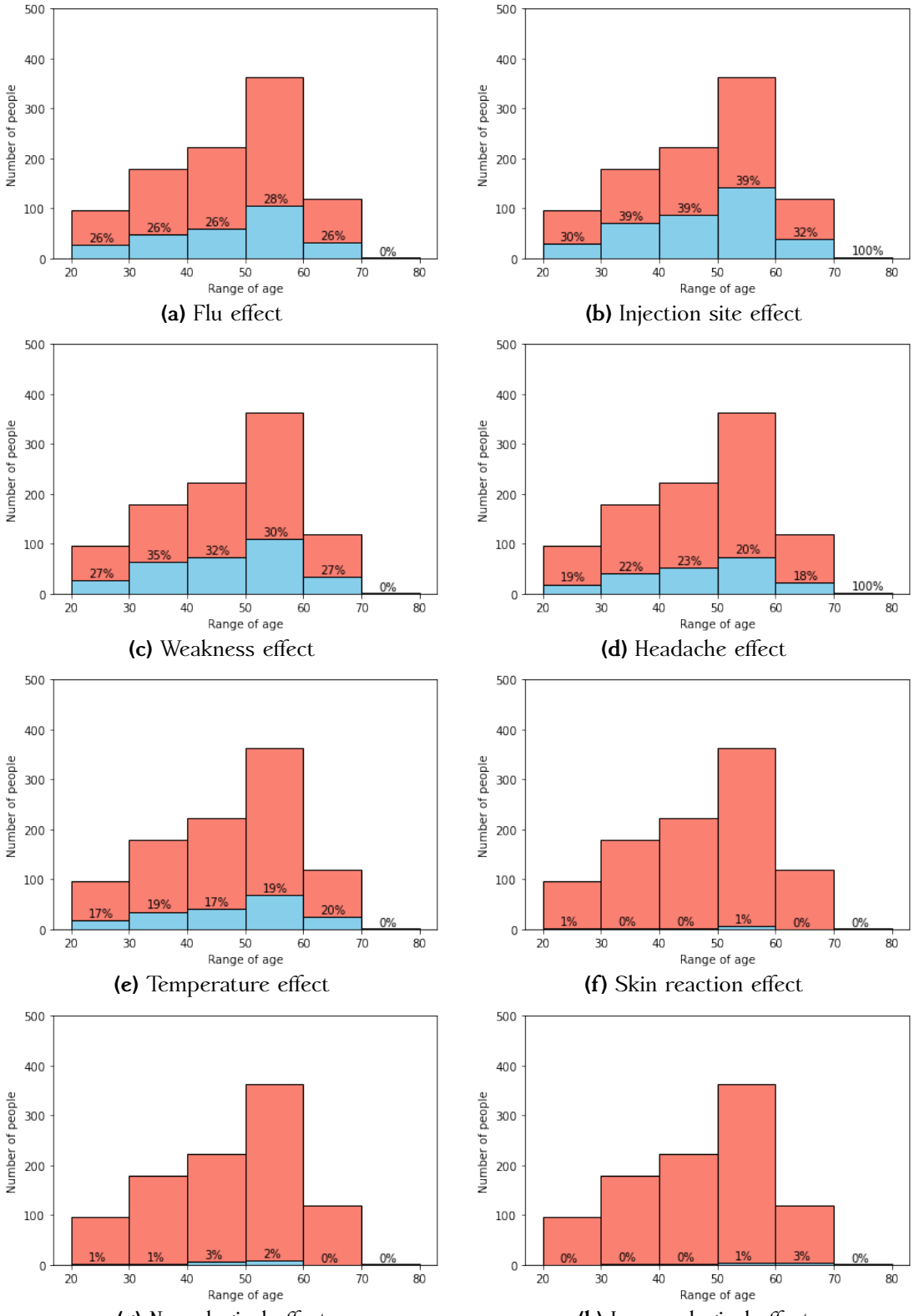
(d) Headache effect

(e) Temperature effect

(f) Skin reaction effect

Figure 21: Side effects after the first dose of vaccine.

In the second questionnaire, two new categories of side effects were added: neurological and immunological.
The results are shown in figure 22.



(a) Flu effect

(b) Injection site effect

(c) Weakness effect

(d) Headache effect

(e) Temperature effect

(f) Skin reaction effect

(g) Neurological effect

(h) Immunological effect

Figure 22: Side effects after the second dose of vaccine.

3.1.2 Correlation between personal data and IgG anti S antibodies

The main purpose of this thesis is to find, if any, a correlation between the variables available and the quantity of antibodies produced. The first blood test was performed approximately 30 days after the second vaccination to evaluate the maximum values of S-type antigens.
The second blood sample was taken about 90 days after the second vaccine injection to evaluate the persistence of S-type antigens.
Subfigures 23a and 23b show two scatterplots representing the values of type S antibodies produced by each individual, after 1 and 3 months.

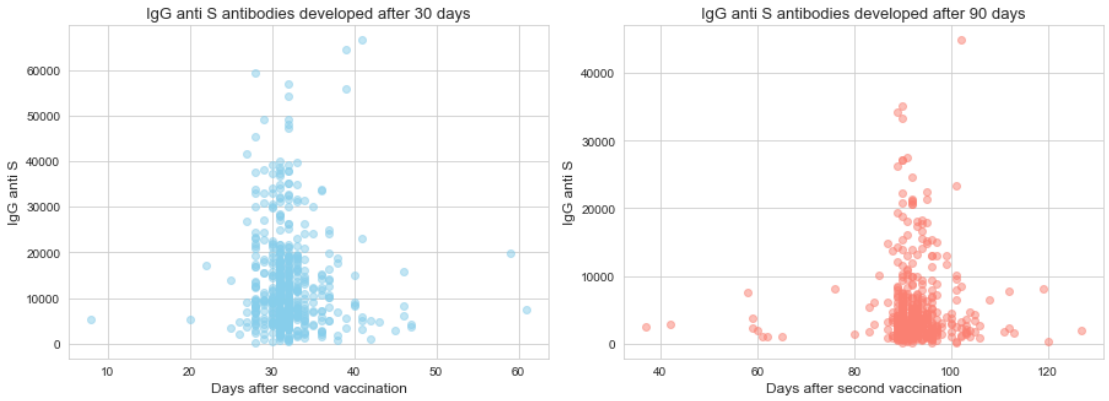


Figure 23: IgG-anti-S antibodies developed 30 and 90 days after vaccination

In the second figure it is easy to see the significant lowering of the values, although in both cases most of the values are concentrated below the 20000 value.
Unfortunately the results obtained after the first blood test did not show a significant correlation with any variables.
As for the second test, it is interesting to show the correlation with three different variables, which have shown noteworthy results.
Figure 24 shows a scatter-plot that links the S antibodies with the Body Mass Index (BMI):

