ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

CICLO XXXIII

Settore Concorsuale: 01/B1 - INFORMATICA
Settore Scientifico Disciplinare: INF/01 - INFORMATICA

On the implications of big data and machine learning in the interplay between humans and machines

Presentata da
GIOVANNI DELNEVO

Coordinatore Dottorato
Prof. ANDREA CAVALLI

Supervisore
Prof. MARCO ROCCETTI

Co-supervisore
Prof. SILVIA MIRRI

ESAME FINALE ANNO 2022

This dissertation is submitted for the degree of Doctor of Philosophy in Data Science and Computation

January 2022

Any sufficiently advanced technology
is indistinguishable from magic
— Sir Arthur C. Clarke

After all, to the well-organized mind,
death is but the next great adventure
— A. Dumbledore

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 50,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 40 figures.

January 2022

Acknowledgements

The Ph.D. has been a long journey and I would like to thank all the people I have collaborated with and who have shared part of this journey with me. First, I would like to thank my supervisor, Prof. Marco Roccetti, for his guidance during this research. His dedication to work and his attention to detail were inspiring. Then, I would like to thank my co-supervisor, Prof. Silvia Mirri, for her thoughtful comments and recommendations and for the ongoing support.

I would also like to thank all the other members of the research group, in particular, Prof. Paola Salomoni for her mentorship, Prof. Catia Prandi who is an inexhaustible source of ideas, and for her critical feedback, and Dr. Chiara Ceccarini for the countless chats going to get water.

Then, I would like to thank all the people and companies with whom I have had the opportunity to collaborate in this long journey. A special thank goes to Prof. Sobrero and Prof. Lipparini for their brilliant point of view. And even if we shared a short part of this journey, I would like to thank Nicolò Zagni, for his enthusiasm in facing the work.

I'm deeply indebted to my family, for always supporting me in my choices and for constantly encouraging me to improve and to my friends, for always having been there and for the happy distractions.

Finally, I could not have completed this long journey without Elisa, who always supported me with her love.

Abstract

Big data and machine learning are profoundly shaping social, economic, and political spheres, becoming part of the collective imagination. In recent years, barriers have fallen and a wide range of products, services, and resources, that exploit Artificial Intelligence, have emerged. Hence, it becomes of fundamental importance to understand the limits and, consequently, the potentialities of predictions made by a machine that learns directly from data. Understanding the limits of machine predictions would allow dispelling false beliefs about the potentialities of machine learning algorithms, avoiding at the same time possible misuses. To tackle this problem, completely different research lines are emerging, that focus on different aspects. In this thesis, we study how the presence of big data and artificial intelligence influences the interaction between humans and computers. Such a study should produce some high-level reflections that can contribute to the framing of how the interaction between humans and computers has changed, since the presence of big data and algorithms that can make computers somehow *intelligent*, albeit with some limitations. In the different chapters of the thesis, various case studies that we faced during the Ph.D. are described, chosen specifically for their peculiar characteristics. Starting from the obtained results, we provide several high-level reflections on the implications of the interaction between humans and machines.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

In this chapter, we present the background and context of this thesis and we introduce the research questions that drove the study. Finally, we detail the outline of the thesis.

1.1 Research Background

In 2009, Schmidt and Lipson [132] were able to distil natural laws directly from experimental data, without any prior knowledge about physics, kinematics, or geometry. The experimental data consisted simply in the position of an air-track oscillator and a double pendulum over time. Employing symbolic regression and varying the parameters considered, the authors were able to discover Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation. Similar results have contributed to fuel the current of thought according to which, big data, coupled with machine learning algorithms, are engendering a paradigm shift from a knowledge-driven science to a data-driven one [77]. Knowledge-driven science is the one of the scientific method, based on hypothesis testing. Scientists hypothesize models and then conduct experiments that confirm or refute such models. In many fields, such as physics or biology, these experiments can no longer take place for several reasons (e.g., the energies are too high, the accelerators too expensive, ...). Hence, many are putting into

question the scientific method in favour of massive amounts of data and applied mathematics, since they offer a whole new way of understanding the world [8]. In particular, Anderson begins his reasoning starting from George Box's famous maxim "All models are wrong, but some are useful" [17], stating that, in the Petabyte Age, there is no need to settle for models at all. It is clear to all scientists that correlation is different from causation, as the countless examples of spurious correlations demonstrate [80, 138]. According to Anderson instead, such an abundance of data, in the order of petabytes, allow to say that correlation is enough, making further explanation unnecessary: "With enough data, the numbers speak for themselves".

As opposed to these positions, many scientists are claiming the importance of conceptual insights even when working with big data and Artificial Intelligence (AI) [25, 36, 66, 64]. Calude and Longo [25], in their work, start from Anderson's statements with the aim of documenting the danger of subsuming and replacing the scientific approach with the search of correlations in big data. Their approach is based on ergodic theory, Ramsey theory, and algorithmic information theory. They proved that arbitrary correlations are present in very large databases but they do not appear due to the nature of the data, but due to their size. They conclude that too much information tends to behave like very little information and that the analysis of big data can enrich the scientific method, but can not replace it. In [36], focusing on biology and medicine, the authors point out the weaknesses of pure big data approaches which fail to provide conceptual accounts, regardless of the "depth" and the sophistication of the algorithms used. Instead, Hosni and Vulpiani [66], taking as a case study the weather forecasts, concluded that a context-dependent balance between modelling and quantitative analysis stands out as the best forecasting strategy. Finally, Holzinger, Haibe-Kains, and Jurisica [64] affirmed that automated machine learning can work only in narrow contexts. In other, like the medical one, human experts are still required since they are the only ones able to understand the overall context. According to their vision, AI-technology has to empower/augment humans, rather than substitute them, following the human-in-the-loop or the human-in-control paradigms.

This conflictual situation does not regard only science, but it is reflected in different areas of society. In fact, big data and machine learning are profoundly shaping social, economic, and political spheres, becoming part of the collective imagination [44]. In recent years, barriers have fallen and a wide range of products, services, and resources, that exploit Artificial Intelligence, have emerged [103]. Within this context, it becomes of fundamental importance to understand the limits and, consequently, the potentialities of predictions made by a machine that learns directly from data [22]. Understanding the limits of machine prediction would allow dispelling false beliefs about the potentialities of machine learning algorithms, avoiding at the same time possible misuses.

