POLITECNICO DI TORINO

Master's Degree
in Data Science and Engineering

Thesis

Data Science for predicting SARS-CoV-2 mortality

Supervisor
prof. Roberto Fontana

Author
Maria Francesca Turco

Academic Year 2021-2022

To my parents, for
their love and support.
Also dedicated to the
COVID-19 victims,
fighters and survivors.

ABSTRACT

It has been a little over two years since all of our lives were completely changed after a new, as yet unidentified strain of coronavirus spread to all parts of the world. With the diffusion of the SARS-CoV-2 pandemic, the scientific community took immediate action to first sequence the virus and find drugs that could adequately treat those affected, and then switch to vaccines to prevent the spread of the disease.
Data Science, which has proven to be very reliable in the medical field, played its role in the fight against this pandemic. Using Data Science to predict the probability of death offers a great opportunity to optimize the allocation of medical resources, which is crucial in responding to a large-scale outbreak of an emerging infectious disease such as COVID-19.
The main goal of this work was therefore to develop a Machine Learning model that can identify whether a patient with SARS-CoV-2 is at risk of death. For this purpose, the CDC COVID-19 Case Surveillance dataset was used. This is a very large American private dataset from which ten thousands of COVID-19 patients were extracted to start the research activity. Once the first outcomes were available, we moved on to the second phase, which took into account a total of more than 3 million patients, in order to further expand the analysis and assess the reliability of the initial results. Based on the literature review, the most commonly used and effective algorithms for predicting mortality in people affected by SARS-CoV-2 were selected. These include Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Artificial Neural Network and Bayesian Network. They were then all subjected to a performance evaluation in order to determine which produced the best results. After carrying out several experiments also the most alarming symptoms and patient characteristics that need closer monitoring were detected. The model which has shown the highest accuracy that is 97.20 ± 0.8 was the Random Forest classifier.
Forecasting a patient's mortality with Machine Learning could definitely help the healthcare system in all countries of the world to give more attention and medical care to people who are most at risk. In this way, they will receive appropriate

treatments in a shorter time and there is hope in this manner to reduce the overall mortality rate in the world.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Roberto Fontana for mentoring me throughout the course of this thesis and for providing prompt and helpful feedback.
I am very grateful to the Centres for Disease Control and Prevention (CDC) organization for allowing me to use their restricted dataset. CDC assumes no responsibility for the scientific validity or accuracy of the assumptions and methodology, results, statistical analyses or related findings within this work.
Last but not least, I would like to thank Dr. Maria Stella Gianfiori and all the SAS staff for giving me the opportunity to use a previously unknown environment, namely SAS Viya.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

We should recognize that they reflect things that we're doing, not just things that are happening to us.
We should understand that, although some of the human-caused factors may seem virtually inexorable,others are within our control. [D.Quammen, Spillover]

Chapter 1

Introduction

COVID-19 disease caused by the beta-coronavirus SARS-CoV-2 was first identified in Wuhan, China, in December 2019, and then the World Health Organisation's (WHO) reported a cluster of pneumonia cases of unknown aetiology in Wuhan city and China's Hubei province on 31 December 2019[7],[8]. In a press conference held by China's National Health Commission on 20 January 2020, human-to-human transmission of coronavirus was confirmed [9]. The virus spreads mainly through proximity to humans (less than one meter apart), via saliva particles or by touching previously contaminated surfaces. The most difficult thing about the spread of the virus is that it can often infect people who are free of symptoms for days. After exactly one month, 'Patient Zero' has been identified in Italy and the beginning of the nightmare has also taken place in our country [10]. On 11 March 2020, a global pandemic will be declared on WHO that will affect the whole world [11]. Figure 1.1 briefly presents the main events that have happened over time. To date, there are nearly 617 million cases of disease and about 6.54 million deaths in the world. But what should be most worrying is that various estimates based on excess mortality in a large sample of countries suggest that these numbers may be underestimated by a factor of two to four. In our country, one in four people has already been infected and the number of deaths is over 160000 [12].

The following chapter is mainly based on the book (Spillover: Animal Infections and the Next Human Pandemic), the article "Why Weren't We Ready for the Coronavirus?" published in the New Yorker and many interviews released by David Quammen.

1.1 COVID-19 origin

Everything comes from somewhere and dangerous, new viruses that pass to humans and cause new diseases come from wild animals. "How do we know?" because

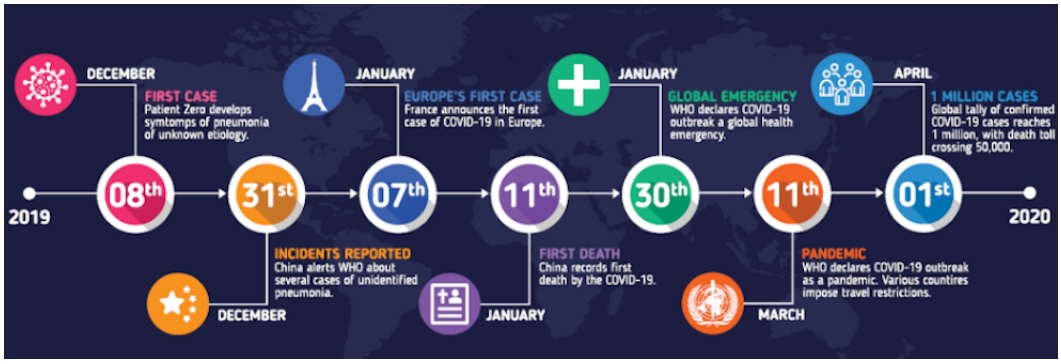


Figure 1.1: Timeline of COVID-19 most relevant events (Data source:[1])

viruses can only reproduce in cellular living things by making copies of themselves, they can only function within the cells of cellular creatures. Viruses are not cells, they are just little protein-enveloped packets, little capsules that contain genomes. The genome can either be a genome of the famous DNA, the double helix molecule, or a genome of another genetic molecule, RNA, which, unlike the double strand of DNA, usually consists of a single strand with the letters of the genetic alphabet. Viruses, such as the SARS-CoV-2 shown in Figure 1.2, replicate themselves by attaching to the cells of cellular organisms and then inserting their genome to hijack the cell's machinery and make copies of themselves.

Animals, plants, fungi, bacteria and all other living things are all cellular crea-

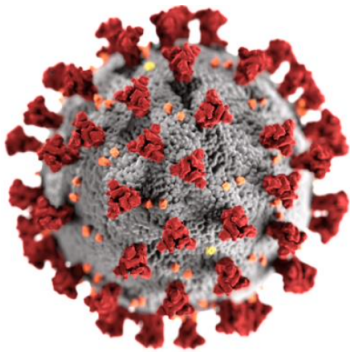


Figure 1.2: SARS-CoV-2 virion structure

tures. Some of them are simple cells, like bacteria, while others are made up of many complex cells that perform different functions, like animals and plants. We live in a world of viruses, they are everywhere, there are several viruses in every kind of wild and domestic animal on the planet and there are viruses in every kind of plant and in every kind of fungus. Some of the viruses have performed important functions throughout history. Viruses actually move genetic material

around and sometimes they introduce genetic material into a species that actually benefits that species. However, we tend to think of viruses in terms of disease. Viruses cause disease when they enter us and multiply, which in some cases has consequences for the body in which they multiply.
"How does this happen and how are they transmitted from non-human animals to humans?" Viruses are essentially passive, they do not seek us out, viruses cannot walk, run or fly, but they do "ride" inside living things. If the host in which a virus rides is disturbed, then the virus has the opportunity to move from its usual host, perhaps a monkey, rodent or bat, to another kind of host, perhaps a human, but not because they are looking for us, just for opportunity, and this opportunity arises simply because we as humans come into contact with wild animals all the time. Particularly by capturing or killing wild animals for food, but not only in this way, but also by disrupting diverse ecosystems that contain many different animals that carry many different viruses, so that the viruses have the opportunity to pass from their non-human animal host to the first human victim and multiply in him. After that, there is a possibility that they can be transmitted from one person to another.
A group of Chinese researchers has published on "Cell" an analysis of viruses found in wild animals traded and used as food in China. This study shows that 102 viruses were identified from nearly 2000 specimens belonging to 18 different species and five orders of mammals. Among them are 65 that were described for the first time and 21 that pose a high risk of infection to humans. Coronaviruses are a spillover event [1] . The conclusion is obvious: wildlife captured for commercial use is an extreme threat to human health. Yet the exotic animal trade remains the fourth most lucrative illegal trade in the world, after drugs, weapons and humans. So we ignore this kind of information and leave the virus to use its sheer strategy to transmit itself from one human to another. For example, like this: when a virus infects the cells in the respiratory tract, in the windpipe, and accumulates there and causes irritation, a kind of throat causes pneumonia. A person starts coughing, the virus is expelled when they cough and then has a chance to get into another person. Typical clinical signs are shown in Table 1.1 [13].

We have heard many stories about this terrible virus: it could have come from a seafood market in the city of Wuhan, or perhaps from a bat that brought it to that seafood market. These are the first stories that came up about the possible origins of this virus. But these stories become more complicated as time goes on. When this virus was isolated from human victims in the city of Wuhan in December 2019, and scientists sequenced the genome of this virus, they were able

[1] "Spillover event" refers to the overcoming of the many naturally occurring barriers that allow a pathogen to pass from members of one species as hosts to members of another species

Table 1.1: COVID-19 Symptoms

Most Common Symptoms	
Fever	87.9 %
Dry Cough	67.7 %
Fatigue	38.1 %
Sputum Production	33.4 %
Less Common Symptoms	
Shortness of Breath	18.6 %
Myalgia/ Arthralgia	14.8 %
Sore Throat	13.9 %
Headache	13.6 %
Chills	11.4 %
Rare Symptoms	
Nausea	5.0 %
Nasal Congestion	4.8 %
Diarrhea	3.7 %
Hemoptysis	0.9 %
Conjunctival Congestion	0.8 %

to determine that it was a coronavirus, which belongs to the coronavirus family. And why? Because coronaviruses have infected people in the past.
Back in 2003, when SARS broke out, the first SARS virus appeared in southern China and spread from the international airport through the city of Hong Kong, then made its way to Toronto, Beijing, Hanoi and Bangkok, making people seriously ill. It killed one in 10 infected people and was then fortunately stopped. In the end, only 8000 people became infected and about 800 people died. That was the first SARS, the first really dangerous coronavirus that spread to humans.
After that, scientists began studying coronaviruses to find out where this particular virus came from. Scientists from China, in collaboration with other international scientists, looked for the hosts from which the first SARS virus came and found very similar viruses to this one in horseshoe bats [14]. A team of scientists led by Dr Zhengli Shi, who maintains a laboratory at the Wuhan Institute of Virology, began studying the viruses, particularly coronaviruses, that live in bats in China, looking for the RNA signatures of these viruses. They found a number of coronaviruses living in bats in caves in southern China and began publishing about them. They found a virus in 2013 and gave it a code number. In 2016, they published a warning about the diversity of bats, their diversity of coronaviruses in southern China. In 2017, they published another warning about all the coronaviruses they had found in bats. In 2019, they published another warning about

the coronaviruses. All of these warnings were about the risk of coronaviruses in bats that could be transmissible to humans and cause infection in humans. Then, in late 2019, this virus shows up in humans in Wuhan and starts making them sick, the genome is sequenced and it is another coronavirus and then, in early 2020, she and her team publish another paper explaining that there is some evidence of where this virus comes from. In fact, in 2013, a virus was found in a mine shaft in southern China that was 96 per cent similar to this new COVID-19 virus and that there is very strong evidence that this virus, the SARS2 virus, also came from a bat. In 2013, she found the same virus that became the COVID-19 virus. It was 96 per cent similar at that time, meaning it represented 40 or 50 years of independent evolution. This indicated that the virus found had probably been living in a bat population in southern China for thousands of years and that there was another bat population separate from it in which this type of coronavirus also lived.
The separation between these two bat populations occurred perhaps 40 or 50 years ago, perhaps due to habitat destruction or human activity that prevented these bats from circulating from one group to the other and transferring their viruses back and forth, so these two viruses. Because of this separation, the coronaviruses evolved separately in them. This part of the population, which she did not sample, is thought to contain the immediate precursor of SARS COVID. The virus jumped from a bat to a human and it was probably already circulating in humans in Wuhan in early December 2019 and the case that occurred had no known contact with this particular market. Perhaps the virus was carried to the market by another animal brought to eat, or by a human who was already infected, and from there spread to other humans, creating the cluster. All we know with great certainty is that the virus came from a wild animal, possibly a bat, evolved naturally, passed to humans, then entered the city of Wuhan and from there spread around the world.
Another hypothesis was drawn up during the first year of the pandemic COVID-19: "the lab leak theory". The US President Donald Trump claiming that SARS-CoV-2 was originated in a laboratory in Wuhan, China. However, he never presented any proof of this and few scientists took this hypothesis seriously. Then a "joint mission" by the World Health Organization (WHO) and a Chinese scientific team investigated the possible origins of COVID-19. This it was concluded by describing the lab accident as "extremely unlikely" [15].

1.2 COVID-19 variants and open challenges

Although it has faded into the background, especially after the war in Ukraine and with contagious but less aggressive variants, this pandemic is far from over.
In fact, hundreds of variants of this virus have been identified worldwide in the past year. WHO and its international network of experts are constantly monitoring the

changes so that if significant mutations are detected, WHO can notify countries to take action to prevent the spread of this variant [16].
Here are the variants defined by the WHO as "Variants of concern" (VoC):

1. Omicron variant (variant B.1.1.529), first detected in South Africa in November 2021. Currently distributed mainly in Italy and Europe.

2. Delta variant (variant VUI-21APR-01, also known as B.1.617) was first discovered in India in April 2021.

3. Gamma variant (variant P.1) originated in Brazil in January 2021.

4. Beta variant (variant 501Y.V2, also known as B.1.351) discovered in South Africa in July 2020.

5. Alpha variant (variant 202012/01, also known as B.1.1.7), first identified in the UK in November 2020.

Viruses always mutate when they replicate, it is just routine, they mutate all the time, these variants are clusters of mutations that have spread and then clustered together into one variant. Some people say that these variants are more transmissible, but it is not true. It just means that they are more successful,that they are fitter to survive. There is concern that these variants may escape the existing vaccines. As evidence, there is evidence that the Brazilian variant, for example, may have already evolved to be able to re-infect people who have already survived infection. The variations are the sign of evolution, and if evolution is taking place, commonly referred to as Darwinian evolution, the more people are infected and the more the viruses are replicated, the more they can evolve. In other words, it is impossible to take anything for granted. No solution can be adopted directly by all states. One must be able to analyse the available data as well as possible. If something goes wrong because the analysis is not adequate, or if the scientists forget even one variable, it will make all the difference and the final models will not work properly. This can happen because even though data is now "the new oil", a huge collection of data is of no use if the statisticians and data scientists are not able to manage it properly. Instead, if the research group has the right core skills to overcome this problem, the analysis carried out will lead to estimates, hopefully with a small confidence interval. In other words, this means that the results can be trusted and acted upon accordingly. All this happens because even if there is a common divisor, in reality for COVID-19 there is no generalisation that can be applied to every condition, and the approaches to this disease are really heterogeneous.
There are two real proofs of this: the case of a vaccine without patents and the case of Cuba.

CORBEVAUX is a vaccine that is completely new in the sense that there are no patents blocking its production and that it can be produced very quickly. Finally, the first vaccine that would escape the egoism of countries and the greed of established companies. While there is talk of the 4th dose in Western countries, poor countries like Cuba are looking for new solutions. The main cause was the lack of the large sums of money needed to buy vaccines per capita. Her personal solution to solve this problem was to develop a new vaccine. Thanks in particular to Iran's help in the experimental phase, where the number of variants ranged from one to another, they found a vaccine by using conjugate vaccine technology. This is known as Soberana 2 and differs from all the others in that it is used in children. The results obtained are excellent. With 97% of the population vaccinated, the infection has actually decreased.
The first unequivocal proof over the whole world that the vaccination of children is an absolutely right course to pursue [17]. There is no rule except one: "No single option is excluded to defeat this pandemic, everything is possible!".

1.3 The lifelong consequences of the pandemic: how it changed the world

The main consequences of this outbreak will be the worsening of inequalities even in developed countries like the US, but especially in developing countries. In 2020 to 2021, the wealth of billionaires will grow to $ 4400 billion, while in the same period over 100 billion people will live below the poverty line [18]. This is the main reason why the USA is the country with the most victims due to COVID.
This disease has affected people with health conditions highly correlated with poverty and work incompatible with isolation. Many Americans do not take tests to determine if they are infected and go to work spreading the virus, or they have asked for help too late. The poor will become poorer as jobs are lost, especially in the low-wage service sector. Another relevant aspect is the fact that many poor children have educational deficits due to online learning, predicting a strong growth in inequalities that will increase over time. The ability to develop, produce and distribute so many doses of vaccine so quickly is certainly a victory for the scientific organisation, but on the other side of the coin there are these kinds of epic failures. Even with all the technology and resources we had, we were not able to increase the vaccine stockpile enough to distribute it to poor countries.
Markets may be able to solve an economic problem, but they are not able to overcome the legal barriers put in place for the intellectual property rights of vaccine owners, so they effectively have a monopoly on the entire market. Incidentally, these companies have an incentive to limit production and set prices at a value that is many times their fixed production costs. This is absolutely crazy, especially

when you consider that the original production was financed with public money. Unfortunately, the COVID shock is very likely to continue. Not doing everything to counteract this disease everywhere is nonsense. Because doing so will perpetuate COVID in our lives by delaying global recovery and exacerbating already high levels of inequality.

1.4 Will be ever get rid of it?

As can be seen from the previous sections, much has been reported in the scientific literature about the danger. Further evidence of this is the November 2015 in which an article by Ralph Baric et al. said that bat coronaviruses are very similar to SARS viruses. More specifically, they may have been able to make a "species jump" from Chiroptera to humans [14]. Thus, a potential risk to humans was already apparent that year, and yet the whole world was caught unprepared.
Viruses give no warning, they are simply creatures that replicate themselves using a genome of DNA or RNA and they are subject to Darwinian imperatives. The three specimens Darwin drew represent evolution, diversity of complexity and adaptation to the process he discovered via natural selection, shown in Figure 1.3: Where D, B and C represent different species that have evolved. Darwin's imperatives are based on 3 things:

1. Makes as many copy of itself as it can
2. Expand itself in geographical space
3. Persist, avoiding extinction

If the first two rules can be fulfilled, then there is a high probability that the third Darwinian imperative will be fulfilled, so that extinction is avoided and the virus survives.
In early February, "Nature" published a paper based on genetic Big Data, in which there are at least 100,000 previously unknown RNA viruses that have left their signature in the database.
Let us now ask ourselves: "if it had not been the SARS-CoV-2 pandemic, how much attention would we pay to these alarm bells?". Be that as it may, we may never get rid of the virus. This is simply because the virus could evolve and become less harmful to humans. Maybe it will just be another cold, like some other coronaviruses that have infected humanity, but there is no guarantee of that. The measles virus, for example, has been prevalent in humanity for thousands of years, and although there has been a vaccine for almost 60 years, many people around the world have not been vaccinated. Therefore, there is the chance that they will become infected or die from this disease. So far, there is no proof whether COVID

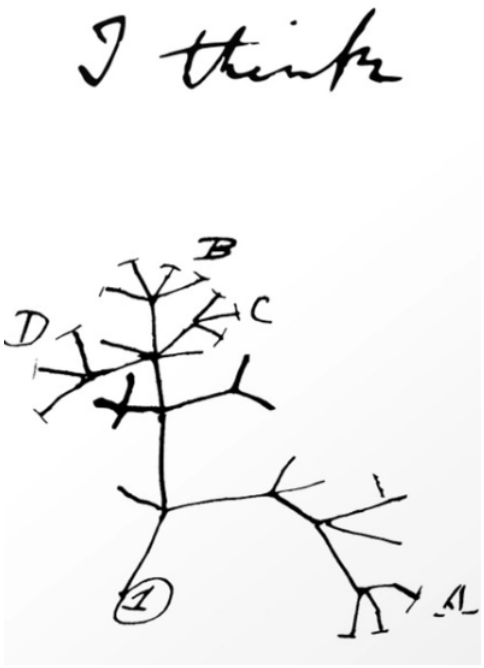


Figure 1.3: The first evolutionary tree sketched by Darwin in one of his notebooks (Data source:[2])

remains in the population forever or whether it is just a nuisance or whether it could remain a threat only to a part of the world's population. This has led to the need to be vaccinated against this virus and other coronaviruses even forty years from now.

1.5 The possibility of further pandemics like this in the future

As already written, nothing is certain about this disease. However, it seems worth investing today to avoid something unpleasant happening tomorrow. This is confirmed by an article by epidemiologists and ecologists published at the beginning of February [19]: Primary prevention is needed against pandemics, which means that three measures can be taken:

1. Monitoring the pathogens that cause the spillover event
2. Better management of the animal trade
3. Stop deforestation

These 3 simple actions would definitely cost one-twentieth less than late intervention. However, if no one intervenes, the only high probability is that there will be such a virus in the future that will completely change our lives.
It is not necessarily a pandemic, but there will be more spillovers of viruses passing from wildlife to human hosts. In some cases, a disease outbreak may affect a dozen people in a small town in the US or in a remote village in the third world countries.
However, it is not inevitable that every outbreak will become an epidemic that spreads across a country and then a pandemic that spreads across the world. It is not inevitable, but something we can manage if we are better prepared next time. All of this is only possible if we recognise in the future that spillover events leading to new diseases have occurred throughout history, with a frequency that has unfortunately increased over the last 60 years. There are many examples of such pandemics in the history of the world. HIV was recognised in 1981, followed in 1994 by the Hendra virus, which was transmitted from bats in Australia to racehorses and passed from them to humans. In 1998, the Nipah virus was transmitted from bats to pings and then to humans in Malaysia, causing a fatal disease. In 1997, avian influenza in Hong Kong was transmitted from wild birds to chickens and ducks and then to humans. In 2003, the original SARS in China was transmitted from bats to humans. In 2012, another coronavirus, MERS, Middle East Respiratory Syndrome, was transmitted from bats to camels and then to humans in Saudi Arabia. And finally, in 2015, the Zika virus, a viral disease transmitted to humans through the bite of infected mosquitoes of some species of the genus Aedes.
A drumbeat of these events has already taken place, in some cases with only a few dozen deaths, in others like the COVID pandemic with millions of deaths.
But the question is always one of "why?" The answer is simply the fact that there are more people today than ever before, causing disruption to wild and diverse ecosystems. There are 8 billion of us on the planet. We are draining resources from the natural world for ourselves and for all the choices we make, even the most habitual ones, like what we eat, wear, buy, how much we travel, how much fossil fuel we use, how many children we want to have. All these choices we all make add up to our collective footprint in nature, and it is this footprint that displaces the viruses that naturally occur in wildlife. It dislodges them from wild animals and gives them the opportunity to infect humans and become a pandemic.
We can do something about it, we are smart, but we are also hungry and numerous. We can develop vaccines, which we have already done, and in this way prevent this type of virus from spreading even more dramatically than it already has. We can prevent future spreads by identifying new ways to mitigate them and also our impact on the world. In particular, we can better prepare for pandemics by developing prototype vaccines that can be used through international virus surveillance

and detection networks for new viruses. It is about human responsibility as members of our community, but also as members of our own families. We must believe in science and persuade those who deny it to do so.

1.6 Science and technology against the viruses

We are in an evolutionary process, including vaccines, in terms of the different virus variants, and we are disrupting the ecosystem so that we increase the likelihood of pandemic outbreaks. From certain points of view, we can certainly do more to minimise our impact on these ecosystems, but on the other hand, it is inevitable that our impact on these ecosystems will increase. The projections are that population growth, even with extreme efforts by countries and individuals to reduce it, is likely to be between 9 and 11 billion people. Even our current figure of 8 billion is almost four times the number of people there were a hundred years ago at the time of the 1918 flu. But it not just about size, it is about total population multiplied by consumption (individual and collective).
We can reduce the impact of consumption and technology even as the population continues to increase through individual responsibility and improved technology.
Technology is one part of this equation. It can either exacerbate the impacts we have or on the other hand there are inventions like the one that allows recycling of plastics, which tends to decrease our consumption. Technology and scientific knowledge play a key role in managing this pandemic to minimise the risk of people being exposed to the virus or becoming seriously ill. Therefore, it is necessary to use science and technology in collaboration to reduce the human impact on the world, with positive technology being the key object: science is just a method of knowledge, a process of discovery, while technology is its own application.

Chapter 2

Background

Artificial intelligence (AI) and related technologies are becoming increasingly common in business and society, and are beginning to be applied in healthcare [20]. AI is not just a technology, but rather a collection of technologies. Most of them are directly related to healthcare, but the specific processes and tasks they support are very different. Machine Learning (ML) is one of the most common forms of AI. It is a statistical technique for fitting models to data and 'learning' by training models with data. Currently, ML is a rapidly developing and ever-growing field. It programs computers using data to optimize their performance. It learns the parameters to optimize the computer programs based on the training data or its previous experience. It can also use the data to predict the future. ML also helps build a mathematical model using the statistics of the data. Its main objective is that it automatically learns from given data (experiences) by providing the desired results by looking for trends/patterns in the data [21].

2.1 The Analytics Life Cycle

Today's business challenges start with large amounts of complex data. Effective decision-making requires state-of-the-art predictive modelling techniques. The analytics process always begins with a business challenge. In response, the business establishes a precise and measurable analytics objective. The analytics life cycle represents a series of activities whose goal is to extract value from raw data. The definition of value depends on the objectives of the particular organisation. The analytics life cycle, as seen in Figure 2.1, has three phases:

1. **Data** is the foundation of everything. This phase is about exploring and preparing data for analysis.

Figure 2.1: Representation of the Analytics Life Cycle (Data source: SAS)

2. **Discovery** is the act of discovering something new. It is the result of science using scientific and technological knowledge. By creating and refining several models, the final aim is to select the best model for the analysis to be calculated.

3. **Deployment** in which we put the model into practise. By applying the model to new data, which is a process called scoring.

Thus the value of ML is demonstrated throughout the analysis life cycle with actionable insights at each stage. It results on being flexible as things can be done in different order within the same project.

2.2 The 5 V's of big data

It is not always true that "the more data you have, the better it is".
In fact, in the real world, it is typical for data to be dirty, noisy, with some inconsistencies and incomplete records. Good data is the foundation for good models.
The 5 V's of Big Data enable data scientists to get more value from their data:

1. Value
2. Volume
3. Veracity
4. Velocity
5. Variety

The data must also be cleaned and reduced to the optimal size for analysis.