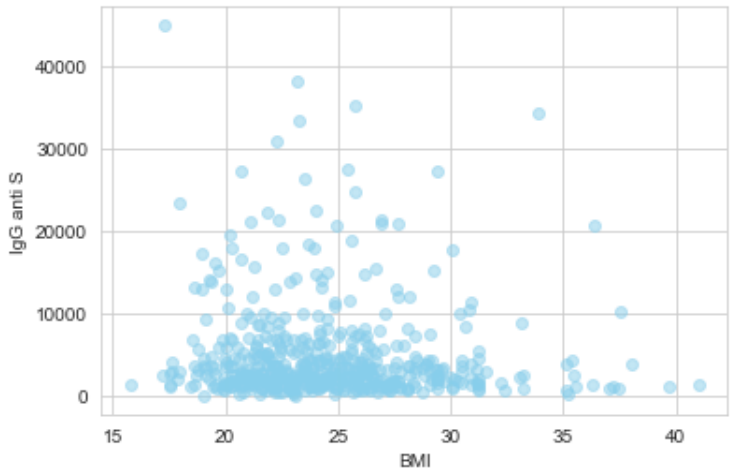


Figure 24: IgG-anti-S vs BMI

Figure 25 instead shows two box-plots that compare the same values based on gender and being a smoker:

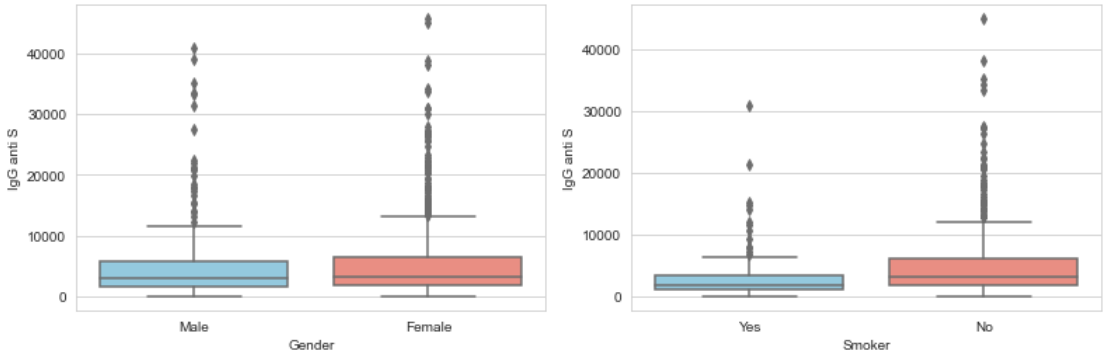


Figure 25: IgG-anti-S vs gender and smoker

While it is clear that body mass index does not have the slightest correlation with an individual's development of antibodies, it appears that being a woman has a slight influence on the permanence of IgG anti S. Being a smoker, on the other hand, is evident how you greatly regulate the amount of antibodies produced and their permanence in the individual.

3.1.3 Data cleaning

In order to make the predictions of the models used as accurate as possible it was necessary to clean the dataset of dirty data.
Table 5 shows the list of all dirty data types divided by variable removed from the two datasets:

Variable name	Dirty data type
vaccinated	Drop those with answer "no"
vaccine_type	Drop non "pfizer" ones
blood_sample_date	Drop NaN values
gender	Drop NaN values
smoker	Drop NaN values
IgG_anti_S	Drop NaN values
BMI	Drop NaN values
relatives_covid	Fill NaN values with 0
positive_swab?	Fill NaN values with 0
positive_serology?	Fill NaN values with 0
positive_after_vaccine?	Fill NaN values with 0

Table 5: Features cleaning

After this operation, both datasets will consist of a total of 538 entries.

3.2 Results - maximum intensity of response

What follows in this section is the result of the analysis of the first dataset, it contains the results of the first blood sample collected between February and March 2021.
With the help of a correlation heatmap it is possible to evaluate the collinearities between each variable. Dependence between two variables is measured using the Pearson correlation coefficient [26] which measures how the value of two different variables vary with respect to each other.
The formula below indicates how it can be calculated:

$$\sum_{i=1}^{N} \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3.6}$$

where N is the number of observations.
The correlation heatmap shown in figure 26 graphically shows the correlation between all the variables taken as input by the first dataset, which are 18.
Each value can vary from -1 to +1. A value close to 1 represents a positive correlation between two variables, a value close to -1 a negative correlation and a 0 indicates that there is no correlation between the two.

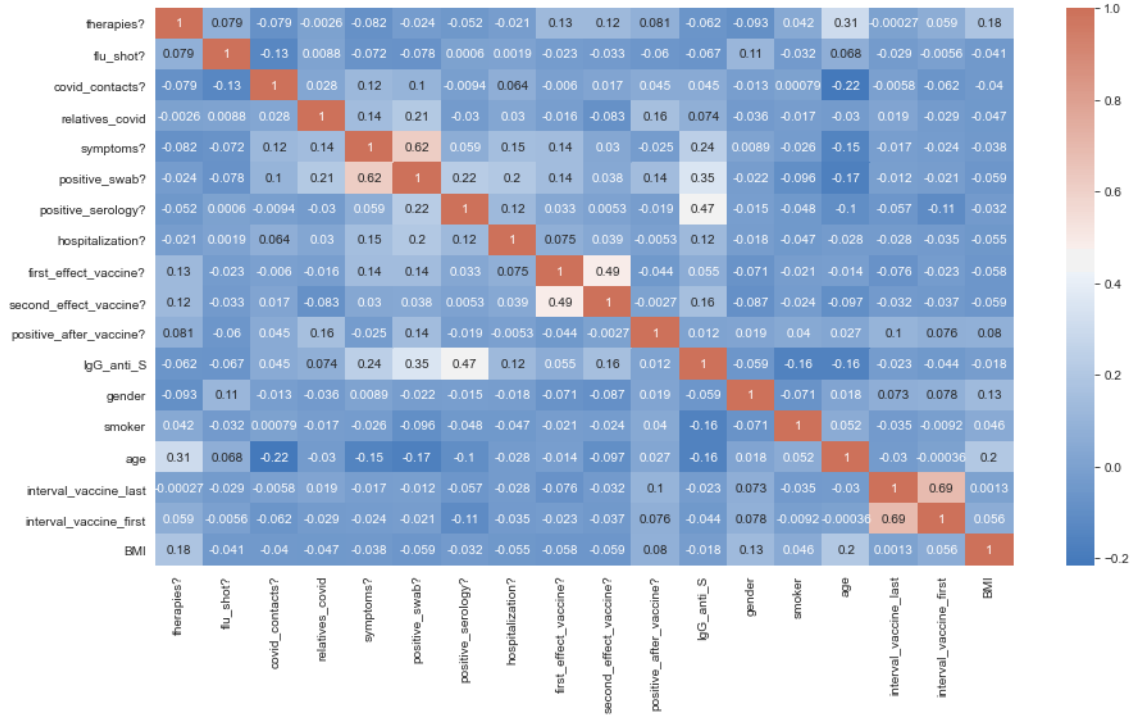


Figure 26: Correlation heatmap for the first dataset

Looking at figure 26, it becomes clear which variables show linear correlation. For example, there is an important positive correlation between "*positive_swab?*" and "*symptoms?*"; this makes sense because as shown above most people who have had symptoms have tested positive for COVID; again, "*positive_swab?*" and "*positive_serology?*" seem related to the value of "*IgG_anti_S*". The "*age*" variable is clearly related to "*therapies?*". Except for these elements, there is no noteworthy sign of collinearity.
The variable chosen as regressand is "*IgG_anti_S*" and is expressed in AU/ml (arbitrary units), as there are currently no international reference standards available for measuring the amount of antibodies produced. The remaining variables will all be used as regressors. The variables "*IgM_anti_S*" and "*IgG_anti_N*" are not used as regressors since they are also obtained from the same blood sample. The need to predict "*IgG_anti_S*" value is useful to understand if some factors influence the maximum antibody development, 30 days after the administration of the second vaccination dose.
Below are the graphs and results for each of the cases.