To tackle this problem, completely different research lines are emerging. Some researchers are trying to determine the limits of such predictions, studying the mathematical foundations of machine learning [118]. An example of this research line is reported in [12], where the authors treat the problem of identifying the learnable using a mathematical framework. They focused on estimating the maximum problem and they found out that, in some cases, a solution of such a problem is equivalent to the continuum hypothesis, making the learnability undecidable (i.e., impossible to prove nor refute). A completely different research line is the one proposed by Rahwan et al. in [120], where the authors frame the emerging interdisciplinary field of machine behaviour, whose purpose is to study the behaviour exhibited by intelligent machines, to be able to control their actions, reap their benefits and minimize their harms. The main motivations behind it are the ever-increasing role of algorithms in society, their complexity and opacity, and the challenge of predicting their effects on humanity. The authors suggest investigating different dimensions (i.e., function, mechanism, development, and evolutionary history) at different scales of inquiry (i.e., individual, collective, and hybrid). Finally, in contrast to the two research lines described above, there is an unbridled use of machine learning, without any a priori theoretical reflections. There is not a theory to prove, but there are data. Starting from these massive amounts of data, there is a tendency to use increasingly sophisticated algorithms and, then, to tune the hyper-parameters

of such algorithms till the point that some relations are extracted from the data. But very often, such relationships exist only in the dataset and not in the real world. Otherwise, it is frequent that the same dataset, analysed by different people, gives completely different results, due to multiple factors, creating the so-called crisis of reproducibility [67]. This tendency upends the scientific world. In fact, while before a theory was generally true under certain conditions, now it is possible to define a different theory for each dataset.

1.2 Research Questions

In this broad spectrum between theorists and those who use machine learning indiscriminately, in my thesis I want to focus on how the presence of big data and artificial intelligence influences the interaction between humans and computers. Such a study should produce some high-level reflections that can contribute to the framing of how the interaction between humans and computers has changed, since the presence of big data and algorithms that can make computers *intelligent*, albeit with some limitations.

RQ-1 *Can an imperfect deep learning algorithm still be useful?*

Machines must be perfect. The claim of perfection is probably a cultural projection of the western culture and the results of the Christian apocalyptic thinking [27]. This current of thought leads to a binary vision which contrasts technological apocalypse against techno-utopianism. The result of such a vision is a neurosis. On the one hand, there is the desire for a perfect machine, able to have 100% accuracy. On the other hand, a single case of failure is enough to put into question the whole system. When working in the real world, the perfect machine does not (yet) exist. Our intuition is that the key question is not whether or not to rely on an imperfect machine. Rather it is deciding how much autonomy we are willing to give to them. According to this vision, machines have to evaluate all those cases in which they are highly confident while humans

will analyse the remaining cases, the most doubtful ones, in which the machine tends to make mistakes. In this way, we return to have the same concept of machines that humanity has always had, consisting of taking away the useless work.

RQ-2 *How much is the impact of human-in-the-loop approaches in data preparation?*

Since machine learning algorithms learn by processing massive piles of data, they are significantly influenced by the way training data are organized and modeled [115]. No matter how much sophisticated is the algorithm that will analyze a dataset, equally critical is the quality of the data and the way data are organized [158]. Data preparation based on human-in-the-loop approaches can successfully lead to the extrapolation of just the training data that represent the complex statistical phenomenon under observation. Such a phase changed the characteristics of the dataset from a statistical point of view, improving at the same time the performance of the deep learning algorithm.

RQ-3 *How can high-dimensional Pareto-distributed categorical data be handled to train a deep learning algorithm?*

Data can be divided into numerical and categorical. While numerical data are measurable in nature and easily manageable, categorical data, instead, represents a collection of information, that can be divided into groups. They have to be adequately prepared before feeding a deep learning algorithm. They can easily lead to an increase in the dimensions of the space under investigation. This phenomenon may go so fast that the available data become sparse, with a consequent loss of statistical significance. For this reason, it is important the role of the humans involved in the data science processes, since they can provide insights and propose new methods to address possible problems. This is exactly our case, where we were able to devise a method to handle high-dimensional Pareto-distributed categorical data.

RQ-4 *Can statistical approaches combined with machine learning be used on big data to observe different phenomena?*

The big data era has sparked the debate which opposes data-driven to knowledge-driven science. With our experience, we want to contribute to this discussion about the paradigm shift from knowledge-driven to data-driven science. While we are well aware that big data have been one of enabling factors for the resurgence of artificial intelligence, and in particular of machine learning, in our opinion, the theory has still a central role. In particular, we believe that artificial intelligence and statistical methods can be successfully exploited to observe different phonomena with the aim of confirming or refuting hypotheses, based on conceptual insights from the humans involved in the experiments.

RQ-5 *How humans should interact with algorithms to codify and transfer knowledge to them?*

Humans beings make choices and take decisions, which they consider optimal, based on all the contextual information available, that are often limited. Take now machines, which in the end are only classifiers, trained to reach a certain probabilistic accuracy on which the possibility of making errors hangs. Humans demand that machines, once properly trained, are able to reproduce or simulate the human optimal. But obviously, their predictions, and consequently their behaviours, present some limitations and strictly depends on the cognification of the context of interest and on the knowledge transferred from the human trainer to the machines.

1.3 Outline

Given the background concepts exposed above, a detailed overview of the work presented in this thesis is provided below:

Chapter 1 the first chapter introduces the context of the thesis and presents the research questions. Lastly, it summarized the content of the thesis.

Chapter 2 This chapter presents a case study relative to the design and implementation of a deep learning algorithm able to predict water meter failures based on historical data about water consumption provided by a company that supplies water in Northern Italy. Starting from this case study, we draw several high-level considerations about how such a machine could be exploited by the company and the impact of human-in-the-loop data preparation on the dataset itself. Finally, we present an approach to deal with high-dimensional Pareto-distributed categorical data. This chapter aims to answer the research question **RQ-1**, **RQ-2**, and **RQ-3**.