2.3 Machine Learning

Machine Learning (ML) develops systems that iteratively learn from data and make predictions. The three main characteristics of ML are: automation, customization and acceleration (Figure 2.2).

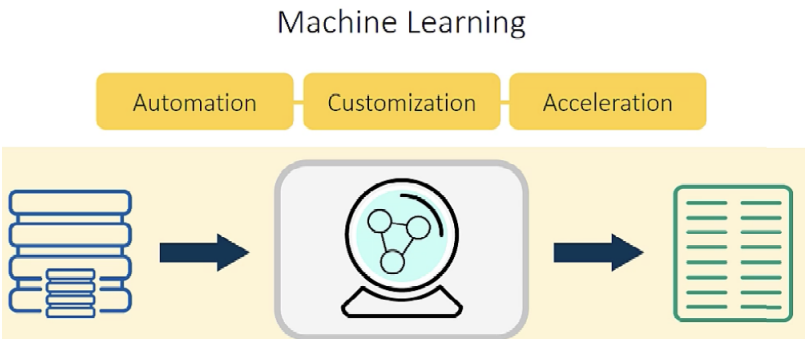


Figure 2.2: Illustration of the main features of Machine Learning main characteristics (Data source: SAS)

ML is largely an automated process. The creation of models requires minimal human intervention.

1. Automation: system learns iteratively from the data, based on algorithms rather than programming. With each pass through the data, the system identifies patterns and makes predictions, continues to learn and improve.

2. The ML process is highly customized. It can draw on many algorithms for training the data, depending on the requirements of the situation.

3. ML speeds up the process of performing sophisticated analyses on large amounts of data so that results can be obtained quickly.

2.3.1 Supervised Learning

Supervised Learning (SL) is a Machine Learning model created to make predictions. SL is also known as predictive modelling or supervised prediction and starts with a training dataset. More specifically, this algorithm is performed by using a labelled dataset as input (independent variables) and it generates responses as output (target). The observations of the training dataset are called examples or instances. SL develops predictive models from classification algorithms and regression techniques. For a given instance, the inputs reflect the a priori knowledge before the target is measured.
- Classification predicts discrete responses. Mainly, the algorithm classifies by

choosing among two or more classes for each example. When this happens between two classes, it is called binary classification and when it happens between more than two classes, it is called multi-class classification. Applications of classification include handwriting recognition, medical imaging, etc.
- Regression predicts continuous responses. The model learns to predict numeric scores or a statistical value. The analysis is done by predictions [22]. The common regression techniques are:
- Linear Regression
- Polynomial Regression
- Logistic Regression
The variables used inside the model can be numerical or categorical. The first is also called an interval variable and can be further classified as continuous or discrete:
- The continuous case can be any infinite number of values within a certain range
- Discrete variables are often counts
Categorical variables take qualitative values, with each value representing a group or category. The number of possible values of a categorical variable is usually finite.
The purpose of the training data is to build a predictive model that relates the inputs to the target. The predictive model is a concise representation of the relationship between the inputs and the target.

2.3.2 Unsupervised Machine Learning

Unlike supervised learning, there is no supervisor, only input data. The basic goal is to find certain patterns in the data that occur more often than others. In statistics, this is called density estimation. One of the methods for density estimation is clustering. This involves grouping the input data into clusters or groupings. Assumptions are made to discover clusters that fit a classification reasonably well. This is a data-driven approach that works better when enough data is available. However, these approximations are usually weak compared to supervised learning [22].

2.4 Algorithms

In this work, I have examined several algorithms with which I have carried out supervised classification.

2.4.1 Logistic Regression

For most applications that have a binary target, logistic regression replaces linear regression [23].Both the set of inputs and the target variable (output) can only take on discrete values when dealing with classification problems. In logistic regression, the expected value of the target is transformed by the logit link function to constrain its value to the range of 0 to 1. Mathematically the logit is the inverse of the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$. Thus it is defined as $logitp = \sigma^{-1}(p) = ln\frac{p}{1-p}$ for $p \in (0,1)$ and it is shown below 2.3.

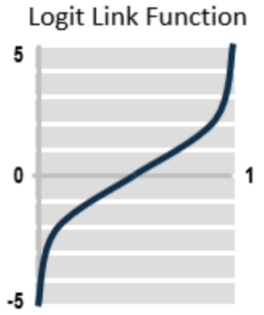


Figure 2.3: Logit link function illustration (Data source: SAS)

The logit link function converts probabilities (between 0 and 1) into logit scores (between negative infinity and positive infinity)[24].The predictions of the logistic model can be considered as primary outcome probabilities. This model infers the probability of P(Y = 1).
A linear combination of the inputs produces a logit score, the logarithm of the probability of the primary outcome, as opposed to the direct prediction of the target by linear regression. For binary prediction, any monotonic function that maps values between 0 and 1 to the real number line can be considered a link. The logit link function is one of the most common. Its popularity is partly due to the interpretability of the model. The functional form of the hypothesis is:

$$Y = C^T \cdot X \tag{2.1}$$

where C is the column vector of regression coefficients and X is the list of the features

$$C = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad X = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_n \end{bmatrix} \tag{2.2}$$

β_i are the regression coefficients or weights for the features present in the data and β_0 is the intercept of the equation. We have:

$$y = c^T \cdot x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.3}$$

In linear regression, the ordinary least square method is used for parameter estimation. However, in the logistic regression model, parameter estimation is complicated by the presence of the logit link function. Therefore, maximum likelihood estimation is used for parameter estimation in logistic regression. In our specific case the record will be defined as survived or death if the value of

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \geq 0 \tag{2.4}$$

The likelihood function is the joint probability density of the data treated as a function of the parameters. A simplified version of the likelihood function for a binary target is:

$$\sum_{i=1}^{n} log(\hat{p}_i) + \sum_{i=1}^{n} log(1 - \hat{p}_i) \tag{2.5}$$

where $\hat{p}_i$ is the probability of die while $(1 - \hat{p}_i)$ is the probability of survive. The mathematical formula (2.5) is the sum of two quantities: the first element on the left represents the training cases with primary outcome while the quantity on the right represents the training cases with secondary outcome. The maximum likelihood estimates are the values of the parameters that maximise the probability of obtaining the training sample. These estimates can be used in the logit and logistic equations to obtain predictions.
One of the attractions of a standard logistic regression model is the simplicity of its predictions. The contours are simple straight lines, commonly known as the isoprobability lines. In higher dimensions they would be hyperplanes. It is possible to get a probability using a straightforward transformation of the logit score, the logistic function. Logistic regression uses a linear combination of the input variables to predict the target and this it performs well for many simple datasets that are linearly separable. It has the limitation that it can only draw straight lines, even if the data could be better separated by more complicated geometry. In higher dimensions, this problem remains, even though the lines are replaced by planes (in three dimensions) or hyperplanes (in higher dimensions).
One way to overcome the rectilinear constraint and introduce curved decision boundaries is to include higher order polynomial terms in the model. A second-order polynomial regression model would include all quadratic terms. By adding these terms to the model equation allows the model to draw quadratic decision boundaries rather than just linear boundaries. By adding third-order terms (such as x^3), the model can draw cubic decision boundaries and so on for higher order terms. These additional terms increase the flexibility of the model but potentially

leading to overfitting. If we know that the inputs are not linearly related to the target, it may be useful to add polynomial terms.

2.4.2 Gaussian Naive Bayes

The Naive Bayes algorithm is an "elementary" probability classifier. This type of algorithms predict class as a function of the probability of appertaining to that specific class [25]. It computes a series of probabilities from the number of frequencies and values in a given data set. It is based on Bayes' theorem by following this formula:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} \tag{2.6}$$

Using this theorem, it is possible to determine the probability of event A (the target event) occurring if a particular event B has occurred. In COVID case the event B can represent for example the presence of a particular symptoms or an individual personal data such as the age or sex. It assumes that variables are all independent and due to this, is a very fast algorithm compared to other complicated ones, which is an advantage in cases in which computational saving time is preferred over higher accuracy. However, this lack of independence does not model the real-world context because it ignores the correlation between variables. Therefore, it is referred to as "Naive" [26].

2.4.3 Decision Tree

This algorithm is a non-parametric supervised learning method. Decision Tree models are easy to explain and interpret. Trees follow a decision split (decision rules) , IF-THEN logic and can be represented in a tree-like graphical structure (see image 2.4) which are inferred from the data given [27].
Decision trees use rules that determine a decision based on the values of the input variables. The rules are expressed in Boolean logic and they are arranged hierarchically in a tree-like structure with nodes connected by lines. The first rule is placed on the base is called the root node. Then the Subsequent rules are called inner nodes. Nodes with only one connection are called leaf nodes. The depth of a tree indicates the number of generations of nodes. The root node is generation 0, the children of the root node are generation 1 and so on.
For the evaluation of a new case, the input values must be examined and the rules defined previously by the decision tree classifier must be applied. By increasing the maximum depth it can cause overfitting while vice versa increasing the number of minimum leaf size can prevent overfitting.
The goal of splitting is always to reduce the variability of the target distribution

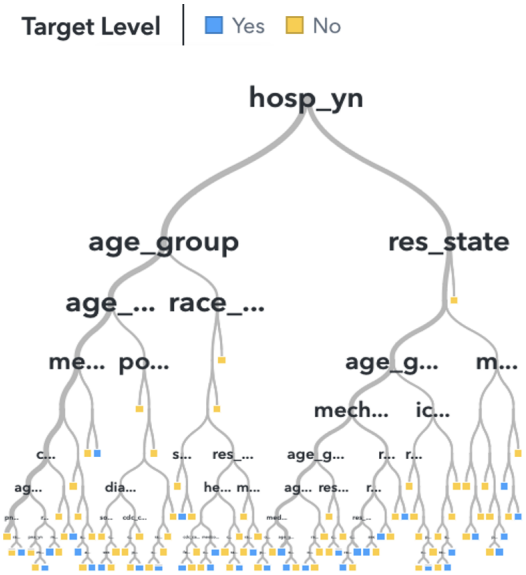


Figure 2.4: Representation of the Decision Tree classifier implemented on SAS Viya

and thus increase the purity in the child nodes. A splitting criterion measures this reduction and there are a variety of impurity reduction measures that can be used. Most of these are applicable to binary or interval categorical targets. The result of the splitting process is called the maximum tree. The maximum tree is built solely on the training data, so it is unlikely to generalise well to the validation data. The maximum tree is the starting point for optimising the initial model through a process called pruning. In bottom-up pruning, we start from the most complex maximal tree and then consider a tree with n-1 leaves until we reach the root of the node. We then compare all these possible candidates and choose the one that fits the validation data better, when two models fit the validation data in the same way then the one less complex will be selected.
Decision trees recursively partition the input space into different regions and classify each region according to the value of the target for the majority of training cases in that region. In this way, a list of rules for classifying new data points is created. Larger and more complicated decision trees produce more distinct regions and longer lists of rules.
The first part of the algorithm to create the tree is called the split search. Split search begins by selecting an input to partition the available training data. Two groups are created for a selected input and a fixed split point. Cases with input values smaller than the split point are branched to the left and cases with input values larger than the split point are branched to the right. The groups, together

with the target results, form a 2x2 contingency table with columns indicating a branching direction and rows indicating a target value. A Pearson's chi-square statistic is used to quantify the independence of the counts in the columns of the table. Large values of the chi-square statistic indicate that the proportion of 0s and 1s in the left-hand branch is different from the proportion in the right-hand branch. A large difference in the proportions of outcomes indicates a good split. Since the Pearson chi-square statistic can be applied to the case of multiway splits and multi-outcome targets, the statistic is converted to a p-value. The p-value indicates how likely you are to get the observed value of the statistic if you assume identical target proportions in each branch direction. For large data sets, these p-values can be very close to zero. For this reason, the quality of a split is given by the log value, which is -log(chi-squared p-value). The best split for an input is the split that gives the highest logworth.

2.4.4 Random Forest

One problem with decision trees is that they are relatively unstable models. This means that a slight change in the training data can significantly change the model created [28]. When a tree is created with the slightly changed data, the decision boundaries are completely different, although the prediction accuracy could remain the same. This leads us to wonder whether we have really created the best possible decision tree, because many trees trained with similar data will also perform similarly. One solution to this problem is to create many different decision trees and combine their results. This random forest method recognises that there are many different trees that could contribute to an accurate prediction of the target, but instead of selecting one, their results are averaged. The Random Forest algorithm randomly samples both observations and variables when it creates the trees in the forest. This random selection means that each tree in the forest is different because it has been trained with a different part of the data. The combination of trees can be done by averaging the predicted value or by each tree voting on a prediction and following the majority rules (Figure 2.5). Random forests not only solve the problem of instability associated with decision trees, but also reduce variance by combining a collection of different models into an average. This reduction in variance can improve model performance and reduce the risk of overfitting.
Although the performance of Random Forest is great, it has its limitations when it comes to understanding why certain predictions are made. It does not start from a data structure and therefore it is difficult to ensure how and why certain features may affect the prediction performance.

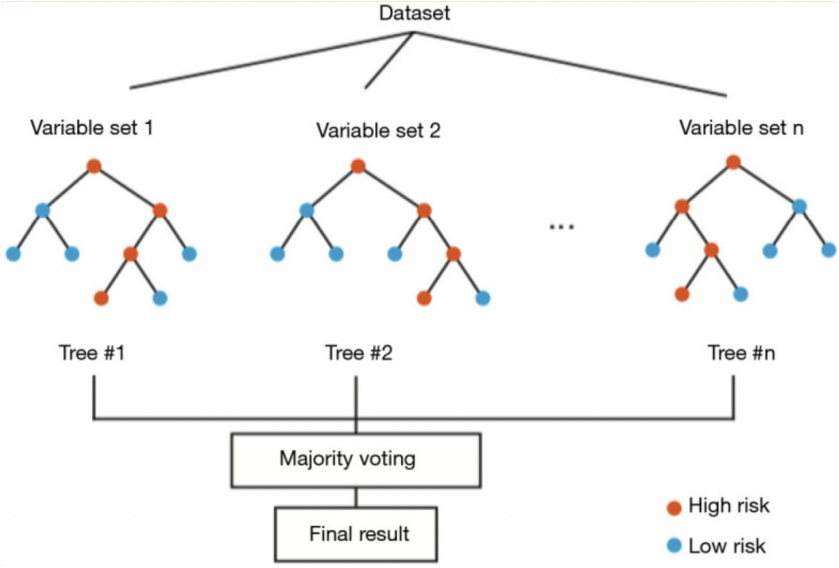


Figure 2.5: Conceptual diagram of the Random Forest algorithm. Averaging many random decision trees significantly reduces variance and bias in individual samples. (Data source: [3])

2.4.5 Boosted Random Forest

Gradient boosting is another method to improve the performance of decision trees by combining them. This time, however, an iterative sequence of decision trees is used for improving the model at each iteration. We start by creating a single decision tree from the training data. This tree will make some errors, so we calculate a column of residuals that simply represent the errors the decision tree made in classifying points. These residuals are calculated as the difference between the true value of the target and the prediction of the target by the decision tree. We then create a second decision tree with the same inputs as the original tree, but instead of trying to predict the target, we only try to predict the residuals of the original model. The cycle continues as we create a series of decision trees. These trees can then be superimposed to produce an ensemble model that has a much smaller error in the training data than the first decision tree we trained in the sequence. This procedure is basically designed to overfit the training data. So if we create enough trees to get good performance on the training data, we weight the trees less and less at each iteration to get good performance on the validation data. Boosting has the advantage of building the ensemble model by focusing on the misclassification errors that occur during each iteration and increasing the weights for each misclassification, which in this way can end up improving the performance of the model [29].

2.4.6 Gradient Boosting

The gradient boosting model uses a partitioning algorithm described in [30]. A partitioning algorithm searches for an optimal partition of the data defined by the values of a single variable. The optimality criterion depends on how another variable, the target, is distributed among the partition segments. The more similar the target values are within these segments, the higher the value of the partition. Most partitioning algorithms further partition each segment in a process called recursive partitioning. The partitions are then combined to create a predictive model. The model is evaluated using fit statistics defined in terms of the target variable. A good model may result from many mediocre partitions.
Gradient boosting resamples the analysis dataset several times to produce results that are a weighted average of the resampled dataset. Tree Boosting creates a series of decision trees that together form a single predictive model. Each time, the data is used to grow a tree and the accuracy of the tree is calculated. Successive samples are adjusted to compensate for the previously calculated inaccuracies. Since each successive sample is weighted according to the classification accuracy of previous models, this approach is sometimes called stochastic gradient boosting.
Like decision trees, boosting makes no assumptions about the distribution of the data. Boosting is less prone to overfitting the data than a single decision tree. If a decision tree fits the data reasonably well, boosting often improves the fit.

2.4.7 Support Vector Machine

Support Vector Machine (SVM) [31], like neural networks, are black boxes, meaning they are difficult to interpret, but they are very flexible and automatically discover any relationship between the inputs and the target, so there is no need to specify the relationship before modelling. Unlike trees and neurons, they are not something that most people can visualise. The classifier model (i.e. the separating line) is symbolically denoted by H and its mathematical formulation is:

$$H = < w, x > + b = 0 \tag{2.7}$$

H has two elements: a normal vector w and a bias term b. The vector w is perpendicular to H. In a two-dimensional input space where H is a line, w affects the slope of the line. The bias term b is a measure of the offset of the division line from the origin. H is defined as the dot product between the vector w and the vector of inputs x, plus b. If a point falls exactly on the line, the value of H is 0 and the result of the dot product gives a scalar as the answer. To train a SVM model , you must choose w and b so that the line separates the values of the target. This also applies to a higher dimensional example that has a hyperplane instead of a line. The predictions done after training for new observations are made using

the formula above, in which H returns predictions as positive or negative values, where the sign, not the magnitude, is the crucial aspect. Points on one side of the line return a positive value and points on the other side return a negative value. To find the best classifier for linearly separable data, the approach is to find the hyperplane with the largest distance. This means that we try to maximise the margin of error on both sides of the separator. In the two-dimensional example

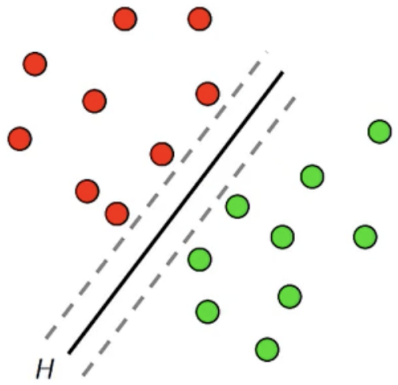


Figure 2.6: Representation of how SVM works in two-dimensional space for linearly separable data (Data source: SAS)

shown in the Figure 2.6, this would mean visualising the thickest line touching the innermost red values of the target and the innermost green values of the target. The thickest line is the largest margin of error on the positive and negative side. The best solution is the exact centre or median of the thick line. If you take the exact centre of the fat line, you get a unique solution known as a hyperplane with maximum margin. This solution, shown as H in the diagram, has the largest margin of error on both sides. The support vectors, also known as the carrying vectors, are the points in the data closest to the hyperplane. Only these points determine the exact position of H.
In the figure 2.6, there are five support vectors: two green and 3 red. Using support vectors avoids the curse of dimensionality because all other points are irrelevant to the solution. When the data is not linearly separable, one possible solution is a support vector machine, also known as a soft margin hyperplane (as in Figure 2.7), which allows for some misclassification. In a two-dimensional input space, soft margin means that a line can separate most points, but some errors occur. We account for the errors by using a penalty term in the optimization process. This penalty is the product of two quantities: an error weighting, which is a regularization parameter and is often denoted C, and the distance between a point in error and the hyperplane. It provides a balance between model complexity and training

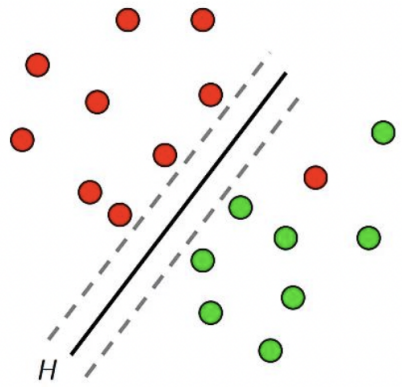


Figure 2.7: Representation of how SVM works in bidimensional space for non linearly separable data (Data source: SAS)

error. A larger value leads to a more robust model with the risk of overfitting the training data. A larger value leads to a more robust model with the risk of overfitting the training data.
If the data points are not linearly separable, we have what is called a soft margin hyperplane. In this case, we have to take into account the errors that the separable hyperplane might make. During the optimization process, the distance between an erroneous point in error and the hyperplane is typically denoted by ξ.
Given the need to account for errors, the optimization problem is solved by minimization:

$$||w||^2 + C \cdot \sum_i \xi_i$$

Under the single constraint:

$$y_i \cdot (< w, x_i > + b) \geq 1 - \xi_i \ , \ \xi_i \geq 0$$

The method used to solve the optimization problem is called the Lagrange approach. The $a_i \geq 0$ are called the Lagrange multipliers and they summarize the problem in a Lagrange function, while the constraints: $\xi_i \geq 0$ and $a_i \geq 0$ are used for finding the saddle point of the Lagrange function. Thus, the optimization problem consists of the following Lagrangian function that is:

$$(w, b, \alpha, \xi) = \frac{1}{2}||w||^2 + C \cdot \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(\xi_i + y_i(+b) - 1\right) \tag{2.8}$$

To detect the saddle point, $L\,(w, b, \alpha, \xi)$ is minimized with respect to w, b and ξ, but maximized with respect to α_i. However, there may be another problem: The red and green points can no longer be separated by a straight line. Here in the

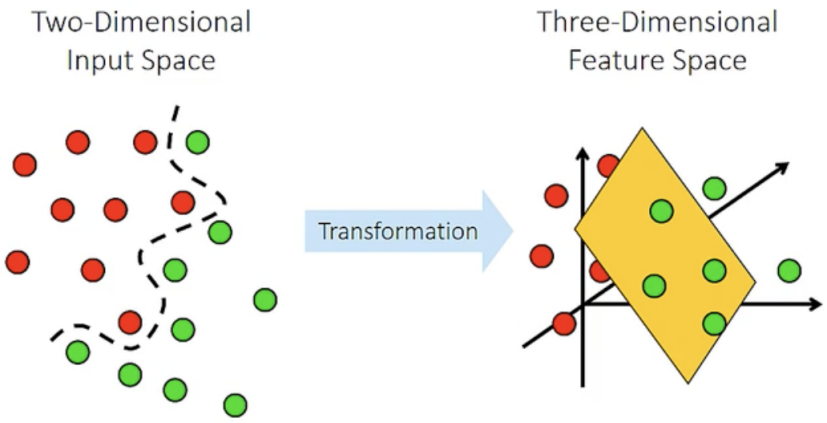


Figure 2.8: Representation of feature space approach in SVM algorithm (Data source: SAS)

Figure 2.8 is represented a possible example of this situation: Even a soft margin classifier would cause too many errors. The solution is the so-called feature space approach. The first step is to project of the data points to a higher dimension using a non-linear transformation called feature space, and then find the hyperplane with maximum margin in this higher dimensional space.
In the red versus green example, it is possible to transform the data into a three-dimensional feature space where a plane separates the red and green points.
Nevertheless, optimizing the complexity of the model in a higher-dimensional feature space is a mathematical challenge. Dot products are required to solve H and evaluate new observations. However, a dot product on transformed data has a different form and is difficult to calculate. To avoid dot products on transformed data, there is a mathematical trick called the kernel function. The kernel function has the following form:

$$K\left(x_i, x_j\right) = \phi\left(x_i\right), \phi(\left(x_j\right)$$

The kernel function exists in the original input space, but it is equivalent to a dot product on the transformed data in the higher dimensional feature space. In this way, it is not necessary to know the transformation and understand exactly what the feature space looks like. It is sufficient to specify the kernel function as a measure of similarity and the geometric interpretation remains the same because the solution is still a hyperplane. If the data points are linearly separable, there is an infinite number of separating hyperplanes. The starting point to reach a unique solution is to imagine a "fat" hyperplane between the points with different target values. This leads to a separator that has the largest margin for error.
Among all these hyperplanes, there is only one that has the maximum margin.
It is essentially the median of the bold hyperplane. The data is not always linearly

separable, so points that lie on the wrong side of the decision boundary lead to a penalty term in the algorithm. This penalty term is the hinge loss function and is used to generalise Support Vector Machines to cases where the data is not linearly separable. This soft margin classifier still looks for the maximum margin of error, but now also minimises the hinge loss penalty associated with the misclassified points. In many real-world situations, the data is not linearly separable, but a soft margin classifier would make too many errors to be a viable solution. One solution to this problem is to transform the data into a higher dimensional space and then find the hyperplane with the maximum margin in that higher dimension. Data points that are not linearly separable in lower dimensions can become linearly separable in higher dimensions, although the calculation becomes more difficult as we increase the number of dimensions.
To make this calculation easier, the kernel trick is used to convert the dot product calculation in higher dimensions into a kernel function in lower dimensions. This non-linear kernel function allows us to find the hyperplane with the maximum margin in higher dimensions without transforming each data point into its higher dimensional representation. Once the hyperplane with the maximum margin has been calculated in higher dimensional space, it can be transformed into a non-linear decision boundary in lower dimensional space, allowing us to generate non-linear decision boundaries using Support Vector Machines.

2.4.8 Bayesian Network

A Bayesian network (BN) is a directed acyclic graph that represents probability relationships and the structure of conditional independence between random variables[32]. A BN can explicitly represent distribution dependence relationships between all available random variables; it allows the discovery and interpretation of dependence and causality relationships between variables in addition to the conditional distribution of the target. In contrast, support vector machines and neural network classifiers are "black boxes" while logistic regression and decision tree classifiers they are able only to estimate the conditional distribution of the target. Therefore, BN classifiers have great potential for real-world classification problems, especially in domains where interpretability matters(see Figure 2.9).