3.2.1 Ridge regression

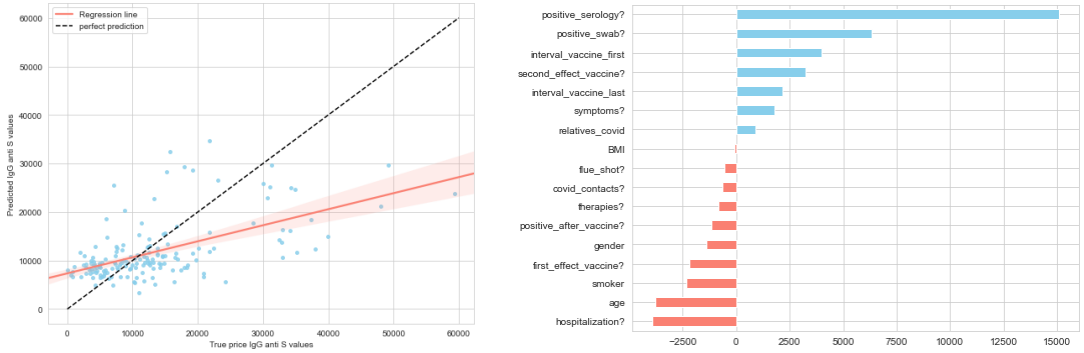
After selecting the model it was necessary to select the "α" parameter that best suited the model. At first the *"RepeatedKFold"* cross-validation method was called [27] from Scikit-learn library, which is a method that repeats the cross-validation several times where in each repetition the folds are divided in a different way.
Subsequently, a dictionary was created to vary the parameter searched. Initially it was made to vary from 0 to 100 with step 1, but noticing that the result was 1, the procedure was repeated by varying the cycle from 0 to 1 with steps of 0.01. The search is performed using the *"GridSearchCV"* class [28] which implements a "fit" and a "score" method. After fitting, the best alpha value is printed, which is 0.99.
The best parameter is chosen on the basis of the best value of R^2.
The code below shows the procedure followed more clearly:

```
model = Ridge()
# define model evaluation method
cv = RepeatedKFold(n_splits=10, n_repeats=3, random_state=1)
# define the grid for alpha
grid = dict()
grid['alpha'] = arange(0, 1, 0.01)
# define search
search = GridSearchCV(model, grid, scoring='r2', cv=cv)
# perform the search
results = search.fit(X_train_scaled, y_train)
# summarize
```

```
print('Config: %s' % results.best_params_)
# best parameter fitting
model = Ridge(alpha=0.99)
results = model.fit(X_train_scaled, y_train)
# test data prediction
predicted = model.predict(X_test_scaled)
```

Listing 3.1: Ridge regression Python code

The results with the Ridge regression are shown in figure 27.
As can be seen, in subfigure 27a the perfect regression line (dashed) is far from the best fit line. Furthermore, the printed points are arranged in a scattered way, not being included in the confidence area (in orange).
A confidence interval shows the probability that the estimated parameter falls within a pair of values, in this case, around the calculated regression line. In this case the confidence interval was set at 95%.
Subfigure 27b graphically represents the values of the coefficients "ω" applied to each variable. Values can be positive (in blue) or negative (in red) depending on whether the variable is directly or inversely related to *"IgG_anti_S"*.



(a) Real vs estimated values **(b)** Feature importance

Figure 27: Ridge regression

3.2.2 Lasso regression

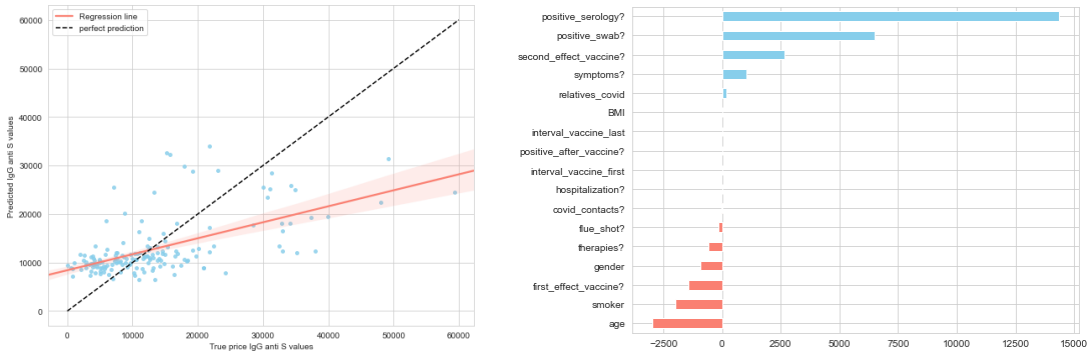
A similar procedure was performed for the Lasso regression. The "α" values were made to vary between 0 and 100 with step 1, because through different running of the code we saw how the model fit better with higher values. In this case the best value was 88.
Code is shown below:

```
# define model evaluation method
cv = RepeatedKFold(n_splits=10, n_repeats=3, random_state=1)
# define model
```

```
model = LassoCV(alphas=arange(0, 100, 1), cv=cv)
# fit model
results = model.fit(X_train_scaled, y_train)
# summarize chosen configuration
print('alpha: %f' % model.alpha_)
# best parameter fitting
model = Lasso(alpha=88)
results = model.fit(X_train_scaled, y_train)
# test data prediction
predicted = model.predict(X_test_scaled)
```

Listing 3.2: Lasso regression Python code

Figure 28 shows the results for the Lasso regression. The regression looks very similar to that seen earlier with the ridge regression, however it can be seen in subfigure 28b how the weight of the less important variables is completely zeroed.