Chapter 3 In this chapter, we aim to contribute to the discussion about data-driven vs. knowledge-driven science, by presenting a case study about the evaluation of the potential relationship between air pollution and CoVid-19 infections. We combine statistical approaches and machine learning algorithms to confirm or refute such a hypothesis of correlation. The goal of the chapter is to provide an answer to the research question **RQ-4**.

Chapter 4 This chapter focuses on knowledge codification and transfer, with the aim of understanding the role of human-machine interaction. The experiments were conducted in collaboration with a company that deals with determining an underwater path for the installation of cables. We designed different machine learning algorithms, modelling different training strategies based on organizational learning theories and evaluating the various implications. This chapter aims to answer the research question **RQ-5**.

Chapter 5 This chapter ends the thesis by drawing conclusions and paves the way for future contributions.

Chapter 2

Imperfect machines and the role of humans in data science processes

> Can an imperfect deep learning algorithm still be useful?
> How much is the impact of human-in-the-loop approaches in data preparation?
> How can high-dimensional Pareto-distributed categorical data be handled to train a deep learning algorithm?
>
> **— RQ-1, RQ-2, RQ-3**

In this chapter, starting from the design of a deep learning algorithm to predict water meter failures using historical data provided by a company that supplies water in Northern Italy, we present several high-level reflections. First, we discuss how such an imperfect algorithm could be exploited by the company and integrated into their processes. Then, we reflect on the role of humans in data science processes, evaluating the impact of human-in-the-loop data preparation on the dataset itself. Finally, we also present an approach to deal with high-dimensional Pareto-distributed categorical data.

2.1 Introduction

If modern Artificial Intelligence (AI) comes often misunderstood, this is mainly due to the fact that, historically, it is solely tied to the way human brains work and think. Machine Learning (ML) algorithms, instead, learn by processing massive piles of data. This process enables machines to adapt to real-world situations, as well as to propose suggestions on how to classify and interpret a variety of different real phenomena. Simply speaking, the deployment of modern ML systems into critical applications is directly influenced by the way training data are organized and modeled [115, 73]. Hence, while those modern algorithms rapidly sift through huge datasets, loaded with millions of information, a thoughtfully designed AI, beyond its ML-based core, should never disregard the fact that algorithms that learn are, for now, just another form of machine instructions, still guided and influenced by the potential and the limitations that training data carry with them. In other words, even when we train algorithms to learn basic associations that can then be used to approximate, or infer, some aspects of a given process, crucial remains the process of harnessing those piles of data into realistic findings. No matter how much sophisticated is the algorithm that will analyze a dataset, equally critical is the statistical validity, the sense, the references, the subtle implications, in one simple word: the semantics, being inherent in those data [158, 46].

This was exactly the case of our controversial experience with a huge real-world dataset, fed with over fifteen million water meter readings, supplied by a company that distributes water over a large area in Northern Italy. In this case study, we were asked to design an ML-based intelligent classifier, able to predict if a water meter fails/needs disassembly, based on a history of water consumption measurements, thus minimizing the number of technical interventions performed by human operators for maintenance and repair. In this chapter, we present several considerations drawn from this experience.

First, our initial attempts to train a recurrent neural network, without specific attention to the quality, and to the limitations, of those data used for training, led to unexpected and negative prediction outcomes. We report on the human-in-

the-loop approach we employed, in terms of statistical tests and semantics of validity, to extrapolate from that large initial database just those training data that could make a sense, as well as that could safely represent the complex statistical phenomenon under observation, with the final target of training a machine able to predict a failure of a water meter, not only in a dataset but also in a real practical case. As a result of these data modeling and re-organization activities, and upon completion of the training process on a safe subset of the initial dataset, our classifier upheld its performance level, from approx. 60% to about 80-90%, in terms of prediction accuracy. Nonetheless, this performance outcome came with the paradox of a statistical transformation of the initial dataset, thus confirming one of our research conjecture in this field: the need for millions of training data can become a non-issue, as compared to a paltrier training set that makes, instead, a learning algorithm much more realistically applicable.

Then, we present our reflections on how such a model could be integrated and employed by the company, even if it is not perfect. In fact, machine learning algorithms do not have to be an all-or-nothing phenomenon [27]. In our vision, the question is not whether or not to rely on an imperfect machine, rather it is deciding how much autonomy we are willing to give to them since there is a wide spectrum of automation from fully normal to fully automatic [121]. In particular, we propose different strategies to exploit the developed algorithm, determining the boundaries of its use.

Thirdly, we also present an approach to deal with high-dimensional Pareto-distributed categorical data. Such an approach is based on the idea of exploiting the categorical values to better select the sets of devices on which the model goes trained. We evaluated the proposed approach and compared the results with more traditional space dimension reduction methodologies and classical machine learning algorithms.

The remainder of this chapter is structured as follows. In the next section, we present the research background on which our study relies on. Section 2.3 lists the research questions while Section 2.4 discusses the dataset and our approach. In Section 2.5, we presents the data preparation pipeline, following a human-

in-the-loop approach. Then, Section 2.6 illustrates the accuracy of the results we have obtained after training a deep learning algorithm that predicts water meter failures and our experience in replicating our approach on a new set of data. Finally, Section 2.7 presents some high-level reflections and provides the answers to the research questions.

2.2 Background and Related Work

This Section is split into three different parts. The first one discusses other studies done in the specific domain of automatic methods for detecting faulty water meters. The second one, instead, illustrates the negative effects we can incur due to a lack of attention in data preparation while instructing machine learning algorithms. The third and final part, instead, discusses dimensionality reduction techniques for categorical data.

2.2.1 Detecting water meter failures

While we are plenty of papers in the literature that employ complex statistical methods, or machine learning algorithms, for individuating anomalies like a leakage or a failure, in water distribution pipelines [142, 116], there is a not surprising scarcity of papers that discuss methods for detecting anomalies in water meters. Pour cause: so far, in fact, smart metering has come into the scene for utilities different from water, like energy and gas, since these latter resources are considered more expensive, in general.