A Bayesian network is a graphical model consisting of two parts (G, P):
- G is a directed acyclic graph (DAG) in which the nodes represent random variables and the arcs between the nodes represent the conditional dependence of the random variables.
- P is a set of conditional probability distributions, one for each node depending on its parents.
BN have two important properties:

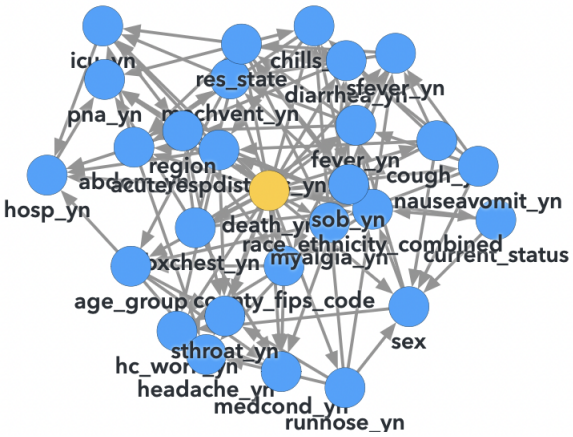


Figure 2.9: Representation of one of the Bayesian Network implemented on SAS Viya

- Edges or the arcs between nodes represent "causality", so no directed cycles are allowed.
- Each node is conditionally independent of its ancestors if its parents are present. This is called a Markov property.
According to the Markov property, the joint probability distribution of all nodes in the network can be decomposed into the product of the conditional probability distributions of each node as a function of its parents, that is:

$$Pr(G) = Pr(X_1, X_2, ..., X_p) = \prod_{i=1}^{p} Pr(X_i|\pi(X_i)) \tag{2.9}$$

where $\pi(X_i)$ are the parents of node X_i

2.4.9 Artificial Neural Network

Traditional non-linear modelling techniques become much more difficult as the number of inputs increases: this phenomenon is called the curse of dimensionality. It is unusual to see parametric nonlinear regression models with more than a few inputs, because deriving an appropriate functional form becomes increasingly difficult as the number of inputs increases. Higher-dimensional inputs are also a challenge for non-parametric regression models. Artificial Neural Networks (ANN) have been developed to overcome these challenges.They are a non-linear model designed to mimic the neurons in the human brain [33]. Although neural networks are parametric nonlinear models, they are similar to non parametric models in one way: neural networks do not require to specify the functional form. This allows to construct models when the relationship between inputs and outputs is unknown.

In a neural network (NN), the weights start close to zero. With each pass through the data, the NN learns more and refines the weights. Neural networks generally work well in sparse, high-dimensional spaces.
A major advantage of this model is the concept of flexibility: neural networks can be used to predict very complex surfaces. It is a universal approximator that can actually model any input-output relationship, no matter how complex. For example, the image 6.6 shows the relationship between two inputs, x_1 and x_2, and a target value y. This complex relationship would be difficult to estimate accurately using conventional regression methods. However, a neural network can take on this task.

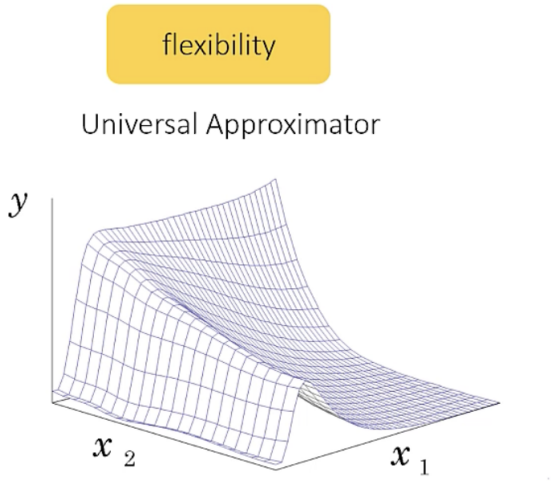


Figure 2.10: Representation of an NN with two inputs (Data source: SAS)

Although they are a very powerful tool to overcome many limitations compared to other models, they also have limitations, in particular the lack of interpretability, which is the basis for the well-known black box objection against neural networks. However, there are methods that make it possible to open the black box. One possibility is decomposition, in which the parameters in a neural network are approximated by a set of IF-THEN rules.
A neural network is constructed in layers that contain neurons or units. A NN is made of 3 main elements: Input Layer, Hidden Layers, and Output Layers [34]. The simplest example of a Neural Network is the multilayer perceptron (MLP), which consists of only three layers: an input layer, one hidden layer, and a target layer. Each unit in the input layer is connected to each unit in the hidden layer, and each unit in the hidden layer is connected to the unit in the output layer.
The hidden units contain an activation function, i.e. a mathematical transformation that is applied to the input layer. However, what makes neural networks so

interesting is their ability to approximate virtually any relationship between the inputs and the target.
A neural network uses a numerical optimization method to estimate the weights that minimise the error function. This process is called learning.
The weight estimates for a binary target are generated by minimizing the error. The error function and the weights in the model define the surface of the error space. The error space can also be referred to as the error surface. The surface can have many valleys.
The goal of numerical optimization is to minimise the error function. This is the same as finding the lowest point in the error space, which is called the global minimum (see Figure 2.11).

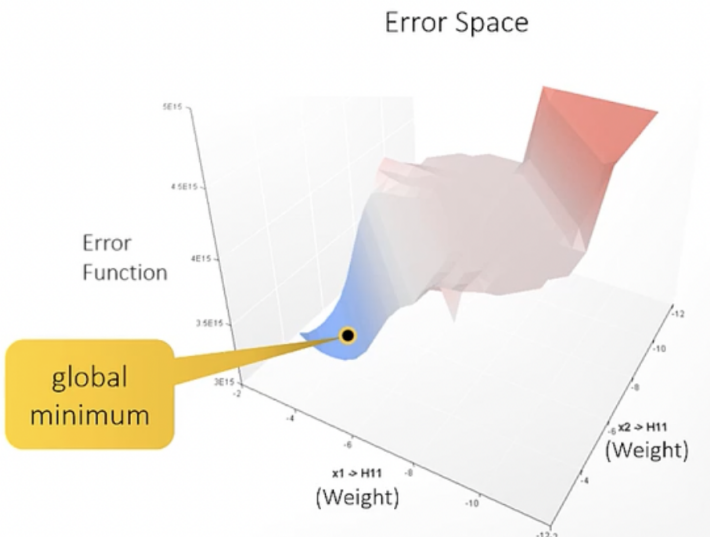
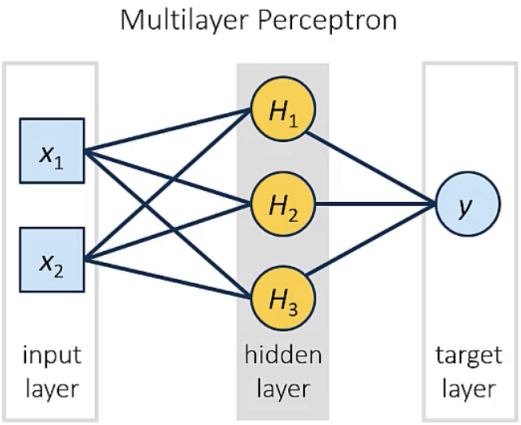


Figure 2.11: Error space representation (Data source: SAS)

Formally, a global minimum is a set of weights that produces the smallest set of errors. A simple strategy to ensure that a global minimum has been found is to perform an exhaustive search of the error space.
Unfortunately, the curse of dimensionality quickly makes this method infeasible. It is more efficient to use a heuristic numerical optimization algorithm to search for the optimal weight estimates.
Numerical optimization algorithms use local features of the error surface to decide how to proceed. However, by using local features, these algorithms are prone to getting stuck in treacherous regions of the error surface that are unlikely to lead to improvement. These regions can be called error plateaus or local minima. The learning process stops when the numerical optimization algorithm determines that it has learned the data well enough. Some modelling situations require a more

flexible and neural network architecture than a simple MLP 2.12.
Neural networks can have additional layers, additional neurons in each layer, and different types of connections.A second layer of hidden units can improve the performance of the model as this allows the MLP to implement discontinuous input-output mappings. On the other hand, adding a second hidden layer also increases



Hidden Units	Result
Too many	• Models noise • Fails to generalize
Too few	• Fails to capture the signal • Fails to generalize

Figure 2.12: MLP pattern (Data source: SAS)

the number of weights considerably. Adding hidden units can improve the performance of the model. However, if the network has too many hidden units, it will model random variations or noise as well as the desired pattern or signal. This means that the model cannot be generalised. Conversely, too few hidden units cannot adequately capture the underlying signal and the model cannot be generalised either.
The optimal number of hidden units is problem dependent. The best way to determine the appropriate number of hidden units is empirically, starting with a default NN and measuring its performance against an appropriate metric. Then increase the number of hidden units by one and observe the impact on the network's performance. Add more hidden units until the performance of the network decreases. As the final model, select the last network that was created before the performance drop. This final model has the optimal number of hidden units.
The network is trained using maximum likelihood estimation to determine the weighting estimates. This is done using a backpropagation algorithm that moves backwards through the network, and updates the weight values to reduce the error in the training data. Once the weights have been trained, the model can be used to classify new data points by simply plugging the new inputs into the model equations. This generates a probability that can be used to classify the new data point. Unlike other models, the neural network generates decision boundaries that are fundamentally non-linear. The ability to model arbitrary nonlinear functions is

very powerful and allows for much more complicated decision boundaries, but this added complexity comes at a cost. Because neural networks are so flexible, it can be difficult to choose the right network architecture or weight training procedure.

Iterative updating in numerical optimization

Numerical optimization algorithms optimize the weighting values by iteratively updating the weights until the algorithm finds a minimum in the error space. The error space is defined by the error function and the weights in the model. The iterative updating process 2.13 comprises four main steps, which are described below:

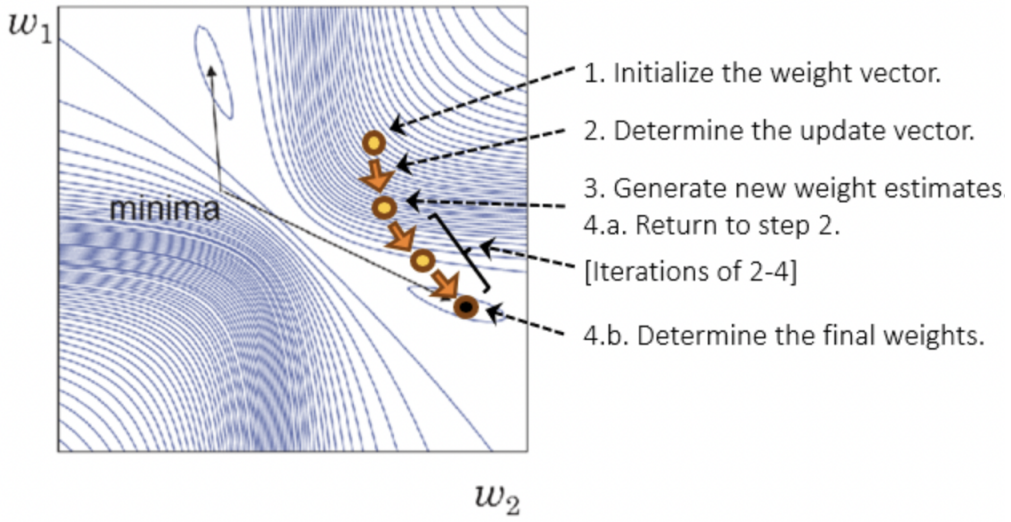


Figure 2.13: The four steps of the iterative update process (Data source: SAS)

1. Initialise the weight vector with small random values. This weight vector represents a point in the error space. (The coordinates are the values in the weight vector).

2. Use a numerical optimization method to determine the update vector (i.e. the values that are added to the weights). This update vector represents a downward step on the error surface.

3. Add the update vector to the weight values from the previous iteration to generate new weight estimates. By adding the values in the update vector to the original weight vector, a new point is reached on the error surface. This point represents new weight estimates for a model that has a lower value of the error function.

4. Determine if any of the convergence criteria are met and decide how to proceed.

If none of the convergence criteria are met, return to step 2 to determine new values for the update vector. If any of the convergence criteria are satisfied, exit the numerical optimization process. The final weighting estimates are for a model that minimizes the error function. These estimates represent one of the minima shown in the contour plot of the error space.

2.5 Model Selection

Among the different models, we must choose the one that generalises well. Choose the right trade-off between complexity, bias and variance, as the model may not be complex enough, leading to under-fitting where the signal is systematically missed. In this situation, the model has a high bias.
A naïve modeller might assume that the most complex model always performs better than the others, but this is not true at all. A model that is too complex might be too flexible, leading to overfitting that takes into account nuances of random noise in the sample. In this case, the model has a high variance. The goal of model selection is to choose a model that neither underfits nor overfits the data. A model with the right amount of flexibility provides the best generalization.

2.6 Data Splitting

It is important to find a way to evaluate the model and say whether the model generalises well. A common method is therefore to split the data into training, validation and test folds.
- The training subset is used during the learning phase. It is the largest subset of the data set and usually about 70-80% of the total data is used.
- The validation fold is used instead to validate the performance of the model during the training phase. It is very useful for obtaining unbiased evaluations during training and for tuning the hyperparameters. The main idea behind the construction of the validation set is to prevent overfitting of the model, i.e. the model becomes very good at classifying samples within the training set, but is not able to classify unseen data correctly. So it is used to get a feel for how well the model is performing. The most commonly used percentage is 20% within the 70-80% previously extracted for the training set.
- The test set: once the training experiments are complete and the parameters have been selected, it is important that you have a completely separate set of data that the model has never seen before. This data is called test data and is only

used for the final evaluation of the model. The proportion of data belonging to this set is usually 30-20% of the total dataset.

2.7 Model Assessment

Not always the best algorithms are necessarily the ones that have achieved the highest reported accuracy. Most algorithms usually require careful tuning and extensive training to achieve the best achievable performance. Choosing the modelling algorithm for machine learning application can sometimes be the hardest part. The decision of which algorithm to use can be guided by answering some key questions proposed by [35]:

1. What is the scale and type of data you have?
2. What do you want to achieve with your model?
3. How accurate does your model need to be?
4. How much time do you have to train your model?
5. How interpretable or understandable does your model need to be?
6. Does your model have a function to automatically set the hyperparameters?

The optimal setting of the hyperparameters is extremely data-dependent. Therefore, it is difficult to establish a general rule about how to identify a subset of important hyperparameters for a learning algorithm or how to find optimal values for each hyperparameter that work for all data sets. Controlling the hyperparameters of a learning algorithm is very important because proper control can increase accuracy and prevent overfitting.

2.8 Performance Metrics

A wide range of performance measures can be used in a classification problem. Depending on the problem and dataset we are facing, choosing the right metrics is essential to fairly compare classifiers and identify the right "champion model", i.e. the best predictive model. These may include:
- Accuracy is the proportion of predictions that the model classified correctly. Mathematically, this means:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

- Precision which is the number of positive predicted instances that the model was able to recognize. High precision is equal to a low false positive rate [36]. This can be represented mathematically as:

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

- Recall is the ratio of positive class predictions with respect to the total positive examples inside the dataset [36]. Mathematically expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

- F1-Score which is the weighted mean of Recall and Precision, by taking into account both false negatives and false positives [36]. Mathematically, this means:

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.13}$$

- ROC Receiver operating characteristic is a graph that relates the sensitivity and specificity of a diagnostic test to the variation in the cut-off value. Its analysis makes it possible to evaluate accuracy and determine the most appropriate cut-off value. By increasing the cut-off value, the number of false-negative re-

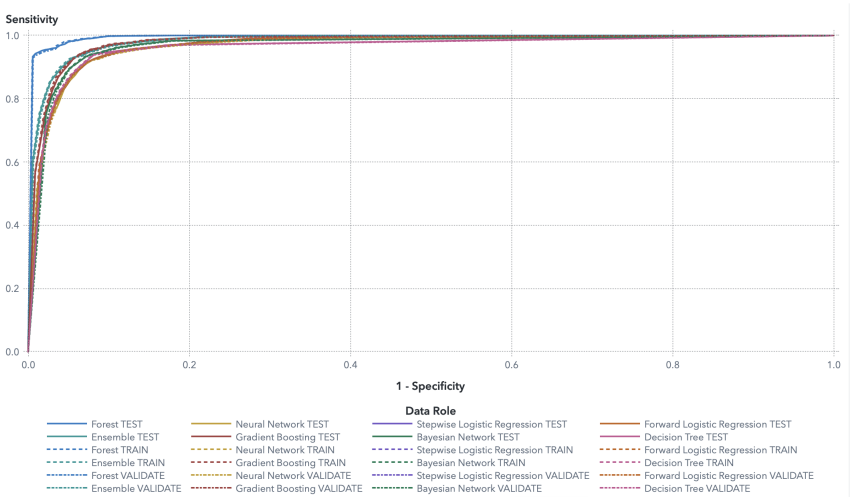


Figure 2.14: ROC curve representation developed in SAS Viya

sults increases, while the number of false-positive results decreases. Consequently, we have a highly specific but not very sensitive test. Conversely, a lower value increases the number of false positives while decreasing the number of false negatives, therefore we have a highly sensitive but not very specific test. It is useful

because it can summarise the required performances in a single graph. On the y-axis we find the rate of true positives (TPR = sensibility). On the x-axis, that of false positives (FPR = 1-specificity). An example of this curve is shown in image 2.14. For each cut-off value, there is a certain sensitivity and specificity value that corresponds to a point in the graph. Once all the different points are connected, the desired graph is obtained [37].
- AUC "Area Under the ROC Curve" if it is 0.5, it means that it is not informative, i.e. the curve would correspond to the bisector and the test is not able to distinguish the survivors from the dead in our case. While, if it is between 0.5 and 1, the test gradually becomes more and more accurate [37].

Chapter 3

Related Work

In the absence of vaccines and lack of information on various aspects of the spread of this disease, governments and researchers began their struggle with great uncertainty. Data from countries such as China, where the virus had initially spread, were considered a valuable source of material for the scientific communities fighting the virus. Over time, however, the amount of data available has become much more consistent, and more policies can now be formulated based on evidence of the "curve" and "peak" of contagion. However, controlling infections caused by this pathogen is a time-critical challenge and the main goal of all those involved in these studies is the health of the world's population.
For this reason, research has focused on finding effective boosters and taking public health measures to contain the spread of the pathogen and restore the desired normality in the period before COVID. In most parts of the world, these measures took the form of lockdowns, forcing people to stay at home and leave only for essential needs, such as work, medical care or taking emergency rations. Although the lifestyle changes required to deal with the virus have had an impact not only in the medical sphere, but also in the economic and social spheres.
In fact, in many cases, the attempt to control COVID-19 has led to a backlog of other medical procedures [38] and a shortage of those procedures. Health care providers have had to find a balance between trying to test and diagnose infected people.
Fortunately, Artificial Intelligence (AI) has proven to be very reliable in the medical field in recent years, which is why the scientific community has been trying to use it as much as possible since the outbreak of the pandemic. Under the term AI we can find, for the present purposes, different branches: Machine Learning(ML)[2] , Natural Language Processing(NLP) and Computer Vision applications that teach

[2]An important class of techniques commonly used is under the name Deep Learning(DL)

the computer to use Big Data based models for pattern recognition, explanation and prediction [4]. This section describes the current literature on Machine Learning, Statistics and COVID-19. The review of the literature is done in the database of Google Scholar, PubMed and IEEE under the keyword COVID-19 Statistical Analysis or COVID-19 and Artificial Intelligence.

3.1 AI applications for SARS-CoV-2

The WHO has created a framework that describes the outbreak of an infectious disease in six different phases [39], briefly shown in Figure 3.1.

1. "Identification": a new virus or bacterium is discovered.
2. "Recognition": where the disease clusters are placed around the world.
3. "Initiation": if there is a permanent transmission to humans.
4. "Acceleration": the number of cases start to increase and several methods are used in order to constrain the disease.
5. "Deceleration": The methods adopted in the 4 phase, lead to a levelling off and a subsequent decline of the number of cases.
6. "Preparation": As the spread slows down, the world entire in this step to prepare itself for the next wave of the pandemic.

AI has the potential to address challenges in all of these phases. A proof of this is that, when there was no potential vaccine, one of the most important technologies that played a crucial role in combating the pandemic was AI [40]. It was seen as helping to contain the spread of the disease through contact tracing, social distancing, quarantine monitoring, trend analysis, symptom reporting and analysis, symptom clustering, symptom severity estimation, disease spread modelling and alerting [41]. AI models have greatly helped in identifying the transmission routes of this virus and in its containment [42], although several issues arise in their use (e.g. initial unavailability of large datasets, ethics related to features, privacy, security, etc.). Accurate prediction of the epidemic was indeed quite difficult in the beginning due to the lack of historical and unbiased data needed for training.This lack of data was explored by [4] in the countries shown in Figure 3.2 by identifying six possible areas where AI could contribute (see Figure 3.3) In the same year, another study by [5] identified seven major application areas of AI for the COVID-19 pandemic, which are reproduced in Figure 3.4. Despite the different labels used by the authors, the areas identified are the same and in a few cases they may be complementary.

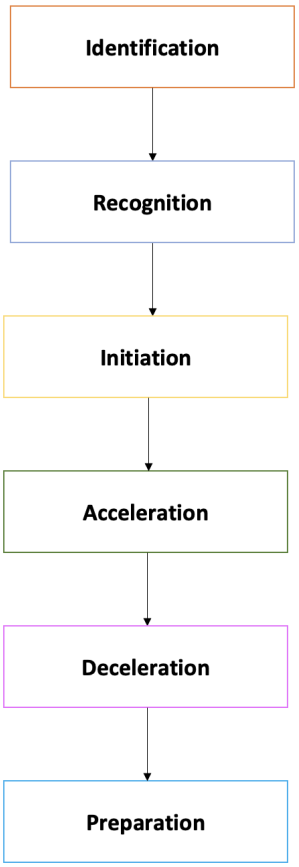


Figure 3.1: The six stages describing the outbreak of an infectious disease

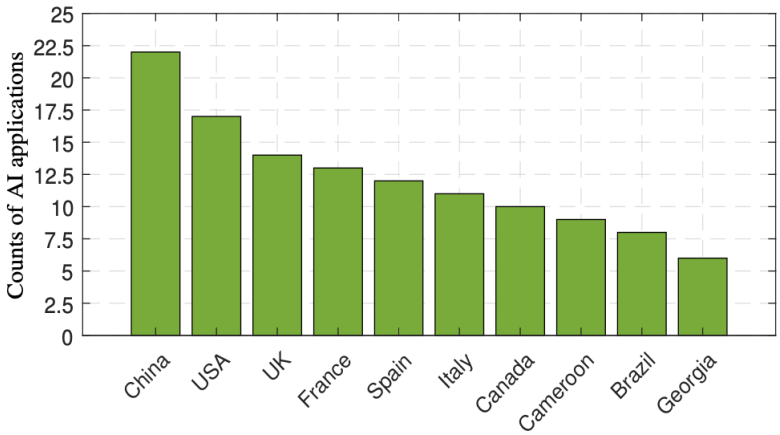


Figure 3.2: Top 10 countries using AI COVID-19 (Data Source:[4])

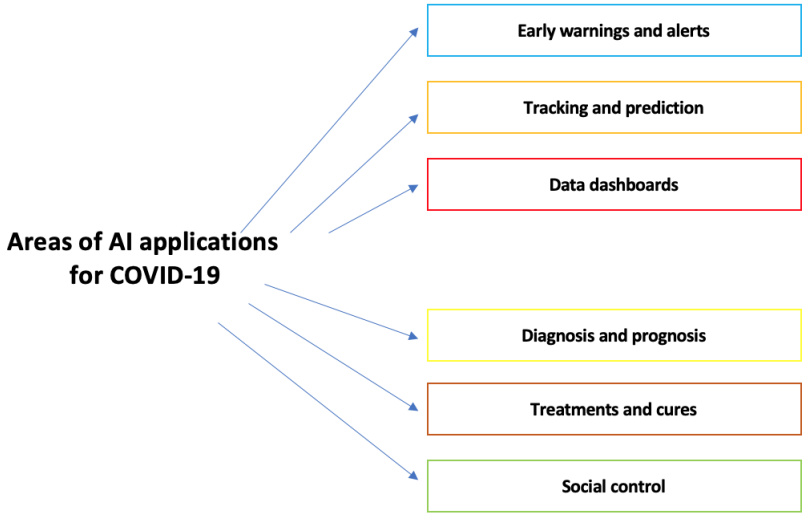


Figure 3.3: Possible AI application areas for COVID-19 (Data Source:[4])

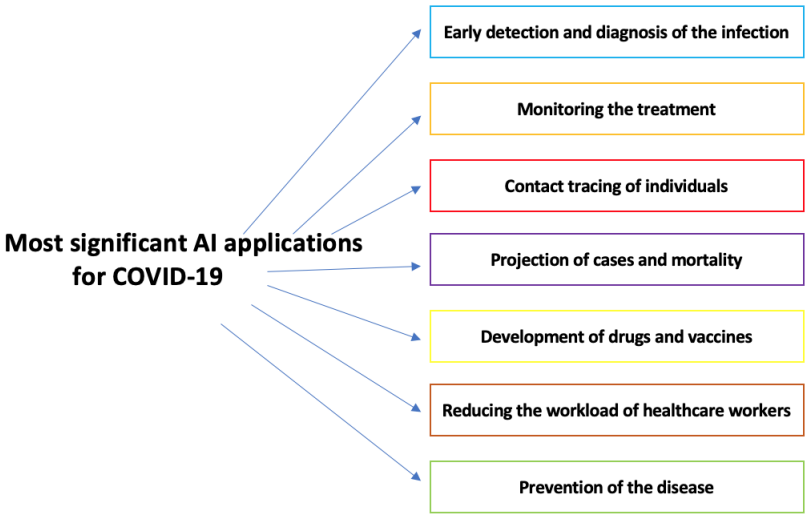


Figure 3.4: Snapshot of key AI applications for COVID-19 (Data Source:[5])

1. Early detection and diagnosis: AI plays an important role as it is useful in detecting irregular symptoms. The early phase of its use in this area concerns diagnosis based on chest X-rays. According to a number of studies such as [43], AI can be as accurate as humans. Another example is [44], in which Dr Rosebrock offers guidance on Deep Learning to automatically detect COVID-19 in a hand-generated X-ray image dataset. So if AI is used in an appropriate way, it can be useful in making decisions that can reduce the overall costs

borne by governments.

2. Monitoring treatment: Neural networks can be used to monitor and generally treat affected individuals.

3. Contact tracing of affected people: AI can help in contact tracing and monitoring of individuals.

4. Projection of cases and mortality: AI can predict the number of positive cases and deaths in each country and region. This is helpful in identifying the most vulnerable regions and people and taking the appropriate countermeasures.

5. Drug and vaccine development: Even before this pandemic exploded, it was claimed that AI provides an insightful glimpse into the discovery of new drugs (see [45]). In the case of SARS-CoV-2, AI was used to analyse the available data for developing appropriate vaccines faster than usual [46].

6. Reducing the workload of healthcare workers: AI helps doctors in the early diagnosis phase to distinguish COVID-19 symptoms from other diseases [47].One example is [43], which says that using AI in X-rays could help radiologists save time and make diagnoses that are faster and cheaper compared to standard tests.