(a) Real vs estimated values **(b)** Feature importance

Figure 28: Lasso regression

3.2.3 Elastic-Net regression

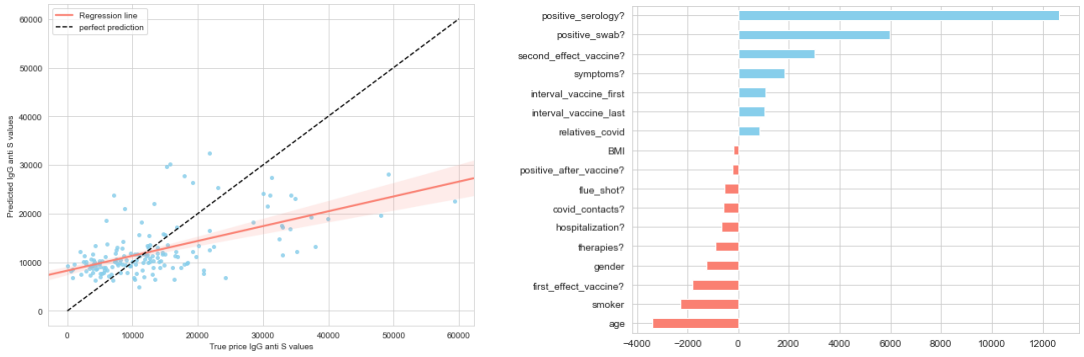
Elastic_Net regression combines the penalties of ridge and lasso regression to get the best of both. In this case "α" is the mixing parameter between ridge (when α = 0) and lasso (when α = 1).
Now, there are two parameters to tune: $l1\_ratio$ and α. The procedure followed was similar to that used for the previous regressions. First of all, cross validation was defined using the *"RepeatedKFold"* class. Subsequently, two dictionaries were defined with values from 0 to 1 with steps of 0.01, for the two different parameters searched.
The model used for the training, created specifically for the Elastic-Net regression is called *"ElasticNetCV"* [29], has returned as results the values: $l1\_ratio$ = 0.98 and α = 0.89.

Finally these values were passed inside the *"ElasticNet"* function [30] to get the results on the test set.
The code below shows what has just been explained:

```
# define model evaluation method
cv = RepeatedKFold(n_splits=10, n_repeats=3, random_state=1)
# define grid for alpha and lambda
ratios = arange(0, 1, 0.01)
alphas = arange(0, 1, 0.01)
# define model
model = ElasticNetCV(l1_ratio=ratios, alphas=alphas, cv=cv)
# fit model
model.fit(X_train_scaled, y_train)
# summarize chosen configuration
print('alpha: %f' % model.alpha_)
print('l1_ratio_: %f' % model.l1_ratio_)
# best parameters fitting
model = ElasticNet(alpha=0.89, l1_ratio=0.98)
results = model.fit(X_train_scaled, y_train)
# test data prediction
predicted = model.predict(X_test_scaled)
```

Listing 3.3: Elastic-net regression Python code

Figure 29 shows the results and in subfigure 29b can be seen how effectively the values of the weights are a cross between Ridge and Lasso.



(a) Real vs estimated values **(b)** Feature importance

Figure 29: Elastic-Net regression

3.2.4 Gaussian process regression

Unlike other algorithms, a GPR does not estimate a value, but the probability density of each value.
First the kernel function is defined, in this case it is the product of the RBF kernel and the Costant kernel which scales the magnitude of the other factors.

The first of the three parameters that is passed to both functions is the constant value which defines the covariance, the others two are the lower and upper bound on constant value.
Then the model is defined by calling the function *"GaussianProcessRegressor"* [31] which implements Gaussian processes for regression and normalizes the regressand training data.
After fitting, the importance of each feature was printed using the *"Permutation feature importance"* function from Sklearn [32].
The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature. The permutation importance function calculates the feature importance of estimators for a given dataset. Importance i_j for feature f_j is defined as:

$$i_j = s - \frac{1}{K} \sum_{i=1}^{K} s_{k,j} \tag{3.7}$$

where:
s: Reference score of the model;
j: Feature taken into account;
k: Number of repetition in 1 ... K.

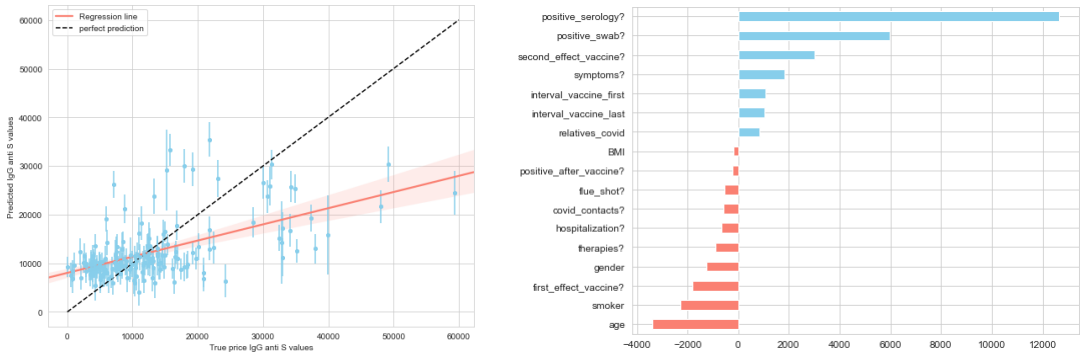
In this case it is not evaluated by the value of the weights but as the difference between the baseline metric, defined as a score and metric from permutating the feature column.
In the last step, the prediction is carried out by extracting the values and the standard deviation of each value.

```
# define kernel function
kernel = C(1.0, (1e-3, 1e3)) * RBF(10, (1e-2, 1e2))
# define model
model = GaussianProcessRegressor(kernel=kernel, alpha=1,
    n_restarts_optimizer=10, normalize_y=True)
# fit model
model.fit(X_train_scaled, y_train)
# print feature importance
imps = permutation_importance(model, X_test_scaled, y_test)
coef = imps.importances_mean
# test data prediction
predicted, sigma = model.predict(X_test_scaled, return_std=True)
```

Listing 3.4: Gaussian process regression Python code

The kernel used is the composition of the constant kernel (C) with the radial basis function (RBF) kernel. Then are also specified, the initial value and the

limits on their hyperparameters.
After specifying the kernel function, the model features were specified. For example, α is the noise variance on the labels and it was fixed at 1 because with lower values the model did not fit well. The number of restarts of the optimizer is used for finding the kernel's parameters which maximize the log-marginal likelihood and was set to 10.
In subfigure 30a the graph is different from the previous ones, each dot represents the estimated average value while the vertical dashes are the standard deviations of the predicted values.



(a) Real vs estimated values **(b)** Feature importance

Figure 30: Gaussian process regression

3.2.5 1st dataset: conclusions

Table 5 shows the valuation metrics for each of the methods used:

	Ridge	Lasso	Elastic-Net	GPR
RMSE	8651.11	8378.73	8368.88	8453.84
R^2	0.298	0.341	0.327	0.329
Variance	0.318	0.343	0.334	0.338

Table 6: 1st dataset: evaluation metrics

As can be seen, the Ridge regression obtained the worst results, it is assumed that this result is due to the fact that the few entries passed to the model did not allow the latter to better estimate the weights of the variables.
However, we cannot be satisfied with any result shown by the other algorithms as well. This shows, for the regression techniques used, that the variables used are not sufficiently informative to estimate the number of type S antibodies developed after 1 month.

3.3 Results - persistence of the response

The second dataset contains the results of the second blood sample collected between April and May 2021. Furthermore, the results of the first blood test have been added to it and as in the previous case in figure 31 a second correlation heatmap is shown.