This motivates the fact why there are a lot of mechanical water meters around, whose main characteristic, different from electrical meters, is that of providing fewer and less frequent readings over time. Hence, even if the number of mechanical meters installed is still high, representing a cheap and well-tested solution, they pose a problem to all the initiatives that are based on machine learning [6]. Indeed, their readings are rare (2/3/4 times per year) and are to be read by a human operator, thus resulting in many imprecisions. Due to

this fact, some of the most relevant papers that illustrate methods that face the problem of detecting faulty water meters, on the basis of an analysis of the amount of consumed water, still resort to traditional approaches, disregarding machine learning. For example, Roberts and Monk developed a simple algorithm that individuates possible anomalies [127], occurring at a given water meter, when a decreasing trend in water consumption is observed along with a series of readings which is updated just quarterly. Monedero et al. [102], instead, propose an approach to detect tampering activities in mechanical water meters that employs a very basic statistical analysis for identifying:

- either a low rate in water consumption,
- or a sudden stoppage of that consumption,
- or simply a decreasing consumption trend.

What is relevant, here again, is the fact that the use of data (readings), which can be, large in quantity yet rare in frequency, does not allow for the use of modern machine learning-based methods.

For sure, the advent of electrical water meters, along with telemetry that can provide water consumption readings on a per hour basis, could significantly alter this picture.

In this unfortunate scenario, our challenge has been precisely that of trying to use learning algorithms, even if trained with data coming from readings of traditional mechanical meters, believing that a large amount of data available, spanning over multiple years and involving more than one million meters, could balance the negative effects of the low frequency in reading.

2.2.2 Inadequacy of the datasets

In this Subsection, instead, we are concerned with the problem that, while machine learning techniques analyze datasets that are very large and expensive, it is often the case when results come up that are inaccurate, or even wrong. Oversimplifying a complex scenario, there are exaggerated anxiety and worry

of developing specific learning algorithms that search through a vast amount of data until they recognize a pattern that finally exists only in (a portion of) that dataset, and not in the reality [5].

As a consequence of these considerations, if we want to discover results that stand the test of time, adequate attention is to be devoted to all the data preparation, cleaning, and transformation activities that must follow their initial acquisition. Disregarding, or simply underestimating, these factors mean crystallizing data inconsistencies and impurities into a shapeless structure that will be inadequate for supporting correct evidence-based decisions.

Drawing upon scientific literature, among all the possible cases we could cite in support of our ideas, we report here just three different examples, where a lack of attention on data used to instruct intelligent machines resulted in negative consequences, as well as into effects diametrically opposed to what we would expect.

The first example is the paradigmatic case discussed by Buolamwini and Gebru in [23]. After a careful assessment, a gender classification system, based on a facial analysis dataset, came up not to be balanced with respect to gender and skin type. It was found out in fact that the most misclassified group was that of the darker-skinned females, with misclassification rates ranging from 20.8% to 34.7%, while, instead, the error rate for the lighter-skinned males' group stabilized at around 0.8%. Such a result was the direct consequence of the dataset that was used for training that system. Ex-post accurate analysis of the dataset simply revealed that it was biased, as overwhelmingly comprised of lighter-skinned subjects.

Another similar example is reported by Bolukbasi and coauthors in [15] that treats word representations. Specifically, the term word embedding intends a representation of words under the form of vectors, commonly used for natural language processing. The popularity of this kind of word representation comes from its ability to capture semantic relationships among words, measurable as linear distances between vectors [114]. An experiment was conducted that tried to train an ML-based implementation of a word embedding representation, only

using articles taken from the Google News service. To simplify this complex matter, we only say that the main result of this specific implementation of a word embedding representation was to let emerge a dramatic case of sex stereotype. For example, the role played by a surgeon was always associated with a masculine subject, while at the opposite the functions of a nurse were always perceived as carried out by a feminine subject. The motivation for the emergence of this gender stereotype was rooted in that specific dataset, as an expensive assessment activity demonstrated. Further, an additional specific procedure was developed that de-biased that representation to the point it finally became gender-neutral.

A final example is drawn from a medical context where a case is reported discussing on a machine trained to learn a prognostic model used to predict adequate medical treatments for patients affected by pneumonia [24]. Surprisingly, upon completion of the training phase, the machine had learnt that patients suffering from both pneumonia and asthma were to be considered at a lower risk of death if compared with those who were afflicted by just pneumonia. The motivation why asthma was (erroneously) considered by the machine (almost) as a protective factor against the negative effects of pneumonia was pretty clear after an ex-post analysis of the dataset on which the machine was instructed. The reason goes as follows. Patients diagnosed with pneumonia, and with a history of asthma, are typically admitted to more intensive care, leading on average to more rapid healing, with respect to patients diagnosed with just pneumonia. Unfortunately, the training dataset was constructed disregarding that relevant fact, with the final paradox of a machine that misinterpreted asthma as a protective variable in that specific domain.

2.2.3 Dimensionality Reduction Techniques for Categorical Data

Data can be divided into numerical and categorical. While numerical data are measurable in nature and easily manageable, categorical data, instead, represents

a collection of information, that can be divided into groups; e.g., black and white; and, as such, they can take on numerical values (for example 1 indicating black, and 2 indicating white), but those numbers do not have a precise mathematical meaning. This is the reason why working with categorical descriptors can easily lead to an increase in the dimensions of the space under investigation. This phenomenon may go so fast that the available data become sparse, with a consequent loss of statistical significance [151].

To understand this phenomenon, take, for example, one of the most common techniques applied to encode categories into numerical values: the one-hot encoding technique [110, 28]. Consider a categorical variable with the following values: Yes, No, and Prefer not to say. They can be encoded with the following vectors [1, 0, 0], [0, 1, 0], [0, 0, 1]. This produces a new, three-dimensional space, with a total amount of twenty-seven points. However, the only interesting points remain three and are orthogonal, equidistant, and sparse. Simply said, we have yielded a three-dimensional vector space, with a new dimension for each original value (yes, no, prefer not to say). Unfortunately, things can even get worse. If we had three categories, each with three values, we would get a nine-dimension space, as this would come with the product of the number of categories times the number of possible values [4, 71].

Hence, a general problem can be posed: how to manage categorical variables, while keeping the dimensionality of the resulting space under control. To this aim, many statistical techniques have been proposed in the literature to face this problem. Typically, the recurrent idea behind all those methods is as follows:

1. Consider a high dimensional categorical space.
2. Apply a procedure for reducing the number of variables, without loss of information.
3. Identify new variables with greater meaning
4. Keep as the ultimate target that of maintaining visible a lot of points, in this reduced space, to be used as representative examples on which a supervised learning model can be trained.