7. Prevention of the disease: using real-time data, AI helps monitor the spread, check the intensive therapy beds needed in a place to avoid overcrowding and unpreparedness on the part of the hospital [48,38], and use thermal images to scan public places to detect possibly infected people and enforce social distancing and possibly lockdown measures [49].

8. Data dashboards: The tracking and prediction of this virus has sparked a veritable sector that creates data dashboards to visualise actual and expected spread (see for example [50]). They are very useful as they provide a general overview of the world and also at country level, as an increasing number of countries already have their own dashboards. Tableau, which is one of the most important visual analytics platform, has also created a COVID-19 Data Hub with a COVID-19 Starter Workbook [3] .

Thanks to this overview, it has been shown that some of the limitations of using AI (e.g. the unavailability of open source data or the lack of historical data needed to train the models) are now partially solved. However, this raises the data deluge problem. It is not always true "the Big Data Hubris" which believes that huge

[3]https://www.tableau.com/covid-19-coronavirus-data-resources

volume of data always leads to better results. Authors Lee et al [51] say: "The conventional doctrine in Machine Learning is that it is always beneficial to collect more training data". This is true if the data collected represents the same underlying phenomenon, but in medicine, it is often difficult to make this assumption, mainly due to the huge variability in patient/clinical characteristics as well as our limited understanding of complex human health and disease trajectories. This is particularly true in the case of COVID-19, where there is much doubt about the characteristics that determine the incidence of severe cases across the population. This is exacerbated today by the presence of many variants. We can no longer even speak of much correlation between the age of a patient and the disease, since the people most susceptible to the latter variants are the children. The presence of outliers in the data and the enormous amount of scientific evidence must be assessed before diagnostic and treatment options are offered [52] . In the end, the scope of AI can be summed up in a single phrase: "Use Big Data to understand the unknown"[52].

3.2 The Role of "Open Science"

This unprecedented period has also seen the emergence of an interesting phenomenon: the emphasis on the fact that science must be conducted in such a way that it is transparent, reproducible and robust. The concept that knowledge must be public and freely accessible to everyone has always been known. This is especially true when it comes to technological knowledge with many applications, as in the case of the search for a solution against SARS-CoV-2. In fact, it offers a lot of advantages in this way over the more traditional thinking of a closed form of knowledge and the reasons for this could be many [53]:

1. any interested party can look at the actual implementation of the code and eventually criticize, add to or even contribute to it;
2. improving transparency;
3. ensuring better results through public scrutiny;
4. certainty about the reproducibility of the results achieved.

Concrete examples of this are the internet, e.g. the flood of open source powered dashboards [53], open data repositories, etc. Another example is the number of scientific papers related to COVID-19 published since the start of the pandemic [54], the amount of data and tools developed to track the evolution of the virus, etc. [55]. Given the increasing number of publications, some community initiatives have used Machine Learning techniques to help identify the most relevant sources ([56], [57]).

3.3 The Pandemic Numbers

Algorithms and models have made it possible to understand much about SARS-CoV-2, often in advance. But for the future, there is a need to increase predictive capacity. It is clear from the other sections of this chapter that the role of ML and AI in general is very important in confronting and responding to this disease. First, the mathematical models based on the available data describe the spread of the epidemic. These models use real-world data: the demographic structure of the population, the mobility networks, the clinical course characteristics of the disease. The computer becomes the laboratory where data are generated that can then be used to understand the characteristics of the pathogens, plan effective measures to control the spread and make predictions. The epidemiological mathematical work and the computational work are additional tools. They are no substitute for the real heroes of this pandemic, namely the medical equipment, the health workers and the many volunteers who are still on the ground. The common divisor is common, but mathematics and AI are fighting a different battle, one that can be managed through numbers and information.
First, it is important to distinguish between the different types of methods that can be used to analyse epidemiological data. On the one hand, there are mechanistic models, in which numerous numerical simulations are carried out in the background using powerful supercomputers, and on the other hand, there are models and data that describe the transmission processes of viruses between individuals. These types of models are called mechanistic because they explicitly describe the mechanism of the infection process and the transmission of the virus from one individual to another. The more detailed the people, their relationships and the factors involved in the transmission of the virus, the better the model can represent reality. In addition, these types of models have the advantage of allowing the identification of equations and algorithms that describe the dynamics of the disease in a population. For example, a fairly relevant number is the reproduction number R_0, which indicates how many people on average can contract the disease from each infected individual.
On the other hand, there are models that are based on the lecture of statistical data without assuming any individual-level processes that determine the contagion phenomena. AI algorithms are found in the most sophisticated of these models. The core idea of these approaches is that the algorithm itself learns to recognise precise relationships and patterns in the data and automatically adapts to "learning by doing" by processing the information. Each type of model has its own advantages and disadvantages depending on the type of data available and the information we are interested in. Different mechanistic techniques and approaches are used depending on the different phases of the disease, the level of knowledge about the mechanisms of infection and the extent to which the data have been

analysed. Then, several initial conditions are simulated using the computer and different parameters of the models are varied within the experimental uncertainty interval. This leads to a collection of plausible trajectories that map the likely next evolution of the disease and describe the probability of events that could occur in the future. It should be noted that uncertainty is always part of the prediction [58].
Another important fact is that all models used at the scenario simulation level and for predictions differ from each other because they use data, techniques or approaches that are very different from each other. Therefore, relying on only one model is never a good choice. The solution is to use what is called the "ensemble multi-model" technique. It consists of combining the results of several models to improve the predictions and make the confidence interval more reliable. This type of model is used extensively by the CDC (Centers for Disease Control and Prevention) in the US, which has used more than 20 models from different independent research groups to predict deaths and hospitalizations. Shuai Wang et al. has identified the radiological changes in CT images of patients with Sars-Cov-2 in China. In this research, he has used Deep Learning methods to extract the graphical features of COVID-19 from the CT (Computed Tomography) scan images and develop it as an alternative diagnostic method. They have collected CT images from confirmed patients as well as patients diagnosed with pneumonia. The results of their work provide a proof of principle for using AI to accurately predict COVID-19 [59]. This research uses CT scan images, which is different from my research because clinical features and laboratory results are used for prediction.

3.4 The pandemic underlined the absence of relevant facilities for the data analysis

"Data is the new oil", as Humby said in 2006. Proof of this is that in today's world, the amount of information and the availability of data is increasing by the day.
But at the same time, as Palmer later said, this sentence was incomplete. It reflected reality, but at the same time there was still a piece missing: "Data is valuable, but if it is not refined, it cannot really be used".
We live in a world where we are surrounded by a multitude of information that is constantly increasing. This results in the need to implement methods that enable us to process it better and transform it into useful information. This continuous evolution requires constant development and integration of data analysis tools.
To meet the demand for data analytics in healthcare, machine learning approaches have become necessary prerequisites [60, 61,62,63]. To facilitate the assessment and modelling of COVID-19 development, ML models have been developed. Liu et al. use ML techniques to predict the spread of the disease in Chinese province [64],

Wang and Wong used a neural network on chest X-ray images to identify it [65]. Beck et al.[66] proposed a list of drugs to potentially combat SARS-CoV-2 using a deep learning model. The ability to predict goes hand in hand with the quality of the data. Unfortunately, this disease shows that even the simplest data collection has some limitations. After all, the data are usually affected by various issues that can change from one state to another, in terms of time period, frequency of testing and delays in reporting by the health system. This has been particularly the case during the stressful period of the pandemic.
So having a permanent infrastructure that works to maintain data quality and reliability is a chore. It is about getting not only public health data, but also information on mobility statistics, genome sequences of pathogens, policies implemented by individual states and measurements of their impact [58]. Instead, during this terrible event, reality has shown that important infrastructures such as these are missing. These facilities are a basic necessity for all humanity because, if built in the right way, they can be a fundamental tool to support governments' decisions by using the best available evidence.

3.5 Mortality prediction in previous works

The importance of ML and DL in many fields has encouraged many researchers to use these methods extensively in the course of the pandemic for automated diagnosis, prognosis and development of tools to predict mortality in severe cases. The Random Forest approach in ML has been repeatedly shown to be accurate in modelling COVID-19 outcomes, using an ensemble of decision tree classifiers to generate predictions.
Ko et al [67] developed an Ensemble-based Deep Neural Network (EDRnet) to predict in-hospital mortality from routine blood samples administered on admission. The dataset consisted of 73 blood biomarkers from 467 COVID-19 patients. A total of 28 blood biomarkers along with age and sex information were selected using analysis of variance (ANOVA) and available data selection (ADR) techniques. An ensemble approach was adopted, combining RF models with a NN that achieved a sensitivity of 1, a specificity of 0.91 and an accuracy of 0.92. In order to obtain additional patient points, the authors also created an online web application (BeatCOVID-19) to predict mortality from blood test results.
Similarly, in [68] the authors applied ML techniques to clinical data from COVID-19 patients in treatment at Mount Sinai Health System (MSHS) in New York City to predict mortality. The training set consisted of 3841 patients, of whom 313 are deceased and 3528 are still alive. The validation set consisted of 961 patients, of whom 78 are deceased and 883 are still alive. Finally, the test set consisted of 249 patients, of whom 25 died and 224 are survived. Recursive feature elimination was

used for feature selection. In addition, several ML models such as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost) were applied and achieved an Area Under Curve (AUC) value of 0.91. It was found that patient age, minimum oxygen saturation in the course of medical treatment and inpatient status were the most important predictive characteristics. The only constraint of this study is the limited number of attributes. The dataset in fact contained only those features which are daily collected during hospital cure. Another study by Das et al [69] in South Korea to forecast mortality in COVID-19 patients with ML. They make use of a public dataset produced by the Korea Centers for Disease Control and Prevention, with 3524 COVID-19 patients from 20 January to 30 May 2020. Five ML models: k-nearest neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB) were used.
LR outclassed the other models with an accuracy of 0.96 and an AUC of 0.83. The limitation of the proposed application was that no clinical information on COVID-19 patients was available. Moreover the authors did not use a hold-out set that had not been previously used for validation thus this might introduce overfitting.
Another study conducted by the scholars [70], used the Korean data to implement ML models to predict the prognosis of COVID-19 patients. The dataset contains the clinical and demographic information of individuals in South Korea. It was created using a combination of different databases of patients with a confirmed positive COVID-19 test. Among them there were 10237 patients, 7772 recovered, 228 died, and 2237 were still receiving treatment. Different models were used, such as LASSO, SVM (Linear), SVM (Radial Basis Function (RBF)), RF and KNN. The top models were LASSO and SVM (Linear). LASSO reached an AUC of 0.963 and accuracy of 0.911. With SVM, L_1 norm feature selection techniques were used to select the most important features. Nevertheless, SVM (Linear) achieved an AUC of 0.962 and an accuracy of 0.919. The study found a strong association between mortality risk and characteristics such as age, sex, disability, previous symptoms, diabetes mellitus and asthma.
In comparison, researchers in [71] analysed the cases of COVID-19 in Madrid using ML and survival analysis techniques to predict mortality. The dataset came from the HER system of HM hospitals and contained 29 variables from admission and clinical data of 2307 patients. Different data analysis methods were exploited, such as LR, Bayesian Network (BN), survival data analysis, DT, RF and bi-clustering. LR achieved the best result with an AUC of 0.89, sensitivity of 0.81 and specificity of 0.81. The study confirms that older people have an increased risk of dying from COVID-19.
In addition, decision rules for predicting the risk of death in COVID-19 patients can be derived from the DTs to help clinicians triage COVID-19 patients. They could identify different patient groups through unsupervised learning, allowing

global analysis of drugs distributed to patient populations. However, the study is limited to data from a specific population collected under complex health conditions.
The researchers in [72] analyzed data from 1955 COVID-19 patients in different regions of Spain. Age, sex, oxygen saturation and the rate of change of both haemogram ratios VNLR (neutrophil-to-lymphocyte ratio) during week 1 from the admission date were used. LR showed the best performance in the analysis of VNLR with an AUC value of 0.891. Ferreira et al [73] analysed patient data from the Portuguese Directorate General of Health by applying different models such as LR, NB, DT and Artificial Neural Networks (ANN). These models aim to predict COVID-19 the outcome of patients whether they recovered or died. The best results were obtained by using all available information on comorbidities, age and symptoms together with oversampling technique and the DT algorithms. A sensitivity of 0.95, accuracy of 0.90 and specificity of 0.86 were achieved. Furthermore, a study [74] used the dataset COVID-19 from 146 countries to detect mortality rates. SVM, ANN, RF, DT, LR and KNN were used. The results of the study show that the model ANN achieved the highest accuracy of 0.8998 with 57 features.
In [75] authors used RF to predict the risk of death in COVID-19 patients. The dataset contains 567 individuals. Gini importance criteria were used with RF to select the most important characteristics. The model achieved an accuracy of 0.655 and an AUC of 0.855. However, the model needs validation and its generalizability is limited as variables such as treatment and intervention are not considered.
In [76] to predict Intensive Care Unit (ICU) admission and mortality in COVID-19 patients using DL and a risk scoring system. There were clinical information from 1108 patients, of whom 837 were for hospital admissions and 271 for the ICU. Of the 837 general admissions (772 were released and 65 died), 271 admitted to ICU (86 were still in ICU, 108 were discharged and 77 deceased). All patients were at Stony Brook University Hospital in New York. Feature ranking was performed using Random Forest to rank the features. The top predictors identified were then fed into a DL model consisting of five fully connected dense layers. The activation function ReLU was used for the hidden layers, while the sigmoid activation function was used for the output layer. To build the risk score model, the generalised additive model was used to graph and illustrate the probability of ICU admission and mortality for each independent clinical variable. For the prediction of mortality, the model achieved an AUC of 0.844 and an accuracy of 0.853. It was found that the significant characteristics for mortality were age, LDH, CRP, cardiac troponin and SpO2. Comorbidities were not a relevant predictors of ICU admission and mortality.
However as it was previously written in the description of RF, the algorithm has certain limitations that can make understanding the model itself difficult. These

limitations could be exacerbated in understanding this disease, as the impact on individuals may be different. For example, if individuals belong to different age groups, as it has been shown that this type of feature is one of the most important in understanding whether a person would die or not [77], [78]. This insight into using patient demographic characteristics to predict health outcomes is a current research trend known as Personalized or Precision Medicine [79, 80, 81, 82].
Iwendi et al. in [29]implemented a Random Forest model boosted by the AdaBoost algorithm. This classifier obtained an accuracy of 94% and an F1-Score of 86%. Their analysis revealed a positive correlation between patient gender and death.
The last and fundamental work, thanks to which I was able to take up my research activity, was the one carried out by [83]. His dissertation was conducted at Illinois State University and uses the same dataset exploited in this research, extracting 10,000 individuals. He demonstrates how the RF classifiers achieves very high performance, reaching an accuracy of 0.95.
For a fast summary of the related studies analyzed in this last part of the literature review, see the table 3.1) on the next page. It shows only those works that are more closely related to the study conducted in this thesis.

Table 3.1: Overview of related mortality studies

References	Technique	Dataset Size	Feature Selection	Results
[84]	RF	176 COVID-19 patients	-	ACC: 0.87,AUC:0.91
[85]	RF	5644 COVID-19 patients	-	ACC: 0.92,AUC:0.98
[86]	RF	2342 COVID-19 patients	RF importance algorithm	ACC: 0.83
[87]	LASSO,SVM,KNN,RF	10237 COVID-19 patients	-	AUC:0.96
[88]	RF	126 COVID-19 patients	Recursive feature elimination	AUC:1.0
[29]	Boosted RF	60000 COVID-19 patients	Recursive feature elimination	ACC:0.94,F1 Score: 0.86
[89]	LR, RF, XGB	287 COVID-19 patients	Extra tree classifiers	ACC:0.95,AUC: 0.99
[67]	EDRnet	467 COVID-19 patients	ANOVA, ADR	ACC:0.92
[68]	KNN,LR,SVM,RF and GB	3841 COVID-19 patients	different types	AUC: 0.91
[69]	LR	3524 COVID-19 patients	-	ACC:0.91,AUC: 0.96
[70]	LR	10237 COVID-19 patients	L1-norm	ACC:0.96, AUC: 0.83
[71]	LR,KNN,BN,DT,RF	2307 COVID-19 patients	L1-norm	ACC:0.96,AUC: 0.89
[72]	LR	1955 COVID-19 patients	-	AUC:0.891
[73]	LR, NB, DT,ANN	- COVID-19 patients	-	ACC:0.90
[74]	SVM, ANN,RF,DT,KNN	2307 COVID-19 patients	-	ACC:0.96,AUC: 0.89
[75]	RF	567 COVID-19 patients	Gini imp criteria	ACC:0.655,AUC:0.855
[76]	SVM,ANN,RF,DT,LR,KNN	146 countries COVID-19 patients	-	ACC:0.89
[29]	Boosted RF	- COVID-19 patients	-	ACC:0.94
[83]	RF	10000 COVID-19 patients	-	ACC:0.95

Chapter 4

Proposed Methods

4.1 Software environments

With the advent of digitization, the excessive use of Big Data has increased. Many programming languages such as SAS and Python are capable of separating usable data to meet today's market demands. As a result, many companies rely on these programming languages for data analysis.
The war between SAS and Python is untamed and in order to rationally decide which tool is ideal, an analyst must first understand the pros and cons of each.
My mentor has talked to me about SAS since I started this work.
So, in order to form a fair and personal opinion about it, we decided to give me the opportunity to use the already familiar Python environment for part of my work and then compare it with the new SAS Viya to see with my eye how different they are and how specifically SAS Viya works.

4.1.1 Python

For the first part of the experiments I used the Python programming language and the Jupyter Integrated Development Environment (IDE).
Python is a sophisticated and effective programming language for general use and has a large standard library that provides tools for performing various tasks. The following Python libraries were used during the work:
- Pandas: It is a Python package that provides expressive data structures that can work with both relational and labelled data. It is an open source Python library that allows the utilization of DataFrame object for manipulation of data arranged by columns [90].
- Numpy: It is an open source Python package for scientific computing. It is extremely useful for working with arrays and also provides functionalities for working

in linear algebra, Fourier transform and matrices [91].
- Matplotlib: It is an open source Python package used for creating plots and 2D representations [91].
- Tensorflow: It has a comprehensive, flexible ecosystem of tools, libraries and community resources that enables the creation and development of DL based applications [92].
- Scikit-learn (sklearn): It is an open source Python machine learning library developed to complement Numpy. It provides various machine learning algorithms for classification, clustering and regression [93].

4.1.2 SAS Viya

For the second part of my thesis I used the potential of a new and great environment, namely SAS Viya.
SAS Viya is an extremely powerful data analytics framework that helps organisations accelerate their data analytics capabilities. The goal of SAS Viya is to help companies access Big Data analytics and accelerate the process at every step from data collection to data delivery [94].

4.2 Dataset

Data collection was a fundamental, difficult and lengthy process. Regardless of the research area, accuracy of data collection is critical to maintain cohesion. As the patients' clinical data were not publicly available, data collection was a rather complicated process.
At the beginning of the study, the aim was to find certain characteristics that could provide a possible explanation for the number of deaths in Italy and in America, and then, if possible, to discover some correlations between the two countries. To achieve this goal, I contacted various hospitals and health institutes in Italy to obtain the most accurate data possible, but no one was willing to provide it.
However, the Centres for Disease Control and Prevention (CDC), a major public health surveillance agency in the United States of America, gave me the opportunity to access their private dataset. The CDC is a United States federal agency that is part of the Department of Health and Human Services and is based in Atlanta, Georgia.
So my research focused on analysing US data to better understand within the dataset what key characteristics are important in understanding when one person with COVID is more likely to die than another.

4.2.1 CDC COVID-19 Case Surveillance Restricted Access Data

The dataset used to train the model to predict mortality was COVID-19 Case Surveillance Restricted Access Data[4]. This kind of data is collected by jurisdictions and, after anonymisation and removal of protected health information, is voluntarily shared by CDC for research purposes under a registration process and data use agreement. This database comprises patient-level data reported by US states and autonomous reporting agencies, including New York City, the District of Columbia, and US territories and affiliates. It contains anonymised electronic medical records (EMR) from over 32 million occurrences of people affected by COVID-19. The observations range in time from 1 June 2020 to 30 April 2022. The unidentified data consists of demographic and geographic information (county and state of residence), and presence of any underlying medical conditions and risk behaviors with a total of 33 features.The predicted mortality outcome in this work is defined as positive if a patient died with COVID-19, otherwise the outcome is marked as negative (alive). The attributes that were considered for the Machine Learning models are shown in Table 4.1.
The data classes are very unbalanced, i.e. the total number of deceased persons in relation to living persons was quite small. A recapitulation of this data in fact shows that only about 5% of the observed individuals are dead due to COVID-19. Data imbalance usually creates problem since it leads models to overfitting.
To obtain accurate and unbiased models, a balanced dataset with the same number of observations for recovered and deceased patients was created for training and testing the models.
The data samples (patients) in the training dataset were randomly selected and are completely separated from the test data. The training data was then further divided into two folds (training and validation) by using GridSearchCV which is a useful tool for tune the parameters of the model and at the same time performing k-cross-validation on data. More specifically, a 3-fold cross validation was performed. Then the parameter selected in this step, are used with test dataset. Figure 4.1 shows the exact number of the American population that died for each month of the observation period. As can be seen, the peak is reached in 2020 December and 2021 January. This event can be explained by the fact that permission for mass vaccination in the United States was granted on 14 December 2020 [95], but considering that 10 days later was the Christmas holidays a relevant number of people did not start vaccination until January. Moreover, as at least 14 days are needed for the vaccine to take effect, it is natural that picks takes place in these

[4]https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t

months and not before, since the virus has probably weakened during the summer months due to weather conditions and is then much more virulent in wintertime. In fact, from March ahead, we can see a decrease in the number of deaths, which could be due to both climatic conditions and, in particular, the effect of the vaccine previously administered. We can then look at how, unfortunately, from 2021 July onwards, deaths begin to rise again. By hypothesis, a plausible for this is that, this is probably due to the COVID variants spreading and by the fact that the vaccines that we have done previously, they were a bit inactive in relation to the new virus forms.
Looking at deaths by race, we can see that this general trend is respected by Blacks, Latinos, Asian and Pacific Islanders people (Figure 4.3, 4.4, 4.2, 4.5). The exceptions is done by Indigenous and White groups (Picture 4.7 and 4.6), in which the number of deaths decreased slightly from December to January, possibly due to the fact that people vaccinated themselves earlier compared to the previously cited races.

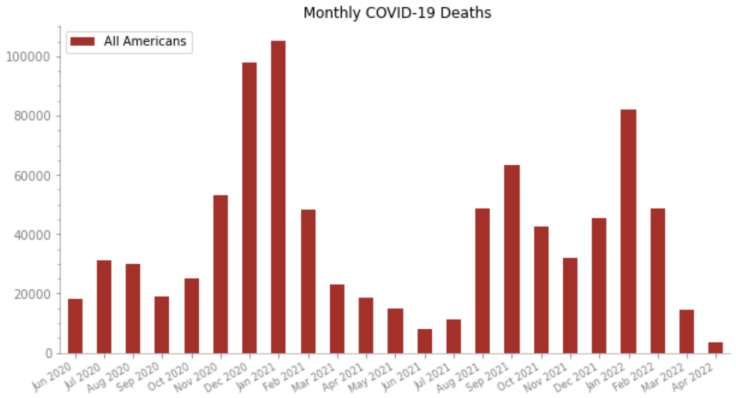


Figure 4.1: Monthly COVID-19 deaths in United States

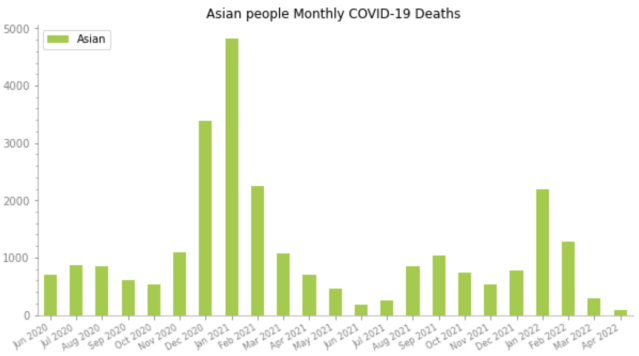


Figure 4.2: Asian race monthly COVID-19 Deaths in USA

Table 4.1: CDC COVID-19 Case Surveillance dataset description

Features	Description	Levels	Data Type
race ethnicity combined	Race and Ethnicity	8	Categorical
cdc_case_earliest_dt	Earliest available data for record	NA	Date
cdc_report_dt	Initial date case reported to CDC	NA	Date
pos_spec_dt	Date of first positive specimen collection	NA	Date
onset_dt	Date of symptom onset	NA	Date
sex	Patient sex	Male, Female	Categorical
hosp_yn	Patient hospitalized	Yes, No	Categorical
death_yn	Patient died	Yes, No	Categorical
medcond_yn	Patient had pre-existing condition	Yes, No	Categorical
res_state	State of residence	50	Categorical
age_group	Patient age group	8	Categorical
current_status	person current status	2	Categorical
county_fips_code	County FIPS Code	511	Numerical
res_county	County of Residence	398	Categorical
icu_yn	Patient admitted or not to intensive care unit	Yes, No	Categorical
hc_work_yn	Patient is a health care worker or not in US	Yes, No	Categorical
pna_yn	Patient developed pneumonia	Yes, No	Categorical
abxchest_yn	Patient had abnormal chest X-ray	Yes,No	Categorical
acuterespdistress_yn	Patient had acute respiratory distress syndrome	Yes, No	Categorical
mechvent_yn	Patient received or not mechanical ventilation	Yes, No	Categorical
fever_yn	Patient had or not fever(temp > 100.4 F)	Yes, No	Categorical
sfever_yn	Patient felt feverish	Yes, No	Categorical
chills_yn	Patient had or not chills	Yes, No	Categorical
myalgia_yn	Patient had muscle pain or not	Yes, No	Categorical
runnose_yn	Patient had or not rhinorrhea	Yes, No	Categorical
sthroat_yn	Patient had or not sore throat	Yes, No	Categorical
cough_yn	Patient had or not chronic cough	Yes, No	Categorical
sob_yn	Patient had or not dyspnea	Yes, No	Categorical
nauseavomit_yn	Patient had or not nausea or vomiting	Yes, No	Categorical
headache_yn	Patient had or not headache	Yes, No	Categorical
abdom_yn	Patient had or not abdominal pain	Yes, No	Categorical
diarrhea_yn	Patient had or not diarrhea	Yes, No	Categorical