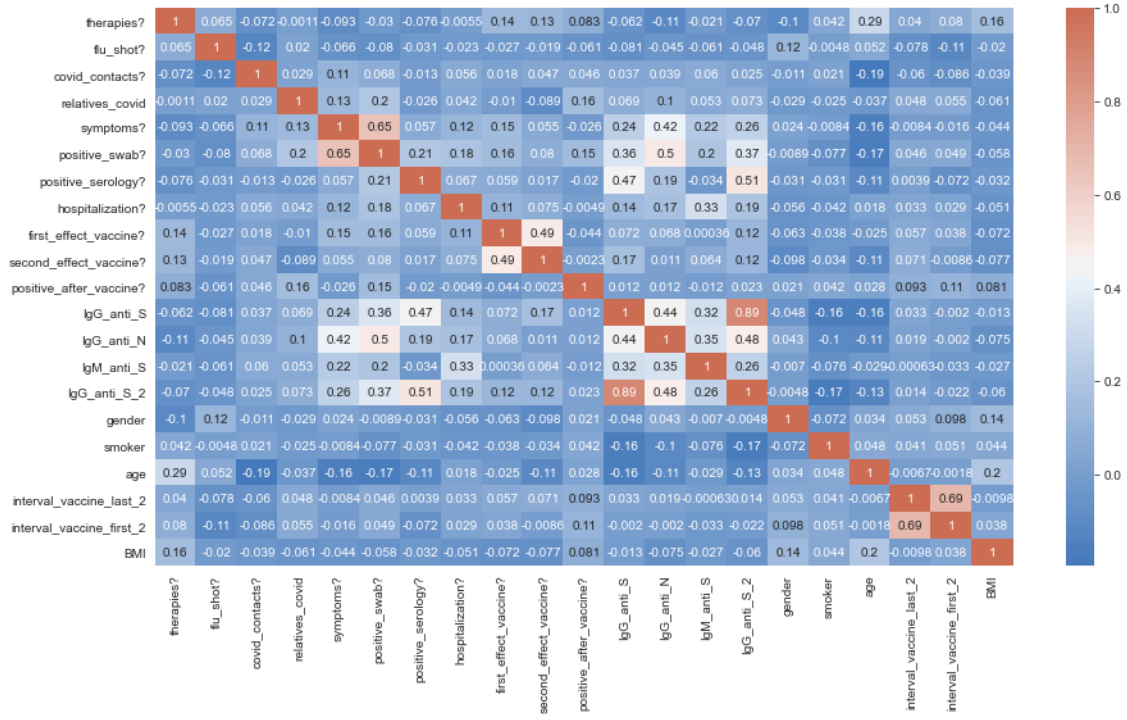


Figure 31: Correlation heatmap for the second dataset

The same correlations highlighted in the previous case are also present in this heatmap, even improved. For example the linear correlation between "*positive_swab?*" and "*symptoms?*" using the first dataset had a value of 0.61 which is now 0.65.
Furthermore, having added the results of the first test, it is immediate to notice how these are extremely correlated with what we have to predict. In particular, the correlation between *"IgG_anti_S"* and *"IgG_anti_S_2"* reaches the value of 0.89, as was to be expected.
The variable chosen as regressand is called *"IgG_anti_S_2"*. The remaining variables will all be used as regressors. Also in this case the variables *"IgM_anti_S_2"* and *"IgG_anti_N_2"* are not used as regressors.
The need to predict this value is useful to understand if some factors influence

the persistence of the number of antibodies 90 days after the administration of the second vaccination dose.
The graph in figure 32 relates the values of *"IgG_anti_S"* and *"IgG_anti_S_2"* this further highlights the lowering of the levels of type S antibodies in each individual. In fact, the values after 60 days from the first blood test are more or less halved.

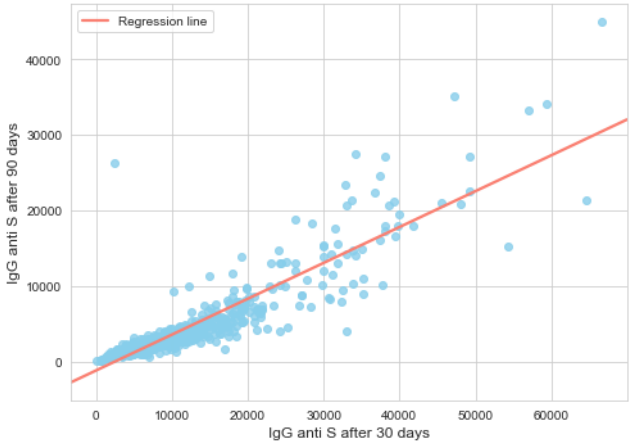
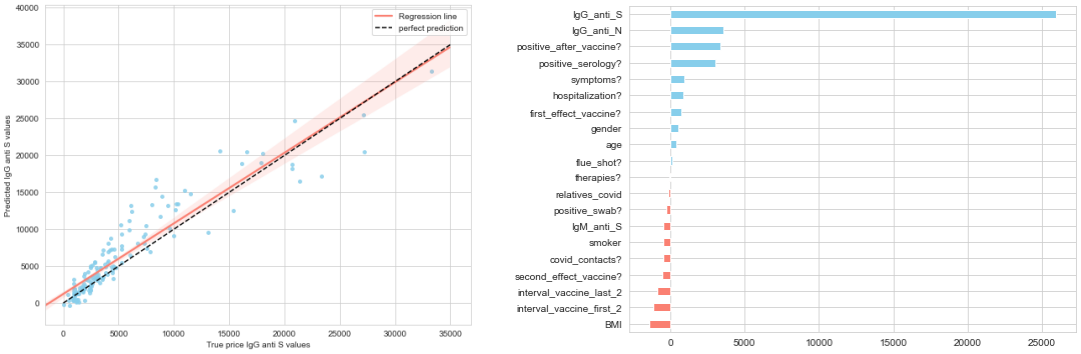


Figure 32: comparison between the values of IgG_anti_S and IgG_anti_S_2

The Python code implemented for the first dataset was used in the same way for this as well.
The next sections include graphs and results.

3.3.1 Ridge regression

After running the gridsearch the best α = 0.42. In subfigure 33a it is evident that the points are largely concentrated around the best fit line, and that the latter is almost superimposed on the perfect regression line. This increase in accuracy is clearly due to the addition of values from the first blood test to the dataset. This is also verifiable in subfigure 33b where the most important variable is exactly what we wanted to predict in the previous section.



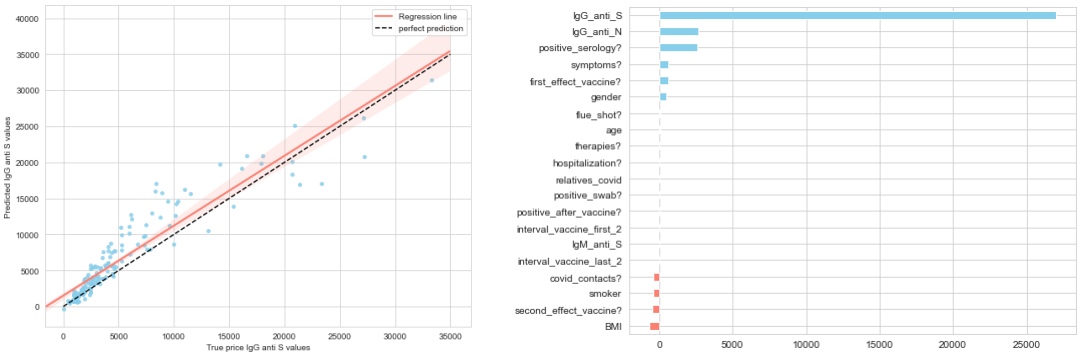
(a) Real vs estimated values **(b)** Feature importance

Figure 33: Ridge regression

Clearly, as already seen above, the value of *"IgG_anti_S"* was found to be by far the most informative of all the variables, however it is unexpected that the most negatively correlated is *"BMI"*, while in figure 24 no correlation was highlighted.