What is also very common is the fact that the procedure for reducing the dimensionality rests upon the idea of representing the categorical space with a few orthogonal (uncorrelated) variables that capture most of its [136]. In the remainder of this Section, we are going to provide a few details on the principal techniques of this family. Before beginning with this review, we briefly anticipate here that our method will be different. We will avoid to use categorical descriptors as input to the model to be trained. Instead, they will be used as a driver for data selection, thus eliminating, from the start, the need for a dimensionality reduction of the categorical space.

Among the traditional methods mentioned above, probably, the Correspondence Analysis (with all its variants) is the most known one. Akin to the Principal Component Analysis, the Correspondence Analysis (or CA) provides a solution for projecting a set of data onto lower-dimensional plots. Essentially, CA aims at visualizing the rows and the columns of a contingency table as points in a low-dimensional space, so that a global view of the data is made available, yet easily interpretable [99, 97]. Identically derived from the Principal Component Analysis, we have the CATegorical Principal Components Analysis (or CATPCA). Here, again, the final goal is to reduce the data dimensions by projecting them onto a low-dimensional plane, with the plus that the relationships among observed variables are not assumed to be linear [131].

Of interest in this field, it is also the so-called Multi-Dimensional Scaling (MDS) technique. Technically speaking, MDS is used to translate information about the pairwise distances among a set of n objects into a configuration of n points, mapped into an abstract Cartesian space. In essence, this technique is proven to be useful to display the information contained in a distance matrix, while providing a form of non-linear dimensionality reduction [162, 130]. Sometimes, also some kind of structural equation modeling is employed to individuate groups, or subtypes, in the case of multivariate categorical data. These are called latent classes, as detailed in the following references [49, 81, 166, 156].

Another interesting technique, in the context of multivariate statistics, is that of Binning. Here the target is somewhat different, since, at the basis, we

have a form of data quantization. Essentially, all the data values falling into a given interval (the bin, indeed) are all replaced by a single representative value. A typical example, which is provided to explain this technique, is that of representing the ages of a group of people with intervals of consecutive years, rather than with each single age value [109]. Needless to say, going for binning is a delicate choice, since some pieces of information can come sacrificed. Nonetheless, it may result in a valid option when dealing with categorical variables, because a large amount of less frequent values, which could increase the dimensions of the resulting space, can be instead all grouped under a unique generic value (e.g., Other). This way, we yield just one dimension for an entire group of categorical values.

2.3 Research Questions

Considering the different aspects analyzed in the previous Sections, this chapter aims to answer the following research questions:

RQ-1 *Can an imperfect deep learning algorithm still be useful?*

RQ-2 *How much is the impact of human-in-the-loop approaches in data preparation?*

RQ-3 *How can high-dimensional Pareto-distributed categorical data be handled to train a deep learning algorithm?*

2.4 Methods

In this Section, we detail the characteristics of the dataset provided to us, we describe the data preparation activities, and the machine learning algorithm.

2.4.1 Dataset Description

We were, initially, provided with a huge dataset comprised of almost fifteen million water meter readings, plus other contextual information. This large dataset spanned a period in time, from the beginning of 2014 to the end of 2018. All these measurements involve more than one million water meters, including those affected by faults, and hence subjected to disassembly and subsequent replacement activities.

Our dataset has fourteen attributes for each water meter reading, as described in Table 2.1. Instead, water meters are characterized by seventeen attributes, reported in Table 2.2. It is obvious that negative examples should be faulty meters (with their corresponding readings) and positive examples non-faulty meters (with their corresponding readings). Such information is reported in the attribute Operation (Faulty/Non Faulty) of the water meter dataset, which essentially indicates if the water meter has been either disassembled or not.

No	Attribute name	No	Attribute name
1	Water Meter ID	8	Reader ID
2	Reading ID	9	Type of Contract
3	Reading Value	10	Reading Validity
4	Reading Value Date	11	Certification on the ERP
5	Previous Reading Value	12	Final Billing
6	Previous Reading Date	13	Reason for Reading
7	Reading Frequency	14	Accessibility

Table 2.1 Reading Dataset Attributes

2.4.2 Data Preparation

Data preparation activities usually consist of normalizing data employing the standard scaling or the min-max one and transforming categorical data into numerical ones, exploiting for example the One Hot Encoding method. Instead, in this case study, the data preparation phase is very elaborate as the dataset provided was directly extracted from the company ERP. Hence, it not only

No	Attribute name	No	Attribute name
1	Water Meter ID	10	Installation Date
2	Serial Number of the Producer	11	Plant
3	Producer Description	12	Type of Contract
4	Material ID	13	Geographical Zone
5	Material Description	14	Accessibility
6	Max/min Reading Value	15	Use Category
7	Meter Type ID	16	Address
8	Meter Type Description	17	Operation (Faulty/Non Faulty)
9	Year of Construction		

Table 2.2 Water Meter Dataset Attributes

describes the phenomenon under investigation (i.e., the consumption of water) but it presents numerous inconsistencies and impurities, whose causes trace down to the fact that it contains also data generated by different business processes that happen in the company. For these reasons, we choose to adopt a Human-in-the-loop approach to the data preparation [42]. In fact, human involvement is instrumental in many stages of data cleaning and preparation such as providing rules or validating computed repairs [126]. Our approach to data preparation based on a human-in-the-loop is described at length in Section 2.5.

2.4.3 Machine Learning algorithm

We employ a deep neural network, specifically designed for our case study. Since there are series of values (i.e., readings of the water meters), we decide to exploit recurrent neural layers. The whole architecture is depicted in Figure 2.1.

Our deep neural network is comprised of two parallel subnetworks. Consider the first one. Its aim is to learn a series of consecutive readings. Hence, it presents a Gated Recurrent Unit (GRU) for each reading in the series. The output of each GRU is passed to a Dense layer of 32 fully-connected neurons. The overfitting phenomenon is avoided by using an in-between Dropout layer, with a keep probability of 0.9, that separates GRUs from the Dense layer.