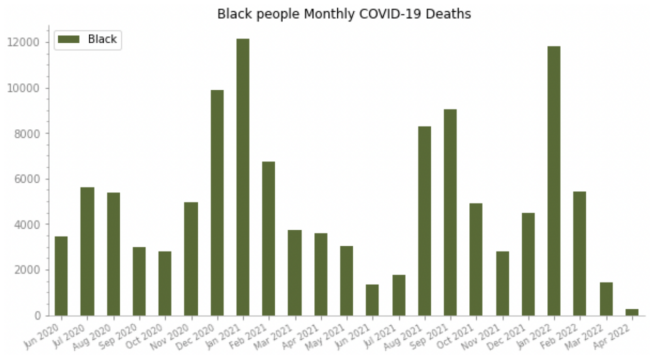


Figure 4.3: Black race monthly COVID-19 Deaths in USA

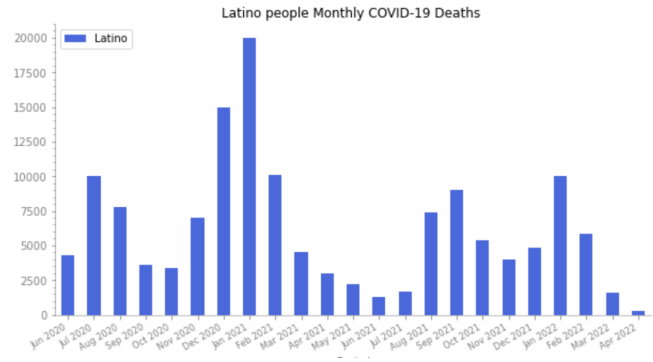


Figure 4.4: Latino race monthly COVID-19 Deaths in USA

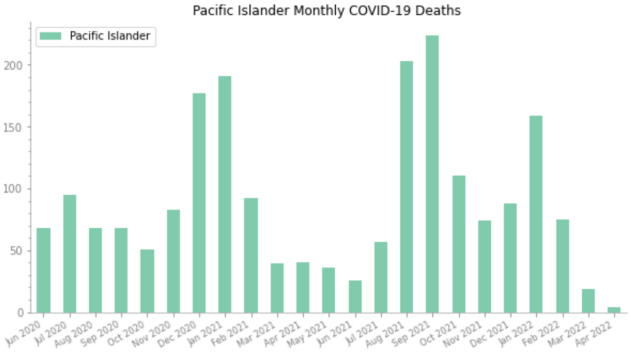


Figure 4.5: Pacific Islander race monthly COVID-19 Deaths in USA

4.3 Data Preprocessing

Preprocessing data is an important step in developing Machine Learning models. The data collected is often poorly controlled and contains many missing values that can be misunderstood by the model and considered noisy. For this reason, such data can falsify the result of the experiment.
- Imputation of missing columns: The CDC started entering data into their database as early as January 2020, but there was an inconsistency between the way data are registered before the month of June 2020. For this reason, I decided to omit the period from January to May 2020 by starting the analysis from June. At the beginning the dataset was contained in a zip file which, when unzipped, was about 16 GB in size. It again contained a folder for each month of each year analysed. Each month file was in turn divided into several csv files. In particular, each month, with the exception of those relating to the year 2020, had 4 to 8 parts each. Once these had been analysed to determine whether or not the attributes used were the same and consistent across all months, they were merged month by

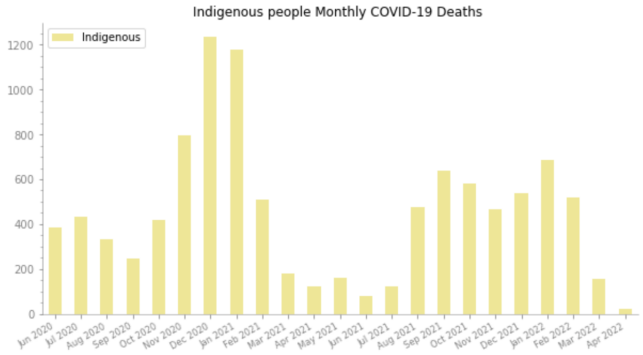


Figure 4.6: Indigenous race monthly COVID-19 Deaths in USA

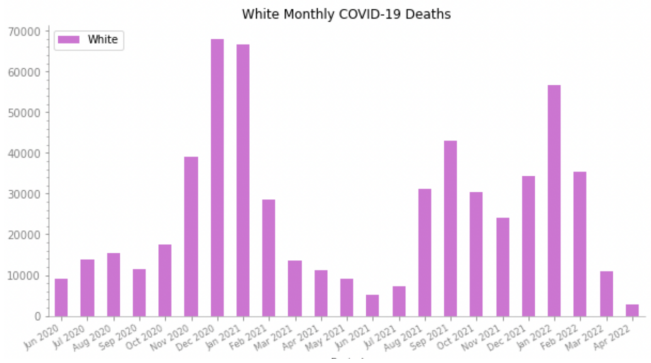


Figure 4.7: White race monthly COVID-19 Deaths in USA

month into a single csv file. As a final step, these monthly files were then used to create the Pandas dataframes.
- Imputation of missing values: After a long analysis, I have assumed that the data under analysis fall into the category of "missing completely at random": missing values appear randomly in the input data, so it is possible to omit the rows with missing values without distorting the model [96].
Usually, where this lack of data is found, are attributes that can be considered "more sensitive", such as the race or sex of the person. Therefore, I have assumed that these features are directly required as optional during the data collection process, rather than mandatory if the person does not wish to provide them. So under this assumption, missing values are not relevant to increase the predictive power of the model. Complete cases (the rows that are complete) are representative of all original cases. Consequently, patients with missing values are simply excluded from the analysis. Moreover, the percentage of complete cases is really high, it is about 85 % of the total cases. For the reasons previously mentioned, and given the abundance of the complete cases, rows with missing values are discarded before

random sampling the rows that are used to train, validate and test the models.
- Categorical data encoding: For the experiments conducted in the Python environment, the OneHotEncoder package was used, which encodes categorical features as a one-hot numeric array. Concerning SAS Viya, this program automatically recognises orkhich are the categorical variables involved and modifies them so that each model is able to use them, i.e. to understand them and to extract valuable information.

4.4 Sampling

Given the large number of individuals inside the original source, a dataset of 10000 samples was first extracted from the files. Then another sample of about 3 million individuals was created. The first sample consisted of 10000 individuals since the idea was to compare the results obtained by [83] with my work. For the second phase of the experiment, a much larger number of lines was selected to prove the quality and generalizability of the models. In both cases individuals are randomly selected by ending up with an equal number of dead and deceased individuals.

4.5 Feature Selection & Dimensionality Reduction

The work is structured in such a way that it was possible to analyse how the models work to detect mortality due to COVID-19 using a different number of features. This was done to understand whether certain types of characteristics are relevant to deceased detection or not. Therefore, four different types of experiments were conducted on the larger dataset:

1. The first experiment consists in the use of all the available features
2. The second experiment consists in keeping features that are not considered the most relevant by the Random Forest Algorithm
3. The third experiment consists of deleting the categorical binary attributes that are negatively correlated with the target variable
4. Use only the race, age group, sex and previous health status of each person to find out if they contain enough information to solve accurately our research question

This type of analysis has been carried out to explain which attributes are important in determining whether or not a person affected by COVID will survive. In

this way, there is the possibility of reducing the dimensionality and complexity of the task[97]. It is important to select only those attributes that contribute to the estimation of the target variable and disregard the irrelevant ones, as they might lead to different problems:
- overfitting: it may be that some of the variables included in the models are noise variables, so that the model cannot be validated with new future samples because it will perform poorly. The overfitting problem is essentially caused by multiple testing, where some noise variables are included in the model by chance;
- productivity: even if we are in a case in which all the attributes are relevant, it is important to consider some things, such as the amount of data available, the storage and computing resources, the time needed to complete the work, etc.). This makes it clear that using all of them is almost impossible for one reason or another;
- redundancy: dealing with many predictor variables increases the possibilities that there are hidden relationships between some of them, leading to redundancy and it can be a huge problem for the quality of the resulting model;
- model understandability: models with fewer predictors are easier to understand and explain. Since data science steps are performed by humans and results need to be understandable to humans, we need to take into account a sort of trade-off. This consists in giving up some potential benefits to the success rate of the data model, while ensuring that the data model is easier to understand and optimize.
Feature selection using Random Forest is called "embedded methods". These are characterised by the fact that they have their own integrated methods for feature selection. One of their greatest advantages is that they are very accurate, generalizable and easily interpretable. More specifically in the case of Random Forest by using boosting and bootstrapping techniques not every tree sees the same features or observations. This ensures that trees are decorrelated and thus less prone to overfitting.
The other type of analysis consists in discarding binary features negatively correlated with the output (see Figure 4.8) and also dropping dates, residence country, residence state and postal code. This led to a reduction in the number of attributes to 15, which were then used by the models for the classification task.
The final idea of using only the four variables came from the fact that the neural network used with the smaller dataset was still accurate. Therefore, I decided to use this result and validate it with a larger part of the population to see if it still works or not.

Figure 4.8: Attribute Correlation with the output

4.6 Implementation

One of the biggest difficulties researchers have encountered in combating COVSARS-2 is the different consequences the virus can have on individuals. This is because the virus does not always have the same effects, rather it depends for example on age group or race[98]. In this work that has the aim of predicting the health status of patients affected by COVID-19, models are developed using the entire corpus of data selected as described previously. As the first step of this work, I have partially adopted the general workflow of [83].
This idea of using the individual characteristics of each patient to prevent health risks has been called personalised medicine in the last decade [79,81,99]. It is the result of numerous innovations in the field of molecular biology, genetics and bioinformatics. There is no really clear definition of Personalized Medicine except that it is a global approach to prevent, diagnose, cure and track diseases according to general characteristics and not just those of one person. This term is more European, while Americans prefer to call it as Precision Medicine.
Regardless of terminology, the basic idea is that each individual's genome, when interacting with the external environment due to complex pathologies, has unique characteristics that can be diagnosed and treated in an efficient and effective manner. There are several aspects of its application. These include oncology (i.e. with the analysis of the genes of an individual affected by a tumour, it is possible to choose a more appropriate therapy) or the diagnosis and cure of neurological and cardiovascular diseases. The interest in our case instead lies more in the use of individual information to avoid health consequences.

4.7 Experimental setup

This section briefly describes how the work was divided. More precisely, this was comprised two phases, which differ in the number of individuals analyzed and the programming language used.

4.7.1 Phase I: sample of 10000 individuals

The first phase of the experiments was conducted entirely in the Python environment using the Anaconda graphical user interface, specifically via Jupyter Notebook. Following what [83] did, a sample of 10,000 individuals was randomly selected from the CDC dataset to reduce the computational burden. Random sampling was done to obtain a balanced dataset in terms of the possible values taken by the target variables. In the end, there were 5000 dead and 5000 surviving individuals.

First, the improved RF model presented by Iwendi in the paper [29] is replicated. After achieving the same performance with the dataset he used in his paper [29], the model was applied to the previously created sample of 10,000 individuals. Then Iwendi's Boosted Random Forest model was compared with other ML models.
From the sklearn library, the GridSearchCV is used to perform an exhaustive search over specific parameter values. The most appropriate parameters have been shown in table 4.2. Different classifiers have been used: Logistic Regres-

Table 4.2: Optimal parameters for the Boosted Random Forest Classifier

max_depth	6
min_samples_leaf	2
min_samples_split	7
n_estimators	300

sion, Decision Tree Classifier, SVM Classifier,Gaussian Naive Bayes classifier. As a second part of this first experimental phase, a clustering technique was applied to the training set to obtain an additional variable called "clusters". This step was necessary to create a "tailor-made" predictive model called Boosted Clustered Random Forest. Then, a new attribute called "region" has been created and added to the dataset. This attribute represents the geographical location in relation to the registered residence. The "region" feature can assumes 4 values: (Midwest, West, South, Northeast). A better understanding about the States division for each region location is visible in the geomap (Image 4.9).

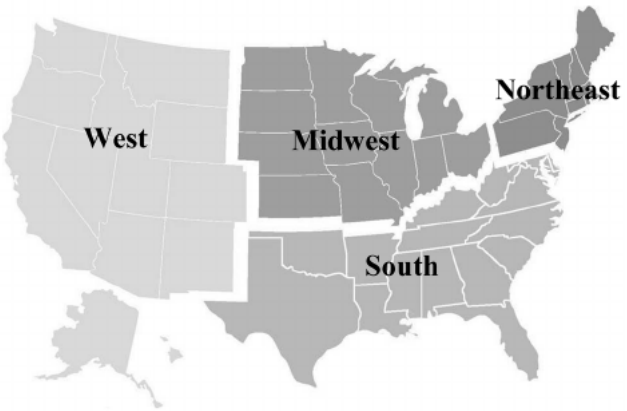


Figure 4.9: US territory division performed by adding the region attribute (Data source[6])

As last step, a neural network model is built to determine the interaction between the levels of different factors. Using the Tensorflow and Keras libraries, the neural network is constructed with 1 hidden layers. The activation function was RELU, ADAM the optimizer and the binary cross entropy function was the loss. For the output layer, a sigmoid activation function is used to output the death probability. The number of epochs used to train the model was 200.

4.7.2 Phase II: sample of 3 millions individuals

After getting the final results from the 10 thousand dataset, it was the turn of the larger dataset. This was because it was relevant to check whether or not the results obtained with the smaller dataset were confirmed with a larger number of samples.
For this part, it was decided to add the feature "region" here as well. The reason for this is that when looking at the feature of the previous dataset, it always comes first compared to all the other geographical variables already present in the dataset.
First of all, all available variables were considered for the classification task. In this part of the work, however, I not only tried to reuse all available attributes, but also to see if it was possible to remove some features. In this way, it was possible to better identify which variables were more relevant for distinguishing between survivors and deceased. More precisely, this kind of feature selection was done under several conditions by adopting different reasoning.

4.8 Performance Metrics

To ensure consistency with all other previous work, a wide range of performance measures were used. These include:
- Accuracy is the ratio of correctly predicted COVID-19 cases to the overall SARS-COV-2 cases. Accuracy is represented mathematically in the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

- AUC "Area Under the ROC Curve" in order to compare models among them.

- Precision which is the ratio of correctly predicted instances for the class of deceased patients to the total predicted dead individuals. High precision is equal to a low false positive rate [36]. This can be represented mathematically as:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

- Recall(or sensitivity) is the ratio of correctly deceased patients with respect to the total deceased patients inside the dataset. Mathematically expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

- F1-Score for giving more weight to the false positive predictions with respect to the false negatives. Mathematically, this means:

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4.4}$$

Chapter 5

Results

5.1 General Overview

Several machine learning algorithms were tested, as the most appropriate algorithm may vary depending on the data structure and task.
In the medical field, a machine that is very good at detecting true deaths (high sensitivity/recall) and a high probability that people identified as survivors actually survived (negative predictive value) is much more valuable than a machine that is instead very good at predicting survivors identified as dead, i.e. with a low false positive rate. In other words, it is better to have a model that has a higher false positive value than a low sensitivity and a low negative predictive value.
When analyzing reality, imagine that we are a person with COVID-19 who has just received the results that the model has predicted. If it is good news, this person is labeled as a survivor. What is matter is that we need to answer the following question: "What is the likelihood that he/she will actually survive?"
Another interesting point is about Feature selection. If successful, it allows irrelevant features to be discarded from the outset, which has a positive impact on computational costs (i.e. training time and less information to be gathered to run tests).

5.2 10 thousand CDC dataset

The results obtained in the test set by comparing the Iwendi Boosted Random Forest [29] for the dataset with 10000 individuals are summarized in Table 5.1. More precisely, Figures from 5.1 to 5.4 represent the performances in terms of accuracy, F1-Score, sensitivity and precision achieved by each algorithm tested.
With regard to what Cornelius did in his work, I took a step back. I decided not to assume that the performance obtained with the Boosted Random Forest

could be the best and therefore to repeat the whole case study implemented by Iwendi. In his paper, the performances of the Boosted Random Forest classifier were compared with those of the other classifiers. The algorithms compared were: Decision Tree Classifier, Support Vector Classifier, Naive Bayes and Boosted Random Forest Classifier.
For my study case the recall metric given by the SVM was the highest. But at the same time, this classifier had shown the lowest performance in terms of precision score. This naturally led to the classifier being ranked last in terms of F1 score. Although the model is positioned as last, it can still be considered a good classifier. This is because it has shown a high sensitivity score. It should be remembered that a high recall score is necessary for predictions that need to be outputs-sensitive. Indeed, in situations such as predicting a cancer or a disease such as COVID -19, a high recall score is necessary because detection of false negatives is mandatory. It is fine if a survivor is marked as dead, but a person who dies should never be classified as a survivor. As indicated in the paper [74], the SVM that performed best on this type of data was the one with linear kernel and regularization parameter of 1.
Then following the ranking list from bottom to top, there are: Naive Bayes, Decision Tree and Boosted RF. The results of these latters are very similar to each other. In particular, for the Naive Bayes classifier, the Precision is higher than for the Decision Tree, but as mentioned earlier, high recall is preferred over a high precision for the aim of this work. The armonic mean (i.e. F1 score), anyway, puts the Decision Tree classifier as the winner among the two. Finally, the Boosted Random Forest outperforms all the others in terms of both recall and precision, making it the best of the lot. Even if the F1-score and Accuracy shown by it and decision tree are quite similar.

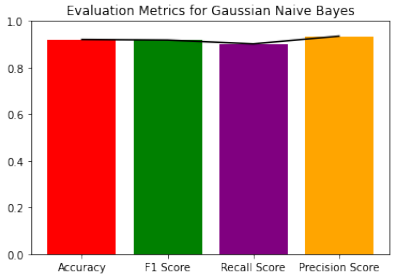


Figure 5.1: Evaluation metrics for GNB on 10 thousand individuals

After confirming that the Boosted Random Forest classifier was the best, I focused my attention on this classifier. This was done by comparing the "simple" Boosted Random Forest with two other types of Boosted Random Forest. This

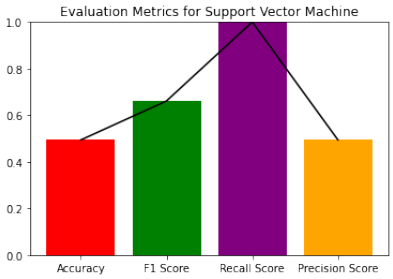


Figure 5.2: Evaluation metrics for SVM on 10 thousand individuals

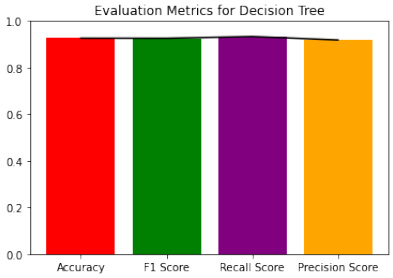


Figure 5.3: Evaluation metrics for DT on 10 thousand individuals

"new" first model was constructed according to the concept of clustering given by Cornelius and the other model is related to the concept of creating a Boosted Random Forest by using Features Elimination. This one was created in order to try to reduce the problem complexity. At the end this model was constructed by discarding 14 features. The omitted features were selected as those considered irrelevant for differentiating among survivors and dead people by the Random Forest algorithm.
Table 5.2 put in comparison the results obtained by the "simple" Boosted Random Forest, the Clustered Boosted Random Forest and the Boosted Random Forest with Feature Elimination.
Looking at the results the Boosted Random Forest classifier with Feature elimination led to the highest performance.
The last step of this experimental phase was the construction of the Neural Network as Cornelius did 83. The NN is trained for 200 epochs by achieving an accuracy of 86%. The mortality probabilities for White males and females given as output by the Neural Network are shown in Tables 5.3 and 5.4.

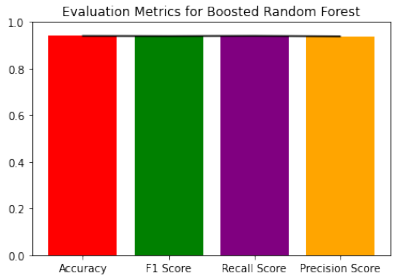


Figure 5.4: Evaluation metrics for Boosted Random Forest on 10 thousand individuals

Table 5.1: Classifiers performance on 10 thousand CDC dataset

Classifier	**Recall(%)**	**Precision(%)**	**F1-score(%)**	**Accuracy(%)**
Gaussian Naive Bayes	90.18	93.49	91.80	92.05
Support Vector Machine	100	49.40	66.13	49.40
Decision Tree	92.61	92.05	92.33	92.40
Boosted Random Forest	92.81	93.10	92.95	93.05

5.3 3 millions CDC dataset

In this part of work, even more classifiers were analyzed than before. These include: Neural Network, Stepwise logistic regression, Forward logistic regression, Decision Tree classifier, Random Forest, Gradient Boosting classifier, Gaussian Naive Bayes classifier and the Ensemble classifier. The latter one is constructed as a function of posterior probabilities (for the target classes) given by all the aforementioned classifiers.
Moreover in this part of the project activity different amount of features were tested. The 3 million CDC dataset is divided into 10 samples of about 300,000 individuals each.
Four different typologies of experiments were conducted for a total of 40 experiments. The overall results are resumed in the tables 5.5, 5.6, 5.7, 5.8. The results inside the tables are given as mean ($\bar{x}$) $\pm$ the standard deviation (σ).
As for the authors of [89] RF outperformed all the other algorithms while the last position was taken by the Logistic Regression algorithm. For all the different type of experiments conducted, the champion model was always the Random Forest classifier. Pictures from 5.5-5.8 represent the number of trees selected for the Random Forest for each of the four experiment conducted. These graphs show how the average squared error changes as the number of trees in the forest increases. The part in which it must be put higher attention is the validation one in which

Table 5.2: Comparison Boosted RF with and without Clustering

Classifier	**Recall(%)**	**Precision(%)**	**F1-score(%)**	**Accuracy(%)**
Boosted Random Forest	92.81	93.10	92.95	93.05
Clustered Boosted RF	94.03	93.84	93.93	94.00
Boosted RF with Feat. Elim.	94.43	94.24	94.34	94.40

Table 5.3: Mortality risk for a healthy White male (Neural Network output)

Age	**10-19(%)**	**20-29(%)**	**30-39(%)**	**40-49(%)**	**50-59(%)**	**60-69(%)**	**70-79(%)**	**80+(%)**
Northeast	0.13	0.30	0.82	2.02	4.91	11.43	24.38	50.64
Midwest	0.18	0.28	0.57	1.39	3.40	8.07	18	42.24
South	0.11	0.38	0.81	2.72	7.05	15.94	32.14	59.02
West	0.09	0.42	0.99	2.37	7.66	21.79	41.03	66.90

the error gives an indication about how well the model is able to generalize. Concerning the training error usually it decreases as the number of trees increases. When all features were used, the number of trees where the minimum error occurred in the validation set was equal to 29 (Figure 5.5). In the case in which features have been selected according to RF relative importance, there were 4 trees (Figure 5.6). While concerning the elimination of the binary attributes that are negatively correlated with the target variables the minimum error occurred at 21 trees (Figure 5.7). The last case under analysis was the one in which only 4 features have been used the number of trees was equal to 87 (Figure 5.8).
The method of class target voting used either in Python or with SAS was the soft voting. Specifically, this means that the predicted class probabilities for a new individual is calculated as the mean of the predicted class probabilities of the trees in the forest. This method was chosen w.r.t. hard voting because the performance achieved is higher. The reason is that this method gives more weight to highly confident votes.