3.3.2 Lasso regression

Even for the Lasso regression the observable results are almost similar. The chosen value of α is 19.



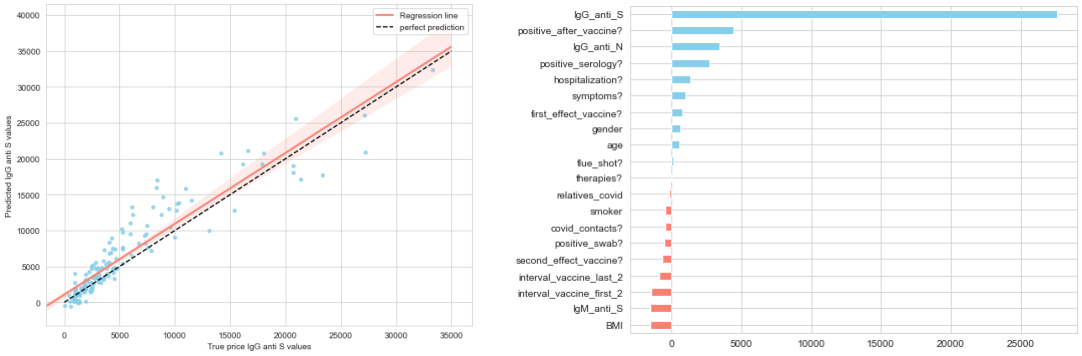
(a) Real vs estimated values **(b)** Feature importance

Figure 34: Lasso regression

As in the previous section, the weight of certain variables has been canceled. Among them there are also some that in the results of the ridge regression added an important contribution, such as: *"positive_after_vaccine?"*, *"hospitalization?"* and *"interval_vaccine_first_2"*.

3.3.3 Elastic-Net regression

To perform Elastic-Net regression the best parameter values are: α = 0.02 and $l1\_ratio$ = 0.98. The values shown in subfigure 35b are somewhere between those of the Ridge and the Lasso.



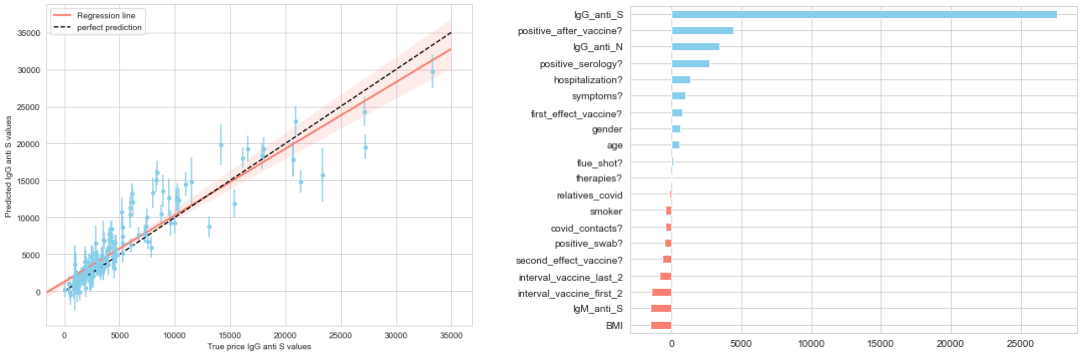
(a) Real vs estimated values **(b)** Feature importance

Figure 35: Elastic-Net regression

In this case the weights of the features vary more gradually than the Ridge regression, moreover this algorithm gives much more importance to the variable *"positive_after_vaccine?"* than to that *"IgG_anti_N"*, unlike the previous cases.

3.3.4 Gaussian process regression

Applying the GPR the same kernel function was used as in the previous section. Also in this case the results are similar to those of the previous models.
Figure 36 shows the results:



(a) Real vs estimated values **(b)** Feature importance

Figure 36: Gaussian process regression

Also in this case the variable *"positive_after_vaccine?"* is considered with greater importance than the first two methods. Furthermore, the lines representing the standard deviation of each point are much smaller than when the same algorithm was used for the first dataset, demonstrating the goodness of this regression.

3.3.5 2nd dataset: conclusions

Table 7 shows the valuation metrics for the second dataset applied to each regression algorithm:

	Ridge	**Lasso**	**Elastic-Net**	**GPR**
RMSE	2436.60	2596.68	2454.69	2398.55
R^2	0.828	0.804	0.825	0.833
Variance	0.857	0.858	0.854	0.850

Table 7: 2^{nd} dataset: evaluation metrics

In this case it is GPR that gets the best results, although they are all very similar. As for the linear regression methods, the one that gave the best results is the Ridge regression. The formula is represented below:

$$\begin{aligned} IgG\_anti\_S\_2 = \omega_1 * IgG\_anti\_S + \omega_2 * IgG\_anti\_N + \dots \\ + \omega_{19} * interval\_vaccine\_first\_2 + \omega_{20} * BMI \end{aligned} \tag{3.8}$$

where ω coefficients values are declared numerically in table 8
The addition of the values of the first blood analysis as regressors made it possible to reach a very high level of accuracy, in particular the *"IgG_anti_S"* feature proved to be by far the most informative in estimating *"IgG_anti_S_2"* values.
The results obtained are certainly positive and with the addition of more features the accuracy could increase further.
Among them, sports activities or the amount of movement that each person does, or still other parameters obtainable from blood and urine tests could be added.

ω	**Feature name**	**Feature importance**
1	IgG_anti_S	25937.81
2	IgG_anti_N	3582.51
3	positive_after_vaccine?	3370.54
4	positive_serology?	3004.71
5	symptoms?	941.86
6	hospitalization?	896.70
7	first_effect_vaccine?	750.81
8	gender	558.92
9	age	417.74
10	flu_shot?	124.41
11	therapies?	-45.76
12	relatives_covid	-134.69
13	positive_swab?	-287.16
14	IgM_anti_S	-465.48
15	smoker	-472.59
16	covid_contacts?	-482.05
17	second_effect_vaccine?	-552.24
18	interval_vaccine_last_2	-894.10
19	interval_vaccine_first_2	-1159.58
20	BMI	-1400.86

Table 8: 2nd dataset: features importance

Chapter 4

Conclusions and future work

4.1 Summary and conclusion

After the discovery in China in early 2020, the SARS-CoV-2 virus has rapidly spread to all countries of the world, causing tens of millions of infections and hundreds of thousands of deaths.
Initially there was no specific medicine or effective treatment for this viral infection.
The COVID-19 pandemic is a global crisis, with devastating health, social and economic impacts and the development of a vaccination was immediately the main goal to be achieved in order to put an end to these limitations.
Another promising treatment and relatively easy to develop are antibody therapies, which use anti-inflammatory drugs to fight the virus. However, both vaccines and antibody therapies are prone to more than 26,000 unique mutations, making the therapy useless in some cases [33].
Pending a safe and effective vaccine, the spread of the virus throughout the year 2020 has been curbed by applying more or less stringent confinement, blocking or closure measures, throughout the Italian territory or in the most affected regions.
As demonstrated in Chapter 1, these restrictive measures were not sufficient. At the end of 2020 the first vaccines were approved and the first administrations also began. After describing the functioning and objectives of vaccinations, in section 1.2.2 some correlations have been shown between the number of vaccinations carried out and some reference parameters.
In this section, making the comparison between the year 2020 and 2021 regarding the trend of some factors resulting from the epidemic in Italy, in a period ranging from 1 March to 10 June, it was possible to reach clear conclusions on vaccines.