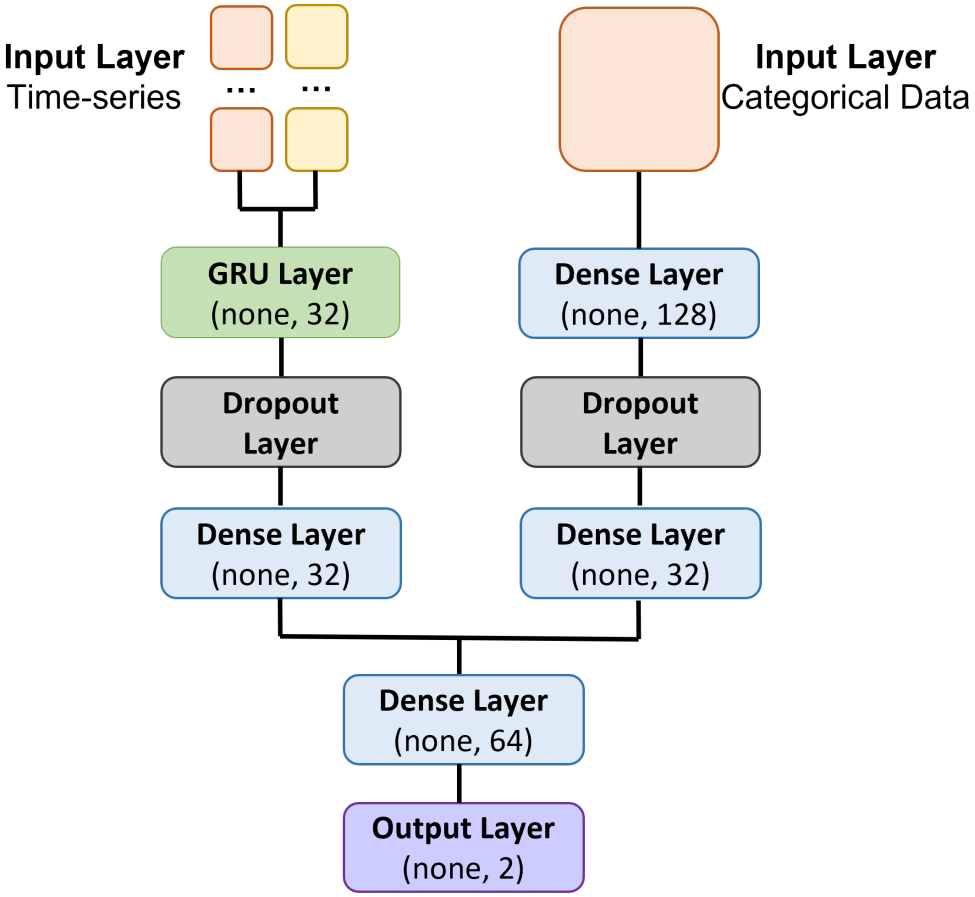


Fig. 2.1 Structure of the deep learning model

Take now the second subnet. It takes as input the one-hot encoded categorical features we have selected and lets them pass through two Dense layers of fully interconnected neurons. The first layer has 128 neurons, while the second one has just 32 neurons. Again, an in-between Dropout layer, with a keep probability of 0.9, separates the two layers of neurons to avoid the overfitting phenomenon.

At this stage, the two parallel outputs of the two subnets are concatenated to form a 64-dimensional vector that passes through a further Dense layer comprised of 64 neurons. The output of this step is a two-dimensional vector (faulty/non

faulty) which is finally delivered to a Softmax activation function that yields the final probability of being faulty or not.

It is worth to notice also that each Dense layer uses a Rectified Linear Unit (ReLU) as the activation function, while we employed a Binary Cross-Entropy function to manage losses, based on the consideration that we had to construct a binary classifier. To conclude the description of our network, we add that, in each experiment, it was trained using the Gradient Descendent Algorithm for twenty epochs, to yield the final optimization.

We implement the deep neural network with the Keras framework [76] and using Tensorflow as the backend [1].

In the experiments, after splitting the data into training and test set, we used the 10-fold cross-validation technique on the training data. In case of class imbalance problems, we take advantage of SMOTE-NC technique [30, 87], to generate synthetic data to oversample a minority target class (in our case the faulty water meters).

To conclude, we come to the evaluation metric we have adopted to measure the accuracy of our classifier. We have chosen the classic Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) [148].

2.5 Human-in-the-loop Data Preparation

As already anticipated in the previous Subsection, we follow a human-in-the-loop approach to prepare the dataset for the training activities. In order to define the semantics of data validity and extract only meaningful data, we took advantage of domain experts, i.e., employees of the company involved in the processes at different levels.

The first step towards a semantics of data validity is taken considering the reading attribute #10 (namely Reading Validity). It corresponds to the case when a human operator reads a value on a water meter and validates it as correct. In the absence of such a positive validation, that reading is to be considered as non-valid, and should not be taken into consideration. Table 2.3 reports the

number of non-valid measurements with respect to the total amount of circa fifteen million readings.

Attribute #10	**# of Readings**
Valid	13,231,251
Non-valid	1,898,128
Total	15,129,379

Table 2.3 Readings: Valid/Non-valid (attribute #10)

Then, we focus on attributes #11 (Certification on the ERP) and #12 (Final Billing). Once selected only valid readings, there are 45 different combinations of attributes #11 and #12. Anyway, just seven of them cover almost 99% of the total amount of readings in the dataset, as shown in the first seven lines in Table 2.4. According to the domain experts, readings to be considered fully valid have to be correctly recorded onto the company ERP system (#11 = 2) and have to be correctly billed to the final client (#12 = 2). Hence, after this step, the total amount of readings decrease from 13,231,251 to 11,856,582.

Attributes		**# of Readings**
#11	**#12**	
2	2	11,856,582
3	2	407,592
2	4	282,527
2	6	132,409
2	5	110,363
2	3	106,742
3	5	105,957
Other		229,079
Total		13,231,251

Table 2.4 Readings: main categories for attributes #11 and #12 (with relative amount of readings)

The third step in the definition of the semantics of validity consists in selecting only real measurements taken on the field by reading a water meter. In

fact, among the readings of the previous steps, many of them are mathematical re-adjustments of estimated values of presumed water consumption values (computed for billing purposes). The balance between real measurements vs. re-adjustments is reported in Table 2.5. Obviously, such re-adjustments bring much noise and are to be removed consequently.