Table 5.4: Mortality risk for a healthy White female (Neural Network output)

Age	**10-19(%)**	**20-29(%)**	**30-39(%)**	**40-49(%)**	**50-59(%)**	**60-69(%)**	**70-79(%)**	**80+(%)**
Northeast	0.21	0.28	0.50	1.06	2.62	6.31	15.91	40.37
Midwest	0.24	0.26	0.40	0.84	1.80	4.38	11.88	32.54
South	0.21	0.27	0.69	1.56	3.81	9.00	20.99	48.72
West	0.18	0.25	0.71	1.92	4.39	9.72	20.16	42.95

Table 5.5: Final results on 3 millions dataset using all the variables

Classifier	**Accuracy(%)**	**F1-score(%)**	**AUC**
Random Forest	96.81 ±0.33	96.73 ±0.35	0.996 ±0.003
Decision Tree	92.97 ±0.23	92.49 ±0.26	0.961 ±0.001
Bayesian Network	92.89±1.34	92.71±1.38	0.968±0.013
Logistic Regression	91.89±0.36	91.92±0.38	0.969±0.002
Gradient Boosting	93.57±1.60	93.89±2.03	0.976±0.020
Ensemble	93.63±0.32	94.21 ±1.01	0.981±0.014

Table 5.6: Final results on 3 millions dataset by using only the features according to RF relative importance

Classifier	**Accuracy(%)**	**F1-score(%)**	**AUC**
Random Forest	97.20±0.34	97.18±0.30	0.997±0.0004
Decision Tree	92.82 ±0.05	92.49 ±0.11	0.956 ±0.003
Bayesian Network	92.83 ±0.31	92.94 ±0.16	0.970 ±0.005
Logistic Regression	91.63 ±0.19	91.65 ±0.08	0.969 ±0.002
Gradient Boosting	93.37 ±0.51	94.23 ±0.20	0.979 ±0.003
Ensemble	93.62 ±0.08	94.32±0.39	0.983 ±0.002

Table 5.7: Final results on 3 millions dataset by dropping binary features negative correlated with the target variable

Classifier	**Accuracy(%)**	**F1-score(%)**	**AUC**
Random Forest	96.65 ±0.36	96.63 ±0.35	0.996 ±0.001
Decision Tree	92.56 ±0.08	93.26 ±0.06	0.961 ±0.001
Bayesian Network	92.30 ±0.29	92.37 ±0.27	0.967 ±0.004
Logistic Regression	91.80 ±0.18	91.72 ±0.22	0.967 ±0.003
Gradient Boosting	93.03 ±0.41	93.39 ±0.39	0.977 ±0.004
Ensemble	93.13 ±0.12	93.27 ±0.11	0.981 ±0.005

Table 5.8: Final results on 3 millions dataset by using 4 features

Classifier	**Accuracy(%)**	**F1-score(%)**	**AUC**
Random Forest	85.79 ±0.81	85.73 ±0.83	0.936 ±0.003
Decision Tree	85.70 ±1.20	85.63 ±1.22	0.932 ±0.005
Neural Network	85.34±1.26	85.21±1.25	0.924 ±0.004
Bayesian Network	85.69 ±1.82	85.64 ±1.78	0.931 ±0.008
Logistic Regression	85.06 ±2.12	85.07 ±2.15	0.927 ±0.013
Gradient Boosting	85.16 ±1.08	84.89 ±1.14	0.929 ±0.005
Ensemble	85.48 ±2.20	85.18 ±2.26	0.930 ±0.007

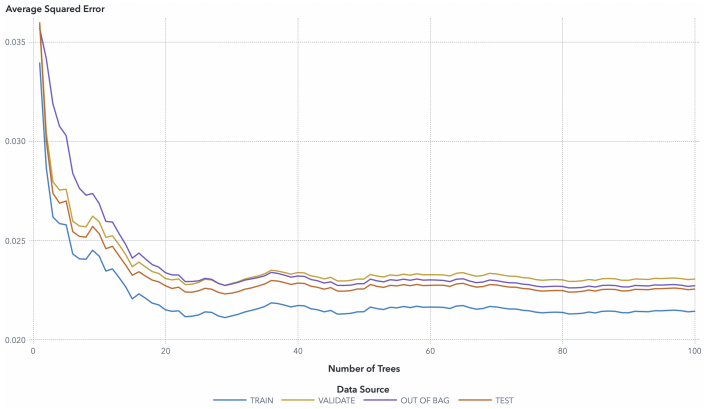


Figure 5.5: Representation of the average squared error according to the number of trees selected for the Champion model using all the attributes

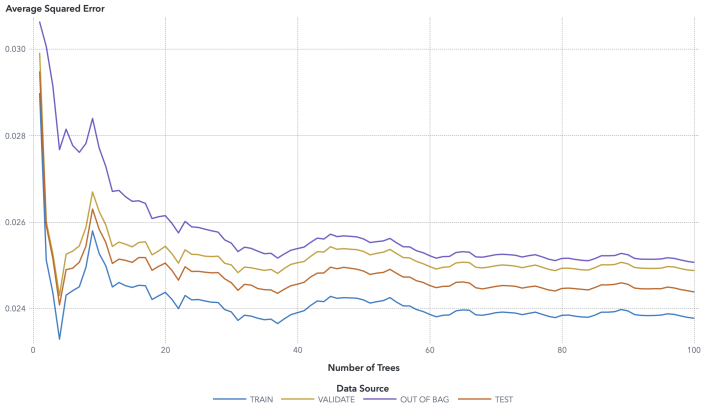


Figure 5.6: Representation of the average squared error according to the number of trees selected for the Champion model following RF relative importance

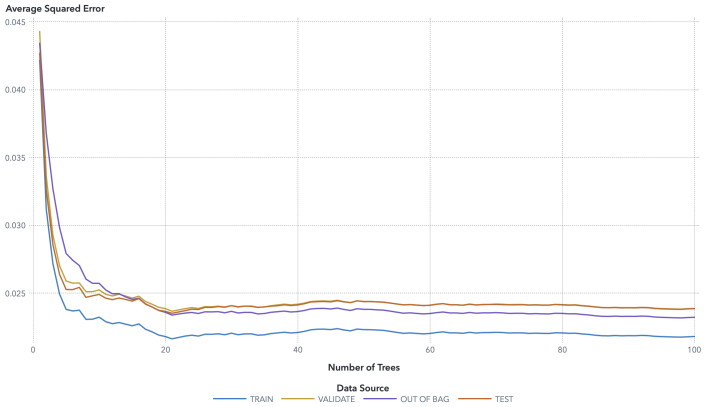


Figure 5.7: Representation of the average squared error according to the number of trees selected for the Champion model for dropping binary features negatively correlated within the target

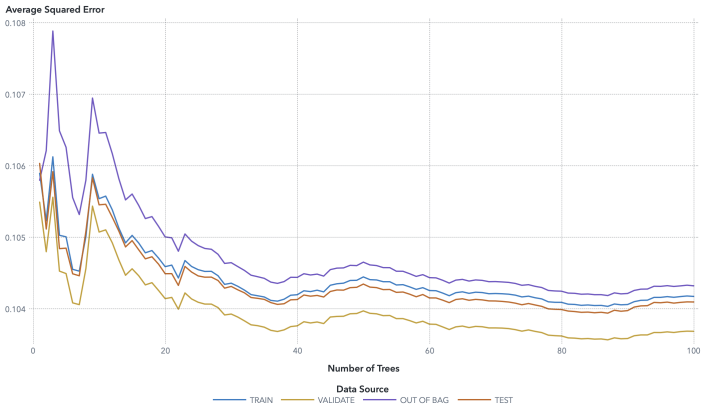


Figure 5.8: Representation of the average squared error according to the number of tree selected for the Champion model in 4 features case

Chapter 6

Analysis and Discussion

6.1 10 thousand CDC dataset

The proportion of men in the sample used was 51% and 49% of women. Temperature was available for 73.3% of the data, with 24.5% of patients presenting with a temperature > 38.0°C.
The distribution of the extracted data sample across the four different geographical areas is as follows 6.1:

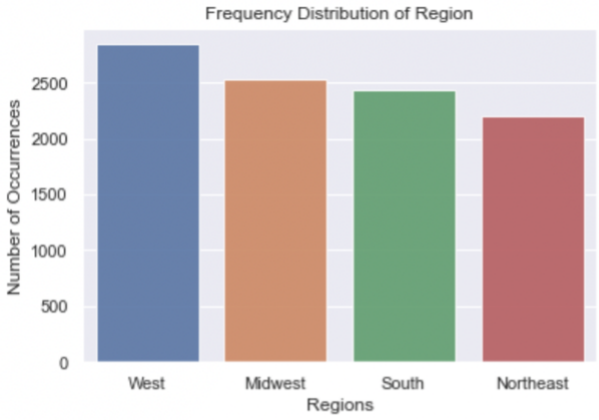


Figure 6.1: Frequency distribution of individuals

Figure 6.2 simultaneously shows the results for the classifiers. The boosted RF model of Iwendi [29] generally proved to be the best performing in terms of the four metrics used and thus a reasonable standard for addressing the predictive performance for COVID-19 patient mortality the results found. The first experiment did not involve any clustering of the data, but simply applied different algorithms

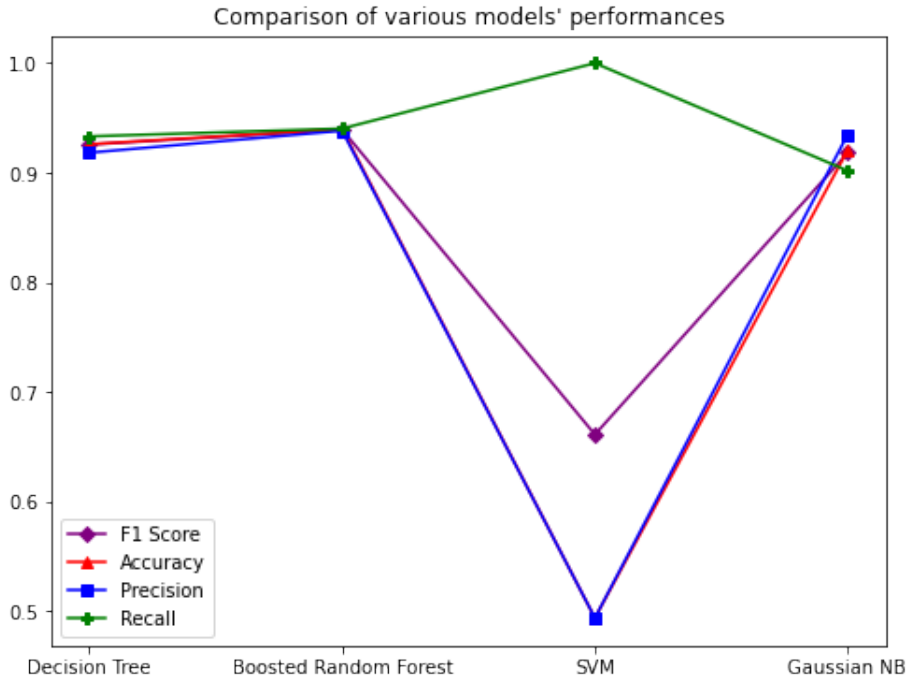


Figure 6.2: Final results in the 10 thousand CDC dataset

to the data after some pre-processing.
Then we moved on to clustering. To build such a model, we first need to define a suitable measure that can capture the degree of similarity between one person and another. Previous work has used a few different approaches to measure patient similarity. For example, in [100], a so-called case-based statistical approach was used to identify similar patients. Panahiazar et al in [101] used a clustering approach to identify similar patients. A RF method can also be used to define a similarity measure, which Lee [51] used to develop a similarity measure for intensive care patients. This is possible by simply running a RF model unsupervised (without a prediction target) that provides a similarity score between individuals. More specifically to find the appropriate number of clusters for grouping the data, both the average silhouette and the elbow methods were compared to ensure that the optimal number of clusters was selected for K-means algorithm.
The images 6.3 and 6.4 show the results obtained. By using these two approaches and trying out the two selected k (k = 9 for the average silhouette score and k = 5 for the elbow method), k = 5 was found to be the best k in the end.
The distribution within the clusters was indicated in 6.1.

After applying KMeans for k = 5, a new feature called "clusters" was added to the dataset, which was used within the Boosted Random Forest, now called

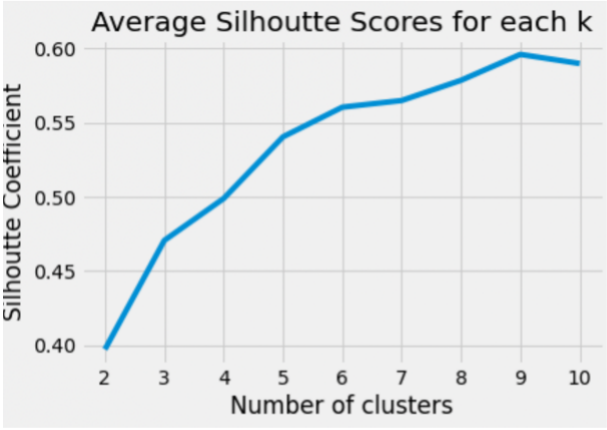


Figure 6.3: Average silhouette applied for choosing the appropriate k for K-means algorithm

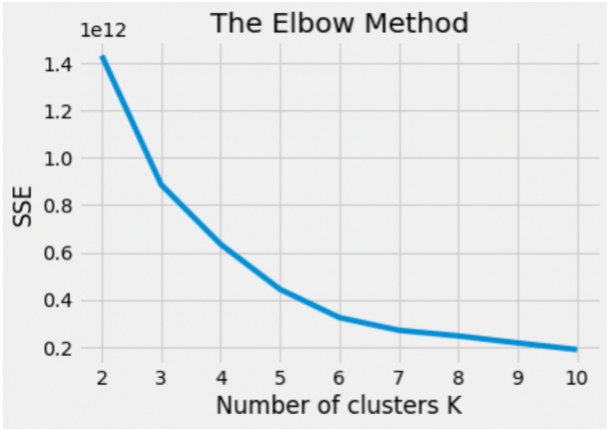


Figure 6.4: Elbow method applied for choosing the appropriate k for K-means algorithm

Clustered Boosted Random Forest.
Figure 6.5 shows the features in order of relevance to the Boosted Random Forest, from the most important to the less important.
The additional attribute "clusters" turns out to be quite relevant and therefore useful for distinguishing between deceased subjects and not for the Random Forest classifier.
As can be seen in the 5.2 table, performance improves when features are selected according to the relative importance assigned to the attributes on RF. By using only the most relevant features and discarding all the attributes after pna_yn (see Figure 6.5), accuracy increases and in general the overall performance also

Table 6.1: Main characteristics of the clusters found in the 10 thousand CDC dataset

Cluster	Number of people	Pre-existing conditions	Hospitalized	Fatality Rate
1	3198	36.68%	41.15%	50.72%
2	1067	55.48%	35.98%	45.07%
3	2513	45.28%	41.03%	49.18%
4	1601	39.30%	33.66%	39.16%
5	1621	45.21%	42.68%	63.78%
Overall	10000	42.72%	39.62%	50.00%

improves.
Hence, the K-means algorithm applied on top of Boosted Random Forest is proved to be useful for this first part of work.

Finally, it was the turn of "Boosted Random Forest with Feature Elimination". It works by dropping the features that the Boosted Random Forest classified as irrelevant (according to the importance of the features represented in 6.5). This was done because irrelevant or partially relevant features can negatively affect the performance of the model. Irrelevant attributes can reduce the accuracy of the model by causing the model to learn noise due to irrelevant features.
The best cut was achieved by considering only the features up to "chills_yn" and eliminating the others below. However, it is important to remember that 14 features were eliminated and the result is therefore significant from a computational cost point of view.
As last step to obtain the interaction between the levels of different factors, a neural network model was built on the data. The neural network model is used to obtain the probability that an individual with certain specified characteristics will die due to COVID-19. The NN was built by inferring new characteristics (these characteristics are thought of as a hidden layer) that are linear combinations of the original features that were then passed through an activation function. Thus the dependent variable is modelled as a function of linear combinations of the derived features.
Only the medical condition, sex and the age group to which each patient belongs were used to build such network.One neural network for each race was created, as there was the desire to obtain a probability for each race.
Using the Tensorflow and Keras libraries, a neural network (Picture 6.6) was constructed with 1 hidden layer and RELU as the activation function, Adam the optimizer and the binary cross entropy function as loss. For the output layer, a sigmoid activation function is used to output the probability of death for each individual. The number of epochs used to train the model was 200.
From this bar chart (Figure 6.7) is evident that almost all people considered are

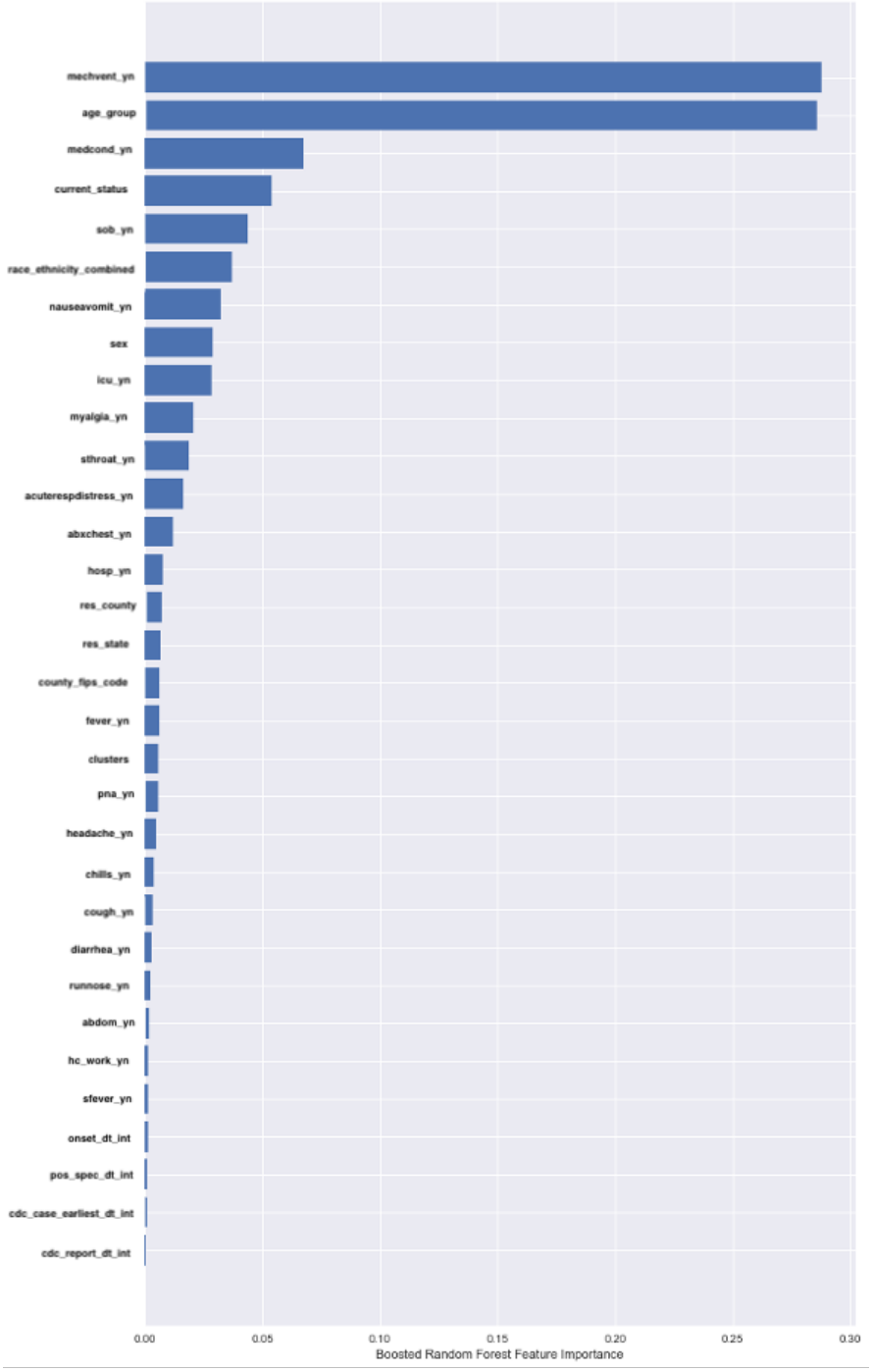


Figure 6.5: Relative variable importance for Boosted Random Forest

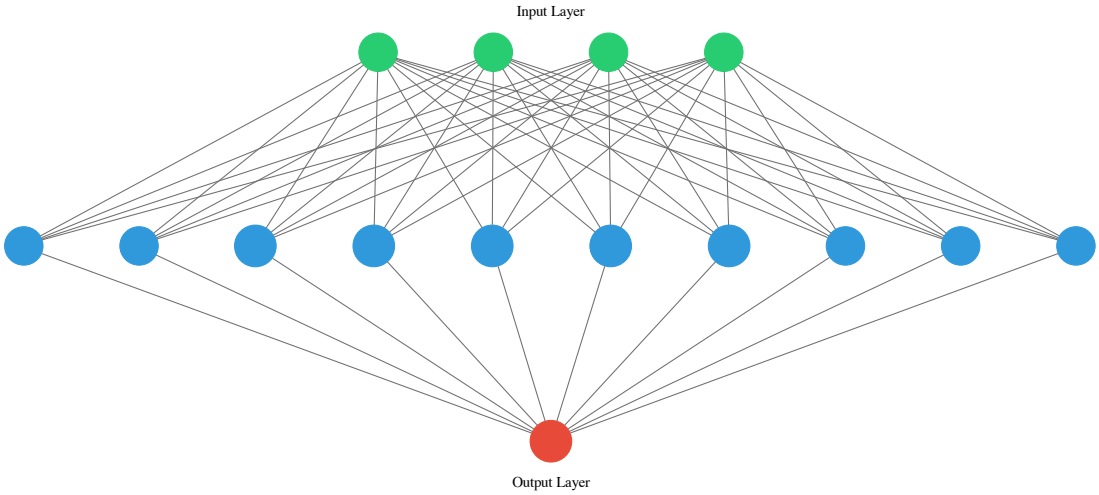


Figure 6.6: Representation of the Neural Network used for last step of the 10 thousand CDC dataset

white. However, this sample is representative of the dataset from which the 10,000 people were extracted. In fact, almost all the individuals registered are white in this case as well. Anyway, it was decided to show only the neural network for white men and women, as the number of elements in the other races was very small and it was not possible to train the network sufficiently. Another peculiarity was the fact that only women and men who had no previous diseases, i.e. healthy people, were considered for this analysis. This choice was made to exclude the possibility that their death was related to other factors rather than SARS-COV-2.
As can be seen in Figure 6.8, the probability of death increases with age for both men and women. In particular, women are generally less likely to die than men, and a few studies show that the general trend is actually that men are sometimes twice as likely to die as women [102].
In the case of the 10,000 randomly selected individuals, women from South states are slightly more likely to die than others. Among men, on the other hand, those with a higher fatality rate are those from West.

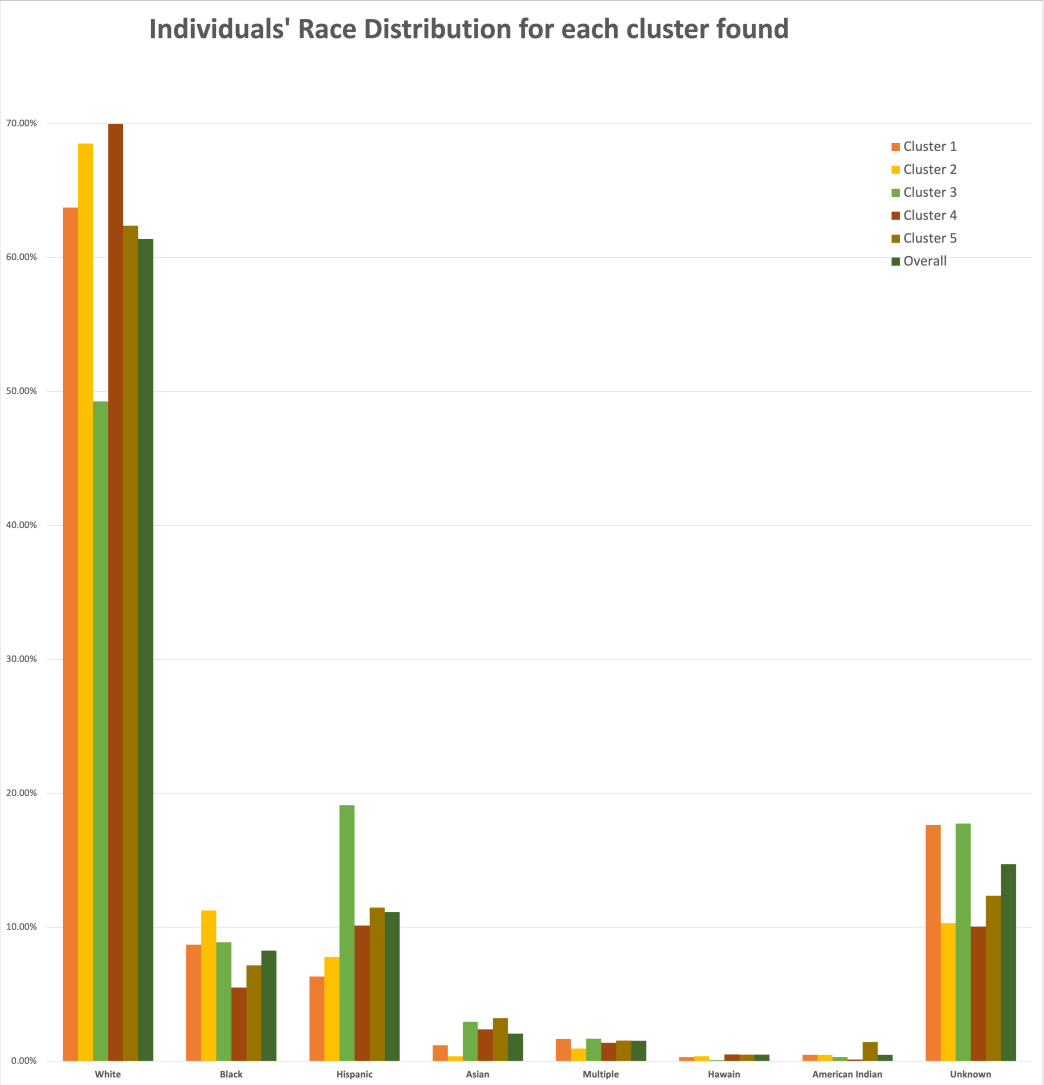
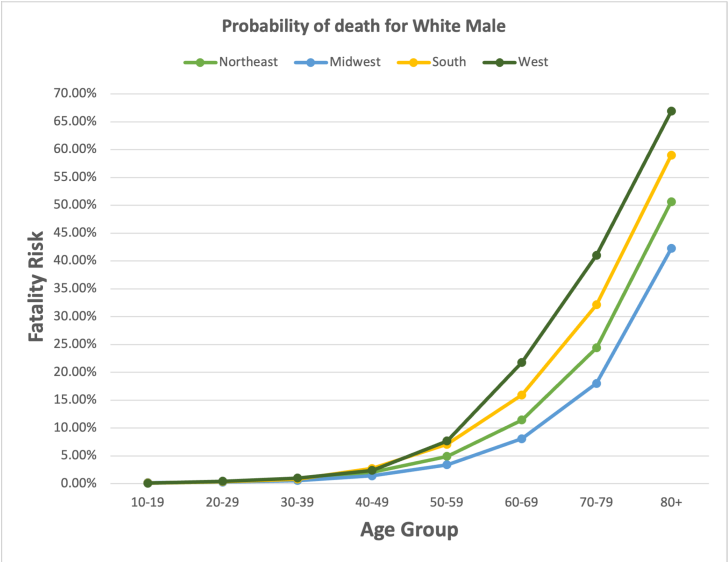


Figure 6.7: Individuals' Race Distribution for each cluster found in 10 thousand CDC dataset

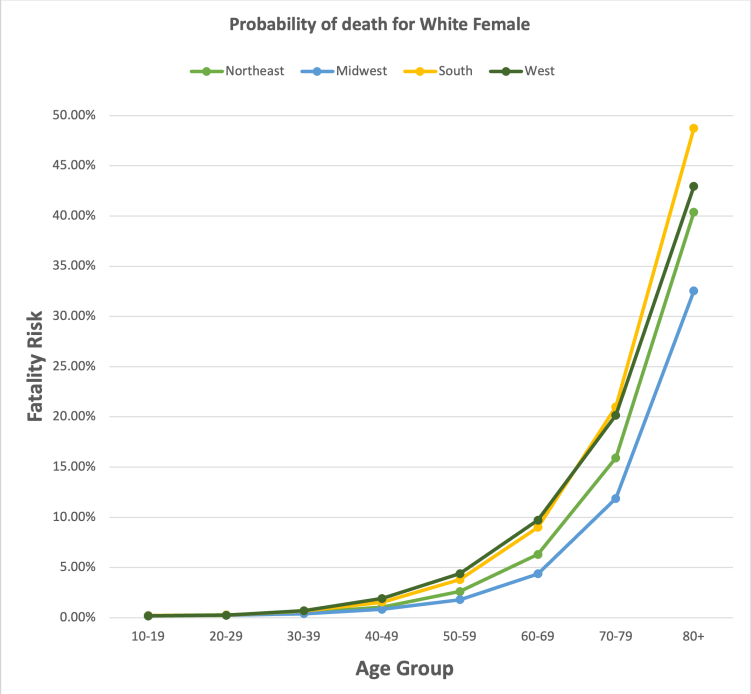
6.2 3 millions CDC dataset

Tables 6.2 and 6.3 represent each characteristic in relation to the subjects under analysis in this part of the research activity.