Analyzing the daily data of: "positive rate", "deaths", "hospitalized patients", "ICU patients" and "reproduction rate", although the trends for the two different years seem very similar, it is necessary to consider that 2020 was characterized by a long lockdown unlike 2021 where an attempt was made to encourage a gradual reopening of public places and commercial activities.
The purpose of this thesis is to understand through an in-depth analysis if certain variables related to a person's lifestyle or his contacts are correlated with the development of COVID-19 virus antibodies. In addition, we also want to estimate the IgG antibody positivity rate for SARS-CoV-2 1 and 3 months after the second dose of the vaccine.
Analyzing the values of what we want to predict with each of the other variables taken individually, it is evident that the only determining characteristic is being or not a smoker. The latter in fact are subject to a lower development of antibody activity.
Subsequently, two different datasets were created to estimate the parameter in the two different periods. Figures 26 and 31 show two correlation heatmaps related to the two datasets. This allows us to understand at what level each feature is related to all the others. What interests us in particular is the correlation with respect to the variable *"IgG_anti_S"* for the first dataset and *"IgG_anti_S_2"* for the second.
To achieve the main objectives different regression algorithms were applied. Three linear regression models (ridge, lasso and elastic-net) and a non-linear Gaussian model were used and compared.
As reported in [34], linear regression is often one of the first algorithms that data analysts are introduced to. The algorithm finds the best line that fits a given data set and assuming a linear equation proceeds to find the best fitting values. The more the equation chosen is suitable for the dataset used, the better the result.
Instead, a Gaussian Process Regressor let the regressor find the best function for us. Gaussian process regression is nonparametric, so rather than calculating the probability distribution of parameters of a specific function, GPR calculates the probability distribution over all admissible functions that fit the data.
Applying the algorithms to both datasets, the following conclusions were reached:

- for the prediction of the maximum intensity of the response, the result obtained was not satisfactory. In fact, even looking at the heatmap it is clear that none of the variables are particularly correlated with the regressand. However, by printing the weights for each algorithm, it is possible to conclude that being positive for serology or a swab is a determining factor.
- the results obtained to predict the persistence of the response, on the other hand, were much more accurate. All algorithms showed similar behavior

when exceeding an r-squared value greater than 0.80. The addition to the dataset of the *"IgG_anti_S"* values analyzed two months earlier was instrumental in achieving this level of accuracy. This is also demonstrated by the weight values shown.

4.2 Future work

As described in the article: [35], the past few months have seen the rapid spread of numerous variants of SARS-CoV-2, including the Variants Of Concern (VOC) variants established by the ECDC, named according to the Phylogenetic Assignment of Named Global Outbreak (PANGO):

- Alpha (B.1.1.7): also called the "English variant" it was first isolated in September 2020 in Great Britain;
- Beta (B.1.351): also called the "South African variant" it was first isolated in October 2020 in South Africa;
- Gamma (P.1): also called "Brazilian variant" it was isolated for the first time in January 2021 in Brazil and Japan;
- Delta (B.1.617.2): also called "Indian variant" was first isolated in December 2020 in India.

In all cases, the virus has mutations on the so-called 'spike' protein, widely discussed in this paper. These variants are characterized by greater transmissibility, while there is still no certainty regarding the effectiveness of vaccinations on them.
A future study could focus on the analysis of these variants and the impact they can have on vaccinated and unvaccinated people.
An interesting development could be to continue taking blood samples from the people involved in this study to analyze how the parameters can vary between 6 months and 1 year.

Appendix A

A.1 Variables in the self-assessment questionnaire for the serological study of February 2021

1. *Data nascita:* date of birth of the patient
2. *Sesso:* sex of the patient
3. *Peso:* weight of the patient
4. *Altezza:* height of the patient
5. *Fumatore?:* is the patient a smoker?
6. *Terapie abituali:* does the patient carry out any therapy?
7. *Quali terapie?:* list of therapies followed
8. *Ha effettuato vaccinazione antinfluenzale?:* did he get the flu shot?
9. *Data vaccinazione:* flue shot date
10. *Professione:* profession performed in the hospital
11. *Struttura di appartenenza:* department of belonging
12. *Assistenza da maggio 2020 (Possibili risposte multiple):* type of assistance carried out by May 2020
13. *Ha avuto da maggio 2020 contatti sul lavoro a rischio (noti) per COVID?:* have he had contacts at risk (known) for COVID since May 2020 on the workplace?
14. *Se si:* with whom did the contact take place?

15. *Data primo contatto a rischio noto:* first contact date at risk in the workplace
16. *Data ultimo contatto a rischio noto:* last contact date at risk in the workplace
17. *Ha avuto parenti/conviventi affetti da COVID?:* positive relatives or cohabitants?
18. *Se Si/Dubbio:* yes/no/maybe
19. *Data esordio primo parente/convivente:* date of onset of symptoms in the cohabitants
20. *Esordio primo parente/convivente prima o dopo di te?:* did the onset of the relative's symptoms occur before or after the onset of the patient's symptoms?
21. *Ha accusato da maggio 2020 sintomi riferibili a COVID?:* has he experienced symptoms related to COVID since May 2020?
22. *Quali sintomi?:* which symptoms?
23. *Data esordio sintomi:* date onset symptoms
24. *Sintomi risolti?:* have the symptoms been resolved?
25. *Data risoluzione sintomi:* date of resolution of symptoms
26. *È stato sottoposto a tampone per COVID?:* has he been swab for COVID?
27. *Ha avuto un tampone positivo?:* result of the swab
28. *Data ultimo tampone negativo:* last negative swab date
29. *Data primo tampone positivo/dubbio:* date of the first positive swab
30. *Infezione risolta?:* did the infection clear up?
31. *Data primo tampone negativo dopo l'infezione:* date of the first negative swab after the infection
32. *È stato sottoposto a sierologia per COVID?:* did he has serology for COVID?
33. *Se Sì:* for what reason?
34. *Risultato:* result of the serology