Valid Readings	**# of Readings**
Real	8,185,163 (69%)
Adjustments	3,671,419 (31%)
Total	11,856,582

Table 2.5 Proportion of real measurements vs. adjustments of valid readings

Such a set of conditions is the results of such an iterative process which saw the involvement of domain experts in the loop and we will refer to it as the **semantics of validity**.

We then move on to select the features to be used as input of our deep neural network. This is a key task since irrelevant or redundant features can significantly impact the training activities [59, 92]. In our case, nonetheless, this thing goes smooth as reading values (current and previous), and relative dates, compose the minimal set of information from which learning algorithms can extract interesting relationships (attributes #3, #4, #5, #6 of the readings dataset). Further, on the basis of precise suggestions provided by the company, we also include the following additional features from the meter datasets: producer (attribute #2), material (attribute #4), meter type (attribute #7), year of construction (attribute #9), and use category (attribute #15). Summarizing, Table 2.6 reports all the aforementioned selected features.

A preliminary experiment with the data enjoying the proposed semantics of validity lead us to obtain an average AUC in the ten-fold cross-validation equal to 0.61. Such a negative result denotes a problem in the semantics of data validity that needs to be fixed.

Our intuition is that semantics disregards the role played by time. In essence, of great importance is the time interval between two consecutive valid readings,

No	Features
1	Reading Value
2	Previous Reading Value
3	Reading Date
4	Previous Reading Date
5	Serial Number of the Producer
6	Material ID
7	Meter Type ID
8	Year of Construction
9	Use Category

Table 2.6 Features used

taken on a given meter. In fact, to use those data to train a neural network, crucial is the regularity of the frequency with which readings, respecting the semantics of validity, are taken over time.

Figure 2.2 below provides insightful information with this regard. On the x-axis, we plot the differences of two subsequent readings, values in terms of cubic meters of consumed water, that respect the semantic of validity, while on the y-axis there are the time intervals (measured in days) between two subsequent valid readings.

It summarizes some millions of reading values taken over a lot of time, and it has to be interpreted as follows. Points, that lie on the y-axis and are very far from zero, correspond to measurements that are not taken without any regularity (while, instead, Italian laws prescribe two/three real readings per year). Points, that lie on the x-axis and are very far from zero, account instead for phenomena where the consumption of water is exaggeratedly high.

For this reason, always in collaboration with the domain experts, we add further conditions to the set of rules of the initial semantics of validity. It is noteworthy that all previous conditions still have to be satisfied: a human operator reads a value on a water meter and validates it as correct, the reading is correctly recorded onto the company ERP, it has been correctly billed to the final client, and it is a real measurement. The further requirement is that two consecutive

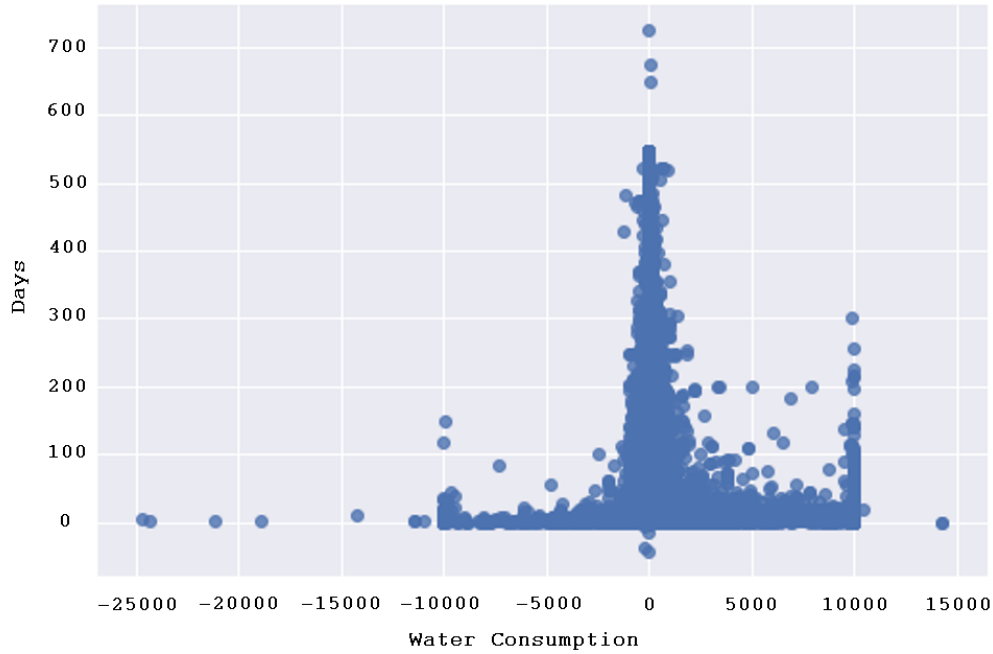


Fig. 2.2 Time intervals vs differential water consumption (two consecutive readings)

readings have to be taken not too far from the others, from a temporal viewpoint. In particular, we set such a time interval to six months. Readings taken more distant in time are not considered valid. We will refer to such semantics of validity as enhanced semantics of validity. After applying it, the overall number of readings falls down to less than two million.

2.6 Results

This Section goes through two different phases. First, we present the accuracy results of the prediction we have obtained using training data enjoying the enhanced semantics of validity, and then we provided a comparative analysis of the results we have obtained with our deep neural network contrasted against those that can be obtained with alternative, more traditional methods. In the latter

phase, instead, we present our experience when experimenting on a different dataset, containing Pareto-distributed categorical data. At the end of such an experience, we propose our approach to handle this type of data.

2.6.1 Predicting Water Meter Failures

We carry out our machine training activities with meters (faulty and non-faulty) whose readings enjoyed the enhanced semantics of validity, always taken in the period beginning 2014 – mid 2018. We assemble a set of positive examples comprised of some 45,000 non-faulty meters, randomly sampled from the million available ones. Along with them, we select all the faulty meters available, 15,000 ones, always with readings respecting the enhanced semantics of validity. Water meters are randomly divided into training and test sets in a stratified way, with the aim of preserving the same percentage of non-faulty and faulty water meters in both sets.