As can be seen, the age factor has always been one of the characteristics that contribute most to whether a person affected by COVID dies or not (Figure 6.9). Indeed, a number of studies looking at the prediction of severity or mortality have found that age is one of the most important characteristics contributing to the



(a) Probability of death for White male



(b) Probability of death for White female

Figure 6.8: Fatality risk for White people with no pre-existing condition

prediction of severity of cases [103,104,105]. As the sample data suggest, the older the patient, the greater the probability of death, while the lower the age, the

Table 6.2: 3 million CDC dataset symptoms

Feature Name	Survivors (n=1677538)	Dead (n=1677538)
Abdominal Pain (No)	1457856	1509905
Abdominal Pain (Yes)	219682	167633
Abx Chest (No)	1604485	830365
Abx Chest (Yes)	73053	847173
Acute Resp. Syndrome (No)	1659164	1112145
Acute Resp. Syndrome (Yes)	18374	565393
Chills (No)	951510	1124676
Chills (Yes)	726028	552862
Cough (No)	592067	518251
Cough (Yes)	1085471	1159287
Diarrhea (No)	1231269	1278802
Diarrhea (Yes)	446269	398736
Fever (No)	1027895	798252
Fever (Yes)	649643	879286
Headache (No)	638055	1189280
Headache (Yes)	1039483	488258
Hospitalized (No)	1559095	379869
Hospitalized (Yes)	118443	1297669
ICU (No)	1656094	978266
ICU (Yes)	21444	699272
Mechanical Vent. (No)	1671740	1207654
Mechanical Vent. (Yes)	5798	469884
Previous Med. Cond. (No)	987029	148531
Previous Med. Cond. (Yes)	690509	1529007
Myalgia (No)	736406	1046980
Myalgia (Yes)	941132	630558
Nausea Vomit (No)	1283161	1322121
Nausea Vomit (Yes)	394377	355417
PNA (No)	1602506	794349
PNA (Yes)	75032	883189
Run nose (No)	914250	1317153
Run nose (Yes)	763288	360385
SFever (No)	949995	971460
SFever (Yes)	727543	706078
Dyspnea (No)	1231178	673937
Dyspnea (Yes)	446360	1003601
Sore Throat (No)	1013976	1373869
Sore Throat (Yes)	663562	303669

greater the probability of survival.
Furthermore, it can also be observed that the number of men in the sample who survived is lower than that of women. From this it can be deduced that men are more likely to die in relation to women (Figure 6.10), and indeed there is a study

Table 6.3: Additional information on 3 millions sample

Feature Name	Survivors (n=1677538)	Dead (n=1677538)
Age (80+ years)	30681	574321
Age (70-79 years)	74804	473688
Age (60-69 years)	163473	301141
Age (50-59 years)	238287	198584
Age (40-49 years)	258471	82345
Age (30-39 years)	289198	31792
Age (20-29 years)	333028	10464
Age (10-19 years)	213603	2921
Age (0-9 years)	75993	2282
Female	932265	706372
Male	745273	971166
Health care worker (No)	1490927	1632528
Health care worker (Yes)	186611	45010
Race (White)	1136221	1240101
Race (Hispanic/Latino)	292674	157492
Race (Black)	159812	180997
Race (Asian)	33301	63883
Race (Native Hawaiian)	10368	13857
Race (American Indian)	4103	5919
Race (Multiple)	41059	15289

on this [106] which suggests that men are about twice as likely to die as a woman. Once again from the histogram 6.11 is evident that almost all people considered are white.

For the abdominal X-ray chest, one can see how the recording of such characteristics can be of fundamental importance, as people who have undergone it are most likely to die. This is easily explained because, especially in the crowded hospitals in the first year and a half, these kinds of X-rays were only taken when the patient's condition was very critical. Having cough, fever, dyspnea [107] and pneumonia [108] seem to be the most common symptoms noticed in dead patients 6.12. As mentioned in [109] the acute respiratory distress is a predictable serious complication of COVID-19 that needs to be recognised early and treated comprehensively.
Furthermore, it is noted that out of the total 231621 health workers, 45010 died while the remaining 186611 survived. Although they were exposed to a higher risk than those who do not work in this sector (one study states that the risk of death is about 7 times higher [110] compared to a person who does not work in this sector), most of them survived thanks to the timely intervention of the States, which were the first to subject this category of workers to compulsory vaccination, which of course offers the possibility that the worst did not happen. Another

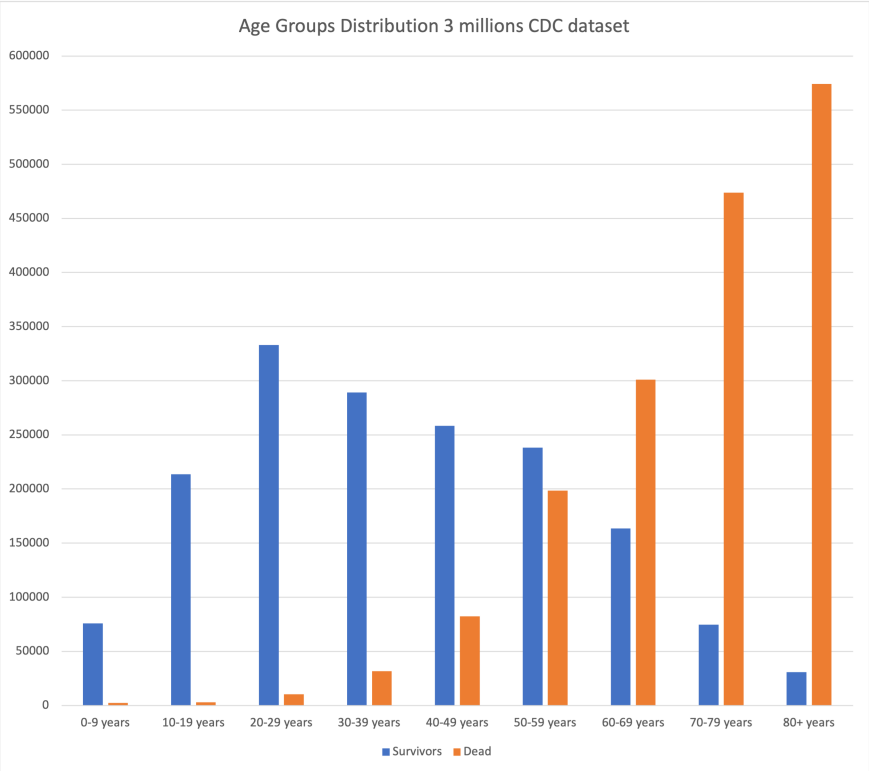


Figure 6.9: 3 millions CDC dataset age distribution

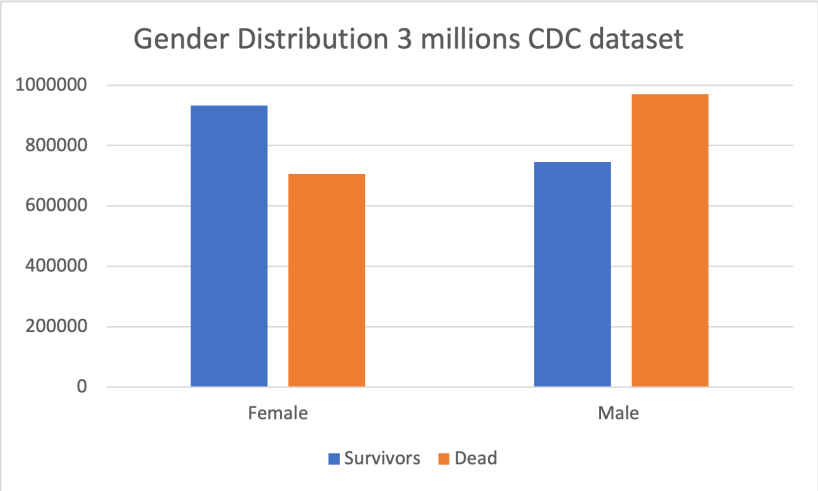


Figure 6.10: 3 millions CDC dataset sex distribution

crucial circumstance that increases the probability of death is the registration of the event in terms of whether or not the patient was hospitalised. Indeed, it is

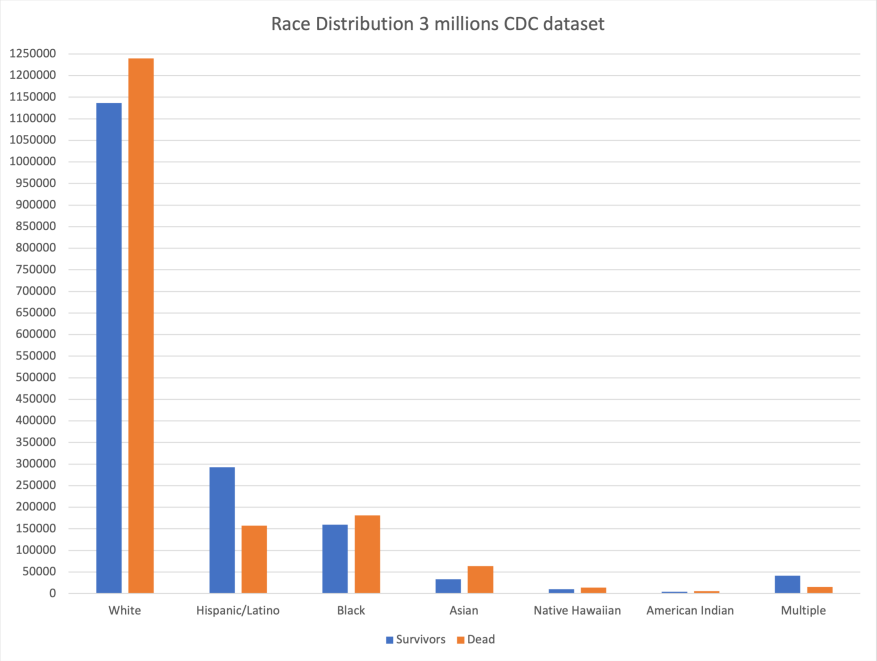


Figure 6.11: 3 millions CDC dataset race distribution

clear that most of those who have been admitted to hospital have died, and this is always due to the fact that at the moment of overcrowding only really sick people had a bed in the hospital. The same holds for those who were admitted to the intensive care unit and / or they were mechanically ventilated.
For those who have pneumonia and/or have pre-existing conditions, these are also essential as they increase the likelihood of getting COVID and dying. Usually, older people and those with chronic pre-existing conditions are at higher risk of severe COVID-19 leading to hospitalization, admission to intensive care and death [111].
Data pre-processing was performed in the Jupyter Notebook system using Python as programming language. After the 3 million individuals were randomly selected, they were split into 10 files, as the size supported by the SAS Viya for student did not allow the examination of all individuals at once. After this step, the data was loaded into the SAS environment, it was possible to proceed to the creation of the pipeline to be used, which is shown in Figure 6.13.
In comparison to the classifiers previously used, additional approaches were tested to those already applied in the 10 thousand CDC dataset: Neural Network, Stepwise logistic regression, Forward logistic regression, Decision tree, Random forest, Gradient Boosting, Ensemble classifier: this is constructed as a function of the posterior probabilities (for the target classes) given by all the above classifiers,

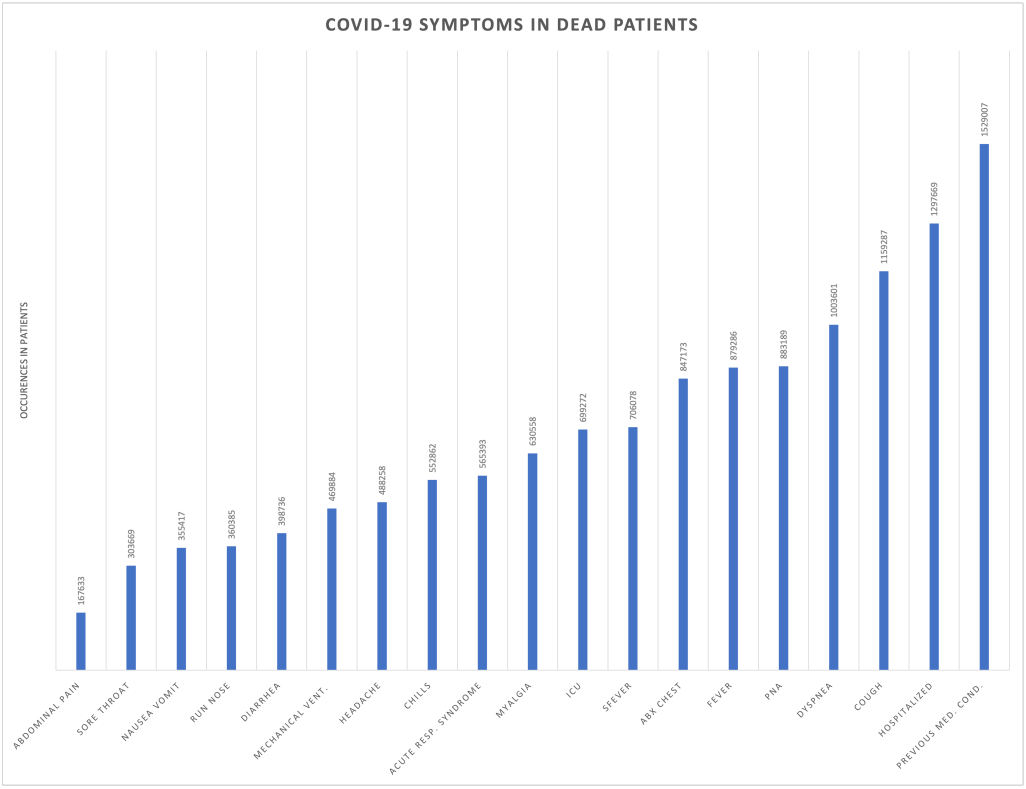


Figure 6.12: 3 millions CDC dataset symptoms in dead patients

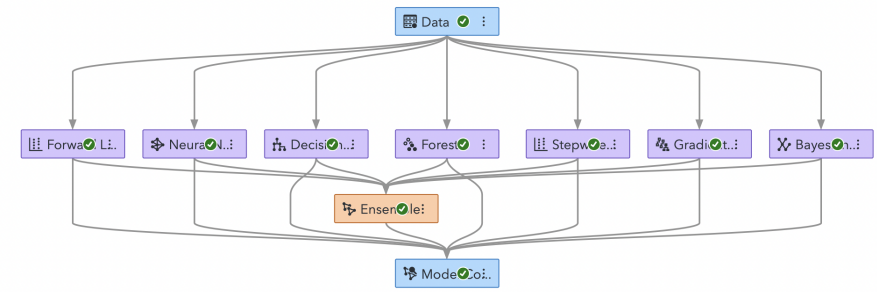


Figure 6.13: Pipeline implemented on SAS Viya

Gaussian Naive Bayes. The difference between backward stepwise logistic regression and forward regression is as follows:
- Backward Stepwise: is a method that starts with a (usually complete) set of variables and then in turn excludes variables from that set until a stopping criterion is met.
- Forward: starts with an empty set of variables and then gradually adds each new

variable and tests for statistical significance.
The original creation of the pipeline was slightly different, as it tried to use 2 blocks belonging to the pre-processing processes: one was that of imputing missing values and another was tested only when analyzing the use case of all the available features, which was used for the "feature selection".
This last block was included because, once its process was completed, it could have led to a computational time that would have been reduced in terms of training time within the classifiers. Also, because sometimes we may filter out information that is simply noise for the models we are using, allowing the algorithm to achieve the same performance it previously obtained by keeping all the information. Sometimes, however, this is explained simply because the information is useless for deciding the target class. However, it was found that its addition to the pipeline led to a large decrement in terms of performance and for this reason it was eliminated this latter block.
In the case of the "imputation of missing values" block, it was found that its elimination improved the overall performance of the classifiers improved. This can be explained by the fact that in some specific cases even missing values per se have a certain meaning and the removal of this knowledge for the models can lead to a drop in performance. Although the common practise is initial or permanent removal of them, removing or imputing missing values before understanding the reason behind them can lead to negative consequences. A particular example is the optional compilation of sensitive data such as race, which individuals were not required to provide if they did not want to. So, after this analysis, it was decided to create the pipeline as it was presented before.
It is possible to divide the experiments conducted into 4 types:

1. First, I started experiments in which all the features were taken into account.

2. Then I tried to consider the possibility of using the relative importance given by the random forest selection as feature selection and launched more experiments. Within the system, the attributes decreased.

3. After another part of the experiments, it was considered to eliminate the binary attributes that are negatively correlated with the target attribute and disregard the dates as they are not useful for the analysis performed. In addition, the specific attributes of the patient's origin were not considered, but only the "region" attribute as geographical data, where there is an indicative geographical alignment of the state of affiliation in relation to its geographical position inside the American continent.

4. Lastly, the models were created using only the 4 variables that had given excellent performance to the ANN used in the 10 thousand CDC dataset.

Based on the results, it can be noticed that by discarding the features with less than 0.012 as relative importance following the relative importance attributes of the Random Forest, the performance remained unchanged compared to using the whole dataset.
However, this has one major advantage: the number of attributes used decreases and so does the complexity of the problem. The removed attributes (see Table 6.4 for more information) are: nauseavomit_yn, sfever_yn, cough_yn, abdominal_yn, currentstatus.

At the same time, however, it is important to emphasise that performance remains excellent even with the idea of do not considering the binary features with negative correlation within the target variable. In this case, another advantage is that the eliminated variables are 14, i.e. only about half were used by the classifiers compared to the original variables. This leads to a further reduction in the complexity of the problem and, at the same time, to reflect on the considerable

Table 6.4: RF relative variable importance on 3 millions CDC dataset

Variable Name	**Relative Importance**
age_group	1
hosp_yn	0.92
medcond_yn	0.31
pna_yn	0.25
icu_yn	0.18
abxchest_yn	0.14
acuterespdistress_yn	0.10
mechvent_yn	0.09
headache_yn	0.07
race_ethnicity_combined	0.045
region	0.044
runnose_yn	0.031
hc_work_yn	0.025
myalgia_yn	0.024
sob_yn	0.023
sex	0.022
diarrhea_yn	0.014
sthroat_yn	0.014
chills_yn	0.013
fever_yn	0.012
nauseavomit_yn	0.011
cough_yn	0.010
sfever_yn	0.010
abdom_yn	0.009
current_status	0.008

amount of information recorded in hospitals that seems to be superfluous for determining a person's mortality.
Then, based on the test performed on the small dataset where the neural network performed well with only 4 attributes to predict patient death , I also tried to test if this is further confirmed in this larger dataset. The number of hidden layers was equal to 1 and the tanh was used as the activation function for the latter. Even though Relu is considered the best and most advanced activation function, able to completely eliminate drawbacks such as the vanishing gradient, the tanh performed better in this case. Increasing the number of hidden layers degrades the performance of the network. This can be explained by the fact that the complexity of the problem is not very high. Moreover, increasing the number of hidden layers beyond the number of layers required to solve the problem caused the accuracy in the test set to decrease, but the performance in the training set to increase. The network has therefore overfitted, resulting in the network being unable to generalize to unseen data. The performance of the NN remains the same obtained with the smaller dataset. Thus, the result is still very interesting. If this result will be confirmed again in the future with other dataset then it will be enough to know only the patient's age, health status, race and sex in order to predict with a high degree of confidence whether the patient will die or not.
An attempt was also made to implement the clustered boosted random forest approach used in the first phase of the work in all the analysis cases of this second phase of experiments to check whether this actually leads to improvements. In this case, no improvement was found: the performances remained the same as those obtained without clustering done on top of the algorithm. Since each additional operation performed within the overall workflow increases its computational complexity (in terms of cost and time), it was decided to discard the cluster attribute, which was therefore not considered relevant to the analysis of the problem.

Chapter 7

Conclusion and Future Work

7.1 Concluding thoughts on the research activity

All of the studies conducted to date, as can be seen in the Related Work chapter, have used a much smaller number of samples to validate their results than those used in my analysis, which tested the models on 3 million patients.
Almost all previous studies reported that age is a strong prognostic factor, and this was also confirmed by the results obtained in this research activity, using either the small or the largest data set. Although the results of 61, 64 were similar to those of mine in terms of performance achieved, they had an advantage which is quite related i.e. much more interesting details related to biological elements and pathological conditions in the patients. This of course helped them in achieving high performance. Other than that, the analysis that I carried out, I think it is very useful and interesting from different points of view. First of all, this concerning the features. The use of a smaller number of features was done step by step to see how the elimination of these attributes might or not affect the quality of the model. Moreover this was performed in different ways and it was found that the 33 characteristics are necessary in order to obtain good performance. Also by reducing the number of characteristics from 33 to 26 and then to 15, no significant decrease in terms of performance was found. Using the technique of eliminating binary attributes negatively correlated with the target attribute produced a important result since no one have never try to do this. The results in fact with this type of "elimination" remained faithful to those obtained using all available features. As noted in many other papers on mortality, RF has always remained the algorithm that gave the best results in all the different case studies, followed by the Ensemble model and Gradient Boosting. The results were amazing and surprising, especially when using only 4 features (sex, race, whether the patient has already suffered from other diseases and the age group to which he belongs). Even if the performances are

lower compared to the other 3 experiments conducted, the quality and efficiency of the algorithms remains very high. However much more validation of this last result must be carried out in order to see if really only this 4 variables are good for predicting mortality in people affected with COVID-19.

7.1.1 Hope for the future

This tragedy, as Bill Gates [112] and David Quammen had predicted in the past, has caught us unprepared not only medically but also organizationally. The crisis has highlighted the importance of having detailed and reliable information quickly in order to make important decisions, such as determining the timetable and modalities of the measures to be implemented. To understand the epidemic, it is necessary to use data, but with method.
" Pas de calcul là où l'observation peut être faite " (No calculation when it is possible to make observations), as the statistician Georg von Mayr said as early as 1895, in response to his Norwegian colleague Anders Nicolai Kiaer, who was the first to propose using a part and not the whole population to obtain social and economic information at national level. Because the reality is that we cannot observe everything directly, and this has never been more evident in modern times, except in this pandemic that has made us all protagonists and made us aware of this obvious limit. It is not easy to obtain data on an entire people and, as a rule, the most severe cases have always been the main protagonists, to the detriment of people with milder symptoms.
It is clear that before the second half of 2021, the data and results available to us seemed unreliable and unrealistic, because in the absence of studies conducted on a representative sample of the whole population, we were unable to measure the prevalence and rate of actual mortality from the virus. Even the number of deaths reported daily by official sources seemed to be grossly underestimated. In the early days, when there were no tampons, doctors often wrote down causes of death that were not true; one of the most commonly used was interstitial pneumonia, although these deaths can now be certified as infections of COVID-19 given the symptoms.
So, with the availability of data increasing daily and the emergence of new collection methods, we need to think about the emergence of new possible information management models. The information system is a complex environment composed of different data collected by several agencies and data that has not yet been collected but has great potential. The pandemic has highlighted this aspect, which is not foreign to those who deal with data and estimates on a daily basis, their integration is therefore a duty.
Even if we cannot do anything to fix the past, we can change the time ahead: A

solution must be found so that a possible new pandemic does not catch us unprepared again.
It will take a special effort from medicine and data science to ensure that this does not happen again in the future, so that the deaths that have occurred have not died in vain.
The last consideration relates to digitization for medical purposes.This health emergency has certainly created the need for more modern and digital healthcare, and the need for closer and more timely care. Digitization for medical purposes is and will be essential in many countries, including Italy, where the sixth PNRR mission foreseen with the allocation of about €15 billion for Digital Health.This will certainly be a major turning point for the use and management of Big Data for the Italian healthcare system, as the information might be used in different ways. For example, for identifying better therapies, for identifying patients at higher risk and, last but not least, for simplifying patient diagnostics. Therefore, the possibility of creating automated systems that are able to respond to specific questions and not just be a kind of "black box model" is certainly the hope future for the whole world.
I hope that my work has somehow helped to promote the visibility that data scientists need to be given in relation to problems like this. Because any solution to a problem as important as this, no matter how simple it may seem, can contribute to the birth of great ideas for the future.

7.1.2 SAS & Python pros and cons

As I have already said, the war between the two environments I have used for this project is very fierce. But my personal opinion is that each instrument is different in its own way. There is no obvious winner in this case. Each instrument has its own advantages and disadvantages.
Python is probably the most widely used programming language among software engineers. Python is a simple, easy-to-learn programming language that has a larger number of libraries.
It is a fast and scalable programming language and includes many useful tools such as visualization, data analysis and data manipulation.
It is certainly a programming language that is great for analysing complicated data in the field of data science. On the other hand, SAS Viya is primarily known for statistical analysis, but it is able to combine AI capabilities with machine learning approaches to create a new product.
It also has the advantage of being a standardized code base that supports programming in SAS and other languages such as Python, R and Java.
The main advantage of Python in my case was that I could use any file size and process it directly, whereas the advantage with SAS Viya was that I did not have

to create much code but only to have in mind what I had to do. I just had to learn and make experiments inside the new environment. However the cons for SAS was that I had to divide my data into several files, as the maximum size allowed per file was 100 MB. Ultimately, I have come to the conclusion that both are valid tools that I will use in the future.

7.2 Further Possible Extensions

A list of possible future extensions is now described to allow others to contribute and improve on the work already done.

7.2.1 Survival analysis

If we want to go beyond this work, an interesting analysis could be that of survival. For this type of analysis it is imperative to have the dates of the first positive test and the time of death for each individual in order to calculate the difference between them. It is useful to understand which factors have a greater impact on the survival time of individuals (usually age is a crucial element in prolonging survival time). Survival analysis, also known as reliability analysis in engineering, aims to establish a relationship between covariates and the timing of an event. The name survival analysis comes from clinical research, where predicting time to death, i.e. survival, is often the main goal. It differs from traditional regression in that parts of the training data can only be partially observed they are censored.
Formally, each patient record consists of a set of covariates $x \in R^d$ and the time $t > 0$ at which an event occurred or the time $c > 0$ of censoring. Since censoring and experiencing and event are mutually exclusive, it is common to define an event indicator $\delta \in 0; 1$ and the observable survival time $y > 0$. The observable time y of a right-censored sample is defined as

$$y = min(t, c) = \begin{cases} t & \text{if } \delta = 1, \\ c & \text{if } \delta = 0 \end{cases} \tag{7.1}$$

Consequently, survival analysis requires models that take into account these unique features of such a dataset. Our main interest should be to investigate whether there are subgroups that differ in survival and whether it is possible to predict survival times.
A key variable in survival analysis is the so-called survival function, which relates time to the probability of survival after a given point in time. Denote by t a continuous non-negative random variable corresponding to the survival time of a

patient. The survival function $S(t)$ provides a probability of survival after time t and is defined as

$$\mathrm{S(t)} = \mathrm{P(T} > \mathrm{t)}$$

Another application of this type of analysis could be to test survival functions for different treatments, for example to test the effectiveness of different types of vaccines or a new treatment being developed in the feature, and so see how many patients received the standard treatment and how many received the new drug. This can be done by having half of the patients receive the new treatment.
Of course, this kind of test requires a new feature: if the patient dies, we need the date of this event to be registered, otherwise this analysis is not feasible. If it is done, the final result could be of great importance for the scientific community.