35. *Data sierologia:* serology date
36. *È stato ricoverato per COVID?:* has he been hospitalized for COVID?
37. *Data ricovero:* hospitalization date
38. *È stato sottoposto a vaccino per COVID?:* has he been vaccinated for COVID?
39. *Data prima dose di vaccino:* date of first vaccine dose
40. *Data ultima dose di vaccino:* date of last vaccine dose
41. *Tipo di vaccino per COVID?:* type of vaccine
42. *Ha accusato effetti collaterali dopo la prima dose vaccinale?:* did he experience any side effects after the first vaccination dose?
43. *Quali effetti dopo la prima?:* which side effect?
44. *Ha accusato effetti collaterali dopo la seconda dose vaccinale?:* did he experience any side effects after the second vaccination dose?
45. *Quali effetti dopo la seconda?:* which side effect?
46. *Ha sviluppato un'infezione da COVID19 dopo la vaccinazione?:* did he develop a COVID19 infection after vaccination?
47. *Data tampone positivo:* positive swab date
48. *DATA PRELIEVO:* date of first blood sampling
49. *PROV:* vaccine injection site code
50. *RICHIESTA:* vaccination code
51. *IgG anti S:* Amount of antibody IgG anti S after a month
52. *IgG anti N:* Amount of antibody IgG anti N after a month
53. *IgM anti S:* Amount of antibody IgM anti S after a month

Bibliography

[1] ECDC. *Immune responses and immunity to SARS-CoV-2*. URL https://www.ecdc.europa.eu/en/covid-19/latest-evidence/immune-responses.

[2] Hitchings M.D.T. Huang, Garcia-Carreras. *A systematic review of antibody mediated immunity to coronaviruses: kinetics, correlates of protection, and association with severity. Nat Commun*, 2020. DOI: https://doi.org/10.1038/s41467-020-18450-4.

[3] Giulio Gottardo Giampaolo Galli. URL https://osservatoriocpi.unicatt.it/cpi-archivio-studi-e-analisi-perche-l-intensita-della-crisi-economica-e-tanto-diversa-fra-paesi-simili.

[4] Hui Zou and Trevor Hastie. *Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301–320, 2005. DOI: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[5] Il Sole 24 ORE. URL https://lab24.ilsole24ore.com/storia-coronavirus/.

[6] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. *COVID-19 Data Repository*. URL https://github.com/CSSEGISandData/COVID-19.

[7] ARS Toscana. URL https://www.ars.toscana.it/2-articoli/4306-nuovo-coronavirus-punto-vaccino-terapie-covid-19-sars-cov-2-trattamenti-sperimentazione-vaccini-cure.html#trial-terapie-covid-19.

[8] AIFA. URL https://www.aifa.gov.it/sperimentazioni-cliniche-covid-19.

[9] WHO. "solidarity" clinical trial for covid-19 treatments. URL https://www.who.int/emergencies/diseases/novel-coronavirus-

2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments.

[10] University of Oxford. *Dexamethasone reduces death in hospitalised patients*. 2020. URL https://www.ox.ac.uk/news/2020-06-16-dexamethasone-reduces-death-hospitalised-patients-severe-respiratory-complications.

[11] ClinicalTrials.gov. URL https://clinicaltrials.gov/ct2/results?cond=COVID-19.

[12] IRCSS. *Vaccini contro il Covid-19: come funzionano? Che protezione conferiscono?* URL https://www.marionegri.it/magazine/vaccini-anti-covid-19.

[13] Ministero della salute. *Vaccini anti Covid-19*. URL https://www.salute.gov.it/portale/p5_1_1.jsp?lingua=italiano&faqArea=nuovoCoronavirus&id=255.

[14] Hannah Jackson Amanda Connolly. Astrazeneca covid-19 vaccine not recommended for those under 55, naci says. URL https://globalnews.ca/news/7726169/astrazeneca-vaccine-safety-canada/.

[15] Our World In Data. URL https://github.com/owid/covid-19-data.

[16] quotidianosanita.it. URL https://www.quotidianosanita.it/studi-e-analisi/articolo.php?articolo_id=82939.

[17] towards data science. *Linear Regression*. URL https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86.

[18] Understanding the ols method for simple linear regression. URL https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc.

[19] Andrea Provino. Ridge regression, . URL https://andreaprovino.it/ridge-regression/.

[20] Andrea Provino. Lasso regression, . URL https://andreaprovino.it/lasso-regression/.

[21] Andrea Provino. Elastic-net regression, . URL https://andreaprovino.it/elastic-net-early-stopping/.

[22] Hilarie Sit. Quick start to gaussian process regression. URL https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319#:~:text=Gaussian%20process%20regression%20(GPR)%20is,uncertainty%20measurements%20on%20the%20predictions.

[23] Scikit-learn. *Supervised learning*. URL https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

[24] Lorenzo Govoni. L'importanza del ridimensionamento dei dati nei problemi di machine learning. URL https://www.lorenzogovoni.com/ridimensionamento-dei-dati/.

[25] sciencebuddies.org. URL https://www.sciencebuddies.org/science-fair-projects/science-fair/variance-and-standard-deviation.

[26] *Coefficiente di correlazione*. URL https://www.jmp.com/it_it/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html.

[27] *Repeated K-Fold*. URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedKFold.html.

[28] *GridSearchCV*. URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[29] *ElasticNetCV*, . URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html.

[30] *ElasticNet*, . URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html#sklearn.linear_model.ElasticNet.

[31] *GaussianProcessRegressor*. URL https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html.

[32] *sklearn.inspection.permutation_importance*. URL https://scikit-learn.org/stable/modules/generated/sklearn.inspection.permutation_importance.html.

[33] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Prediction and mitigation of mutation threats to covid-19 vaccines and antibody therapies. *Chem. Sci.*, 12:6929–6948, 2021. DOI: 10.1039/D1SC01203G. URL http://dx.doi.org/10.1039/D1SC01203G.

[34] *Exploring Gaussian Process vs Linear Regression*. URL https://blog.davidvassallo.me/2019/03/15/exploring-gaussian-process-vs-linear-regression/.

[35] Finlay Campbell, Brett Archer, Henry Laurenson-Schafer, Yuka Jinnai, Franck Konings, Neale Batra, Boris Pavlin, Katelijn Vandemaele, Maria D Van Kerkhove, Thibaut Jombart, Oliver Morgan, and Olivier le Polain de Waroux. Increased transmissibility and global spread of sars-cov-2 variants of concern as at june 2021. *Eurosurveillance*, 26(24): 2100509, 2021. DOI: https://doi.org/10.2807/1560-7917.ES.2021.26.24.2100509. URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.24.2100509.

Acronyms

AIFA Agenzia italiana del farmaco.

ARS Agenzia Regionale Sanitaria.

BMI Body Mass Index.

ECDC European Centre for Disease Prevention and Control.

EMA European Medicines Agency.

FDA Food and Drug Administration.

GPR Gaussian Process Regression.

ICU Intensive Care Unit.

IRCCS Istituto di Ricovero e Cura a Carattere Scientifico.

LASSO Least Absolute Shrinkage and Selection Operator Regression.

MERS Middle East Respiratory Syndrome.

MLR Multiple Linear Regression.

OLS Ordinary Least Squares.

OWID Our World in Data.

PANGO Phylogenetic Assignment of Named Global Outbreak.

RBF Radial-basis function.

RMSE Root Mean Square Error.

SARS Severe Acute Respiratory Syndrome.

SLR Simple Linear Regression.

VOC Variants Of Concern.

WHO World Health Organization.