7.2.2 Performing different Feature Selection

Univariate and multivariate filtering methods are very popular, especially for large datasets, because they are usually very fast and much less computationally intensive with respect to other methods. Specifically, they use a metric to score each individual feature or a subset of them together. The most used metrics in filtering methods include correlation coefficient, Fisher score, entropy and chi-square parameters.
Thus it would be interesting to compare the outputs given by the use of these type of techniques with the results obtained in this work. All of this it will be useful to build a parsimonious workflow by excluding irrelevant features.

7.2.3 Focus on Pre-existing medical conditions

Recall that in the dataset used there was the attribute "medcond_yn "which, when it took the value "Yes", meant that the patient had pre-existing conditions. Instead of just saying "Yes" or "No", it would have been interesting to understand what pre-existing conditions existed. This is because there are many diseases that are associated with a higher probability of death for those who have contracted COVID-19. Among these diseases we remember, for example, diabetes, obesity and in general cardiovascular diseases. In the case of diabetes[113] there is some evidence that marked glycaemic overcompensation may predispose to a higher risk, but these are still few proof and for this it would be useful to do further research. The only evidence so far that tends to support this is that the disease takes a more severe course (pneumonia or respiratory failure) in diabetics.
While, in the case of heart disease [114] , for example, in Italy in March 2020, 75 % of the first 155 dead patients suffered from hypertension due to the new coronavirus infection, while 70 % had ischaemic heart disease. These are numbers

that confirm what has already emerged in China before, and was published at the end of February by the Chinese Centre for Disease Control and Prevention in the medical journal JAMA.
Furthermore, according to the Chinese data, the lethality of the virus - starting from an average value in the population of just over 2 % - increases to 6 % in people with hypertension and even reaches 10 % in patients with heart failure or other cardiovascular diseases or chronic cerebrovascular diseases. People with obesity (even mild) have a higher risk of developing severe forms of Covid-19, which can lead to death. A year after the pandemic outbreak, several studies have confirmed that the more severe the obesity, the higher the risk of those who test positive for SARS-CoV-2 infection. This is also shown by the reports of those who are treated daily in the intensive care unit. Younger patients admitted for bilateral interstitial pneumonia, for example, did not have many of the risk factors associated with Covid-19 severity. Most notably, age. But in many cases, they were people who suffered from obesity. And possibly some of the risk factors and diseases caused by it: such as high blood pressure and type 2 diabetes [115].
If we do so and maybe discover something useful, it may be possible to identify critical factors for mortality and thus to optimize patient treatment strategies.

7.2.4 Comorbidity

The term "comorbidity" has been on everyone's lips with the first studies for referring to people at higher risk of developing adverse symptoms due to COVID-19. 5 groups were formed according to age and disease state[116]:
- category 1: highly frail people;
- category 2: people aged between 70 and 79;
- category 3: people aged between 60 and 69;
- category 4: people with comorbidities aged <60 years, without gravity connotation reported for extremely vulnerable people;
- category 5: rest of the population aged <60 years.

In health care, the term comorbidity refers to the coexistence of several different pathologies in the same person. The data confirm that persons suffering from chronic diseases, when infected with the virus, have a higher risk of manifesting more severe forms of coronavirus (more likelihood of dying), precisely because of the disease for which they have obtained an exemption. These are therefore people who need to be protected and it is essential that they are specially monitored to protect their health and the Health System.
Although Covid-19 can be fatal even without concomitant diseases, according to the ISTAT 2020 study, the majority (71.8%) of those who died and tested positive for SARS-CoV-2 had one or more chronic diseases (comorbidities) that played a crucial role in worsening their prognosis.

Therefore, the ability to have more explicit information within the disease variable would provide the opportunity to identify the most important comorbidities related to the primary endpoint and thus create different clusters with these comorbidities, for example by using discriminant analysis. In this manner the efficiency of the Health care system can be improved and will guarantee to these people a lower chance of death.

7.2.5 Geographical area studies and similar pattern among different states

Based on various datasets with information on the number of people who died who were positive but were then cured, and with geographical connotations related to the various areas, it would be interesting to find climatic or area-related characteristics that are the same or similar in more geographical places of the world. This could be useful as it might explain the impact of different geographical locations on cases and deaths caused by COVID-19 and, if possible, the identification of safe geographical areas.

REFERENCES

[1] Vinay Chamola, Vikas Hassija, Vatsal Gupta, and Mohsen Guizani. A comprehensive review of the covid-19 pandemic and the role of iot, drones, ai, blockchain, and 5g in managing its impact. *Ieee access*, 8:90225–90265, 2020.

[2] T. Ryan Gregory. Understanding evolutionary trees, 2022. `https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-008-0035-x`,.

[3] Song Wan and Jun Liu. Decode a ticking time-bomb. *Journal of Thoracic Disease*, 12(9):4598, 2020.

[4] Wim Naudé. Artificial intelligence against covid-19: An early review. *papers.ssrn.com*, 2020.

[5] Raju Vaishya, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem. Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339, 2020.

[6] Jennifer Drahos, Manxia Wu, William F Anderson, Katrina F Trivers, Jessica King, Philip S Rosenberg, Christie Eheman, and Michael B Cook. Regional variations in esophageal cancer rates by census region in the united states, 1999–2008. *PloS one*, 8(7):e67913, 2013.

[7] Meenu Gupta, Rachna Jain, Simrann Arora, Akash Gupta, Mazhar Javed Awan, Gopal Chaudhary, and Haitham Nobanee. Ai-enabled covid-19 outbreak analysis and prediction: Indian states vs. union territories. *Gupta, M., Jain, R., Arora, S., Gupta, A., Awan, MJ, Chaudhary, G., & Nobanee, H.(2021). AI-Enabled COVID-19 Outbreak Analysis and Prediction: Indian States vs. Union Territories. Cmc-Computers Materials & Continua*, 67(1): 933–950, 2021.

[8] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, 19(3):141–154, 2021.

[9] Lily Kuo. *China confirms human-to-human transmission of coronavirus*, chapter 1. the Guardian, 1 2020. URL `https://www.theguardian.com/world/2020/jan/20/coronavirus-spreads-to-beijing-as-china-confirms-new-cases`. Available on line.

[10] Oriana Liso Alessandra Corica and Mauro Rancati. *Coronavirus, i contagi nel Lodigiano sono 15: i primi sono un 38enne di Codogno e sua moglie. In isolamento 250 persone*, chapter 1. La Republica, 2 2020. URL `https://milano.repubblica.it/cronaca/2020/02/21/news/coronavirus_a_milano_contaggiato_38enne_e_un_italiano_ricoverato_a_codogno-249121707/`. Available on line.

[11] WHO. Who director-general's opening remarks at the media briefing on covid-19, 2020. `https://www.who.int/speeches/detail/who-director-general-opening-remarks-at-the-media-briefing-covid`, Last accessed on 2022-03-30.

[12] Marco Cattaneo. Le sfide aperte della pandemia. *Le Scienze*, 1(645):7, 2022.

[13] World Health Organisation. Report of the who-china joint mission on coronavirus disease 2019 (covid-19), 2020. `https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf`, Last accessed on 2022-05-11.

[14] Vineet D Menachery, Boyd L Yount, Kari Debbink, Sudhakar Agnihothram, Lisa E Gralinski, Jessica A Plante, Rachel L Graham, Trevor Scobey, Xing-Yi Ge, Eric F Donaldson, et al. A sars-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nature medicine*, 21(12): 1508–1513, 2015.

[15] JON COHEN. 'why many scientists say unlikely sars-cov-2 originated lab leak, 2021. `https://www.science.org/content/article/why-many-scientists-say-unlikely-sars-cov-2-originated-lab-leak`, Last accessed on 2022-08-09.

[16] Ministero della salute italiana. Che cosa sappiamo sulle varianti del sars-cov-2, 2022. `https://www.salute.gov.it/portale/nuovocoronavirus/dettaglioFaqNuovoCoronavirus.jsp?lingua=italiano&id=250`,.

[17] Eva Benelli. La strategia anti covid di cuba:vaccinare i bambini. *Le Scienze*, 1(645):10–11, 2022.

[18] Joseph Stiglitz. Le disugualianze sono peggiorate in modo netto. *Le Scienze*, 1(645):28–29, 2022.

[19] Aaron S Bernstein, Amy W Ando, Ted Loch-Temzelides, Mariana M Vale, Binbin V Li, Hongying Li, Jonah Busch, Colin A Chapman, Margaret Kinnaird, Katarzyna Nowak, et al. The costs and benefits of primary prevention of zoonotic pandemics. *Science advances*, 8(5):4183, 2022.

[20] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[21] CR Rao and Venkat N Gudivada. *Computational analysis and understanding of natural languages: principles, methods and applications*. Elsevier, 2018.

[22] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, 3:19–48, 2010.

[23] Susan Li. Building a logistic regression in python, step by step. *Towards Data Science*, 17, 2017.

[24] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[25] S Vijayarani and M Muthulakshmi. Comparative analysis of bayes and lazy classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(8):3118–3124, 2013.

[26] Mucahid Mustafa Saritas and Ali Yasar. Performance analysis of ann and naive bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, 7(2):88–91, 2019.

[27] Skelearn. 'decision trees—scikit-learn 0.24.1 documentation, 2020. `https://scikit-learn.org/stable/modules/tree.html`, Last accessed on 2022-08-09.

[28] Scikit Learn. Random forest classifier. *línea]. Disponible en: https://scikit-learn. org/stable/modules/generated/sklearn. ensemble. RandomForestClassifier. html [Accedido 14 Dic. 2021]*, 2021.

[29] Bashir Iwendi and Peshkar. Covid-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public heath*, July 2020.

[30] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[31] Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14(1):5–16, 1998.

[32] Stefan Conrady and Lionel Jouffe. *Bayesian networks and BayesiaLab: a practical introduction for researchers*, volume 9. Bayesia USA Franklin, 2015.

[33] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

[34] Mahmoud Omid, Asghar Mahmoudi, and Mohammad H Omid. Development of pistachio sorting system using principal component analysis (pca) assisted artificial neural network (ann) of impact acoustics. *Expert Systems with Applications*, 37(10):7205–7212, 2010.

[35] Brett Wujek, Patrick Hall, and Funda Günes. Best practices for machine learning applications. *SAS Institute Inc*, 2016.

[36] R. Exsilio Inc LJoshi. 'accuracy, precision, recall f1 score: Interpretation of performance measures', 2016. `https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation`, Last accessed on 2022-05-09.

[37] Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[38] Kjetil Søreide, Julie Hallet, Jeffrey B Matthews, Andreas Anton Schnitzbauer, Pål Dag Line, PBS Lai, Javier Otero, Dario Callegaro, Shelley G Warner, Nancy N Baxter, et al. Immediate and long-term impact of the covid-19 pandemic on delivery of surgical services. *Journal of British Surgery*, 107(10):1250–1261, 2020.

[39] Ze-Liang Chen, Wen-Jun Zhang, et al. From severe acute respiratory syndrome-associated coronavirus to 2019 novel coronavirus outbreak: similarities in the early epidemics and prediction of future trends. *Chinese medical journal*, 133(09):1112–1114, 2020.

[40] Angela AR de Sá, Jairo D Carvalho, and Eduardo LM Naves. Reflections on epistemological aspects of artificial intelligence during the covid-19 pandemic. *AI & society*, pages 1–8, 2021.

[41] Abdul Majeed and Seong Oun Hwang. Data-driven analytics leveraging artificial intelligence in the era of covid-19: An insightful review of recent developments. *Symmetry*, 14(1):16, 2021.

[42] Rakesh Garg, Anuradha Patel, and Wasimul Hoda. Emerging role of artificial intelligence in medical sciences—are we ready! *Journal of Anaesthesiology, Clinical Pharmacology*, 37(1):35, 2021.

[43] Joseph Bullock, Alexandra Luccioni, Katherine Hoffman Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. Mapping the landscape of artificial intelligence applications against covid-19. *Journal of Artificial Intelligence Research*, 69:807–845, 2020.

[44] Sara Hosseinzadeh Kassania, Peyman Hosseinzadeh Kassanib, Michal J Wesolowskic, Kevin A Schneidera, and Ralph Detersa. Automatic detection of coronavirus disease (covid-19) in x-ray and ct images: a machine learning based approach. *Biocybernetics and Biomedical Engineering*, 41(3): 867–879, 2021.

[45] D Coldeway. Molecule. one uses machine learning to make synthesizing new drugs a snap. *TechCrunch, 3October*, 2019.

[46] Arash Keshavarzi Arshadi, Julia Webb, Milad Salem, Emmanuel Cruz, Stacie Calad-Thomson, Niloofar Ghadirian, Jennifer Collins, Elena Diez-Cecilia, Brendan Kelly, Hani Goodarzi, et al. Artificial intelligence for covid-19 drug discovery and vaccine development. *Frontiers in Artificial Intelligence*, 3:65, 2020.

[47] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*, 2020.

[48] Lawrence O.Gostin. Le istituzioni sanitarie globali sono arrivate al limite. *Le Scienze*, 1(645):28–29, 2022.

[49] A Rivas. Drones and artificial intelligence to enforce social isolation during covid-19 outbreak. *Medium Towards Data Sci*, 26, 2020.

[50] Our world in Data. Dashboards coronavirus pandemic (covid-19), 2022. `https://ourworldindata.org/coronavirus`, Last accessed on 2022-06-18.

[51] Joon Lee. Patient-specific predictive modeling using random forests: an observational study for the critically ill. *JMIR medical informatics*, 5(1): e6690, 2017.

[52] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *science*, 343(6176):1203–1205, 2014.

[53] Marcelo Ponce and Amit Sandhel. covid19. analytics: An r package to obtain, analyze and visualize data from the coronavirus disease pandemic. *arXiv preprint arXiv:2009.01091*, 2020.

[54] Adam Palayew, Ole Norgaard, Kelly Safreed-Harmon, Tue Helms Andersen, Lauge Neimann Rasmussen, and Jeffrey V Lazarus. Pandemic publishing poses a new covid-19 challenge. *Nature Human Behaviour*, 4(7):666–669, 2020.

[55] Ewen Callaway. Will the pandemic permanently alter scientific publishing? *Nature*, 582(7811):167–169, 2020.

[56] Epidy Health Research. Covid-19 risk factors: literature database & meta-analysis, 2022. `https://corona.epidy.com`,.

[57] Artur Kiulian. Fighting coronavirus with artificial intelligence, 2020. `https://www.coronawhy.org`, Last accessed on 2020-05-18.

[58] Alessandro Vespignani. I numeri della pandemia. *Le Scienze*, 1(645):70–75, 2022.

[59] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *European radiology*, 31(8):6096–6104, 2021.

[60] Alison M Darcy, Alan K Louie, and Laura Weiss Roberts. Machine learning and the profession of medicine. *Jama*, 315(6):551–552, 2016.

[61] Kim Sul Gi. Big data in healthcare hype and hope. *CDE*, pages 122–125, 2013.

[62] Nitesh V Chawla and Darcy A Davis. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine*, 28(3):660–665, 2013.

[63] Iñaki Inza, Borja Calvo, Rubén Armañanzas, Endika Bengoetxea, Pedro Larranaga, and José A Lozano. Machine learning: an indispensable tool in bioinformatics. In *Bioinformatics methods in clinical research*, pages 25–48. Springer, 2010.

[64] C.Poirier D.Liu, L.Clemente and M.Santillana. a machine learning methodology for real time forecasting of the 2019-2020 covid-19 outbreak. using internet searches,new alerts, and estimates from mechanistic models. *arXiv*, 2004.04019, March 2020.

[65] L. Wang and A.Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv*, 2003.09871, March 2020.

[66] Y.Choi Beck, B.Shin and K.Kang. Predicting commercially available antiviral drugs that may act on the novel coronavirus through a drug-target interaction deep learning model. *Computational and Structural Biotechnology Journal*, 18:784–790, March 2020.

[67] H Ko, H Chung, WS Kang, C Park, DW Kim, SE Kim, CR Chung, RE Ko, H Lee, JH Seo, et al. Artificial intelligence can predict the mortality of covid-19 patients at the admission time using routine blood samples. *Journal of Medical Internet Research*, 2020.

[68] Arjun S Yadaw, Yan-chak Li, Sonali Bose, Ravi Iyengar, Supinda Bunyavanich, and Gaurav Pandey. Clinical features of covid-19 mortality: development and validation of a clinical prediction model. *The Lancet Digital Health*, 2(10):e516–e525, 2020.

[69] Ashis Kumar Das, Shiba Mishra, and Saji Saraswathy Gopalan. Predicting covid-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ*, 8:e10083, 2020.

[70] Chansik An, Hyunsun Lim, Dong-Wook Kim, Jung Hyun Chang, Yoon Jung Choi, and Seong Woo Kim. Machine learning prediction for mortality of patients diagnosed with covid-19: a nationwide korean cohort study. *Scientific reports*, 10(1):1–11, 2020.

[71] Manuel Sánchez-Montañés, Pablo Rodríguez-Belenguer, Antonio J Serrano-López, Emilio Soria-Olivas, and Yasser Alakhdar-Mohmara. Machine learning for mortality analysis in patients with covid-19. *International journal of environmental research and public health*, 17(22):8386, 2020.

[72] Alejandro López-Escobar, Rodrigo Madurga, José María Castellano, Sara Velázquez, Rafael Suárez del Villar, Justo Menéndez, Alejandro Peixoto, Sara Jimeno, Paula Sol Ventura, and Santiago Ruiz de Aguiar. Risk score for predicting in-hospital mortality in covid-19 (rim score). *Diagnostics*, 11 (4):596, 2021.

[73] Ana Teresa Ferreira, Carlos Fernandes, José Vieira, and Filipe Portela. Pervasive intelligent models to predict the outcome of covid-19 patients. *Future Internet*, 13(4):102, 2021.

[74] Mohammad Pourhomayoun and Mahdi Shakibi. Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*, 20:100178, 2021.

[75] Prathamesh Parchure, Himanshu Joshi, Kavita Dharmarajan, Robert Freeman, David L Reich, Madhu Mazumdar, Prem Timsina, and Arash Kia. Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with covid-19. *BMJ supportive & palliative care*, 12(e3):e424–e431, 2022.

[76] Xiaoran Li, Peilin Ge, Jocelyn Zhu, Haifang Li, James Graham, Adam Singer, Paul S Richman, and Tim Q Duong. Deep learning prediction of likelihood of icu admission and mortality in covid-19 patients using clinical variables. *PeerJ*, 8:e10337, 2020.

[77] Julius R Migriño Jr and Ani Regina U Batangan. Using machine learning to create a decision tree model to predict outcomes of covid-19 cases in the philippines. *Western Pacific surveillance and response journal: WPSAR*, 12 (3):56, 2021.

[78] Mohamad Nikpouraghdam, Alireza Jalali Farahani, GholamHossein Alishiri, Soleyman Heydari, Mehdi Ebrahimnia, Hossein Samadinia, Mojtaba Sepandi, Nematollah Jonaidi Jafari, Morteza Izadi, Ali Qazvini, et al. Epidemiological characteristics of coronavirus disease 2019 (covid-19) patients in iran: A single center study. *Journal of Clinical Virology*, 127:104378, 2020.

[79] Javier Andreu-Perez, Carmen CY Poon, Robert D Merrifield, Stephen TC Wong, and Guang-Zhong Yang. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208, 2015.

[80] Ruo Xu, Dan Nettleton, and Daniel J Nordman. Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65, 2016.

[81] Michael J Joyner and Nigel Paneth. Seven questions for personalized medicine. *Jama*, 314(10):999–1000, 2015.

[82] Joaquín Dopazo, Douglas Maya-Miles, Federico García, Nicola Lorusso, Miguel Ángel Calleja, María Jesús Pareja, José López-Miranda, Jesús Rodríguez-Baño, Javier Padillo, Isaac Túnez, et al. Implementing personalized medicine in covid-19 in andalusia: An opportunity to transform the healthcare system. *Journal of Personalized Medicine*, 11(6):475, 2021.

[83] Erwin Cornelius, Olcay Akman, and Dan Hrozencik. Covid-19 mortality prediction using machine learning-integrated random forest algorithm under varying patient frailty. *Mathematics*, 9(17):2043, 2021.

[84] W.Zhao Z.Tang and D.Shein. Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images. *arXiv*, 2003.11988, 2020.

[85] C.L. de Lima V.A de Freitas Barbosa, M.A. de Santa and R.B Calado. Covid-19 rapid test by combining a random forest based web system and blood tests. *medRXiv*, 2003.09871, 2020.

[86] D.Kumar V.K. Gupta and A.Sardana. Prediction of covid-19 confirmed, death and cured cases in india using random forest model. *Big Data Mining and Analytics*, 4:116–123, June 2021.

[87] H. Lim C.An and S.W.Kim. Machine learning prediction for mortality of patients diagnosed with covid-19: a nationwide korean cohort study. *Scientific Reports*, 10, October 2020.

[88] H.Yu J.Wang and Y.Luo. A descriptive study of random forest algorithm for predicting covid-19 patients outcome. *PeerJ*, 8, August 2020.

[89] Sumayh S Aljameel and Khan. Machine learning-based model to predict the disease severity and outcome in covid-19 patients. *Scientific programming*, 2021, 2021.

[90] Wes McKinney and PD Team. Pandas-powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, 1625, 2015.

[91] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc.", 2012.

[92] Giancarlo Zaccone, Md Rezaul Karim, and Ahmed Menshawy. *Deep learning with TensorFlow.* Packt Publishing Ltd, 2017.

[93] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[94] Kevin D Smith and Xiangxiang Meng. *SAS Viya: The Python Perspective.* SAS Institute, 2017.

[95] Ben Guarino and Ariana Eunjung Cha. 'the weapon that will end the war': First coronavirus vaccine shots given outside trials in u.s, 2020. https://www.washingtonpost.com/nation/2020/12/14/first-covid-vaccines-new-york/, Last accessed on 2022-08-09.

[96] Quantics biostatistics. 'handling missing data in clinical trials', 2020. https://www.quantics.co.uk/blog/post-covid-19-handling-missing-data-in-clinical-trials/, Last accessed on 2022-08-18.

[97] Mohammad Pourhomayoun, Ebrahim Nemati, B Mortazavi, and M Sarrafzadeh. Context-aware data analytics for activity recognition. *Data Analytics*, 2015.

[98] Non Asian. Risk for covid-19 infection, hospitalization, and death by race/ethnicity. *Risk*, 9:27, 1921.

[99] Mario Negri. Medicina personalizzata: che cos'è e come si applica?, 2022. https://www.marionegri.it/magazine/medicina-personalizzata, Last accessed on 2022-08-08.

[100] Yoon-Joo Park, Byung-Chun Kim, and Se-Hak Chun. New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis. *Expert systems*, 23(1):2–20, 2006.

[101] Maryam Panahiazar, Vahid Taslimitehrani, Naveen L Pereira, and Jyotishman Pathak. Using ehrs for heart failure therapy recommendation using multidimensional patient similarity analytics. *Studies in health technology and informatics*, 210:369, 2015.

[102] ISS. 'gender differences in covid-19: the importance of sex-disaggregated data', 2020. https://www.epicentro.iss.it/en/coronavirus/sars-cov-2-gender-differences-importance-sex-disaggregated-data.

[103] Li Yan, Hai-Tao Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li, Mingyang Zhang, Yuqi Guo, Ying Xiao, et al. Prediction of criticality in patients with severe covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in wuhan. *MedRxiv*, 27:2020, 2020.

[104] Liping Sun, Fengxiang Song, Nannan Shi, Fengjun Liu, Shenyang Li, Ping Li, Weihan Zhang, Xiao Jiang, Yongbin Zhang, Lining Sun, et al. Combination of four clinical indicators predicts the severe/critical symptom of patients infected covid-19. *Journal of Clinical Virology*, 128:104431, 2020.

[105] Kenneth CY Wong, Yong Xiang, and Hon-Cheong So. Uncovering clinical risk factors and prediction of severe covid-19: A machine learning approach based on uk biobank data. *MedRxiv*, pages 2020–09, 2021.

[106] Simona Anticoli. 'gender differences in covid-19: the importance of sex-disaggregated data', 2020. `https://www.epicentro.iss.it/sars-cov-2-gender-differences-importance-sex-disaggregated-data`.

[107] Long-quan Li, Tian Huang, Yong-qing Wang, Zheng-ping Wang, Yuan Liang, Tao-bi Huang, Hui-yun Zhang, Weiming Sun, and Yuping Wang. Covid-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. *Journal of medical virology*, 92(6):577–583, 2020.

[108] M Mahendra, Abhishek Nuchin, Ranjith Kumar, S Shreedhar, and Padukudru Anand Mahesh. Predictors of mortality in patients with severe covid-19 pneumonia—a retrospective study. *Advances in Respiratory Medicine*, 89(2): 135–144, 2021.

[109] Peter G Gibson, Ling Qin, and Ser Hon Puah. Covid-19 acute respiratory distress syndrome (ards): clinical features and differences from typical pre-covid-19 ards. *Medical Journal of Australia*, 213(2):54–56, 2020.

[110] BMJ. 'healthcare workers 7 times as likely to have severe covid-19 as other workers', 2020. `https://www.bmj.com/healthcareworkers7-times-as-likely-to-have-severe-covid-19/`.

[111] Marina Treskova-Schwarzbach, Laura Haas, Sarah Reda, Antonia Pilic, Anna Borodova, Kasra Karimi, Judith Koch, Teresa Nygren, Stefan Scholz, Viktoria Schönfeld, et al. Pre-existing health conditions and severe covid-19 outcomes: an umbrella review approach and meta-analysis of global evidence. *BMC medicine*, 19(1):1–26, 2021.

[112] TED. 'la prossima pandemia?non siamo pronti', 2015. `https://www.ted.com/talks/bill_gates_the_next_outbreak_we_re_not_ready?language=it#t-86695`.

[113] Ospedale San Raffaele. 'covid-19 e diabate', 2020. `https://www.hsr.it/news/2020/marzo/coronavirus-diabete-rischi`.

[114] Ospedale San Raffaele. 'covid-19 e malattie cardiovascolari', 2020. `https://www.hsr.it/news/2020/marzo/coronavirus-cardiopatie-hypertension`.

[115] Fondazione veronesi. 'covid-19 e obesità', 2020. `https://www.fondazioneveronesi.it/magazine/articoli/alIMENTO/covid-19`.

[116] Istituto Auxologico Italiano. 'covid-19 e comorbidity', 2020. `https://www.auxologico.it/comorbidita-categoria-prioritaria-vaccini-covid`.