As already mentioned, we employ the ten-fold cross-validation on the training set, using the deep neural network presented in the previous Section. We experimented with series of readings of different lengths, to take advantage of the memory of the network, ranging from two to five readings.

The results of the ten-fold cross-validation are reported in Figure 2.3. Of particular interest are the average validation results (gray bars) that provide AUC values always over the threshold of 80%, precisely in the range [82-88]%. For the sake of completeness, we also report the average training AUC values (black bars).

These results obtained during the validation phase of our deep neural network training process were well promising, yet we wanted to have final confirmation of the efficacy of the deep neural network trained on data respecting the enhanced semantics of validity. Hence, we test the model on the test set. Just series of either two or three readings are tested, for the sake of simplicity. Figure 2.4 portrays those results. The results confirm the efficacy of our approach, with AUC values of 0.86 for two readings and 0.89 for three readings.

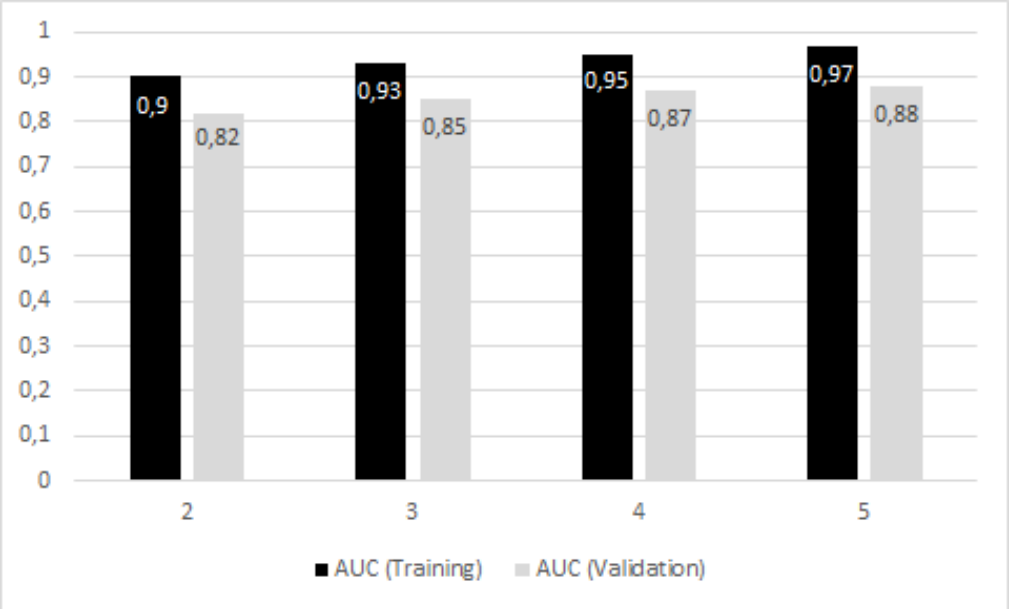


Fig. 2.3 Training and validating with Deep Neural Network: AUC results

Then, we conduct a comparative analysis to contrast the performance of our deep neural network with some of the most common machine learning algorithms that, different from our recurrent deep network, do not use memory. We experimented with all the following traditional machine learning algorithms:

- Linear Regression (LR) [133],
- Lasso (LA) [124],
- Elastic Net (EN) [167],
- Classification and Regression Tree (CART) [31],
- Support Vector Regression (SVR) [117],
- K-nearest neighbors (KNN) [50],
- Adaptive Boosting (AB) [122],
- Gradient Boosting (GB) [74],

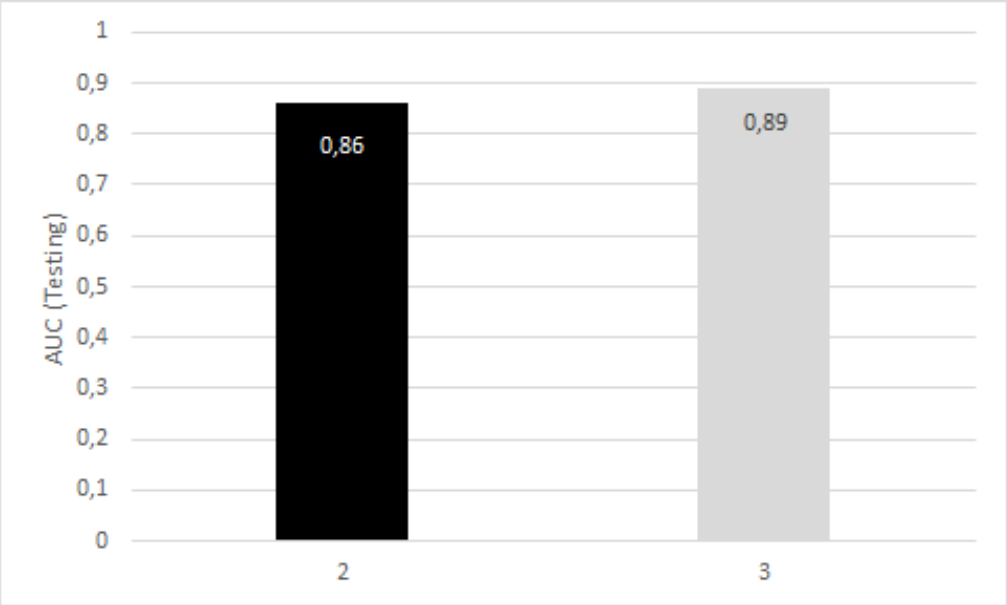


Fig. 2.4 Final testing with Deep Neural Network: AUC results

- Random Forest (RF) [119],
- Multi-Layer Perceptron (MLP), with only one hidden layer with 100 neurons [96].

We use regression algorithms to predict the probability of a water meter being a faulty one. They are implemented using the Scikit-learn library [113].

In Figure 2.5, a plot with the AUC results obtained during the validation of each aforementioned algorithm is reported, contrasted against the result achieved with our deep neural network (DNN) with three readings.

The Figure shows that acceptable average AUC values can be achieved in some specific cases; for example, with the GB and MLP algorithms, that yield almost an 80% value of accuracy. Nonetheless, DNN performs significantly better as it reaches an average accuracy value of 85% with just three readings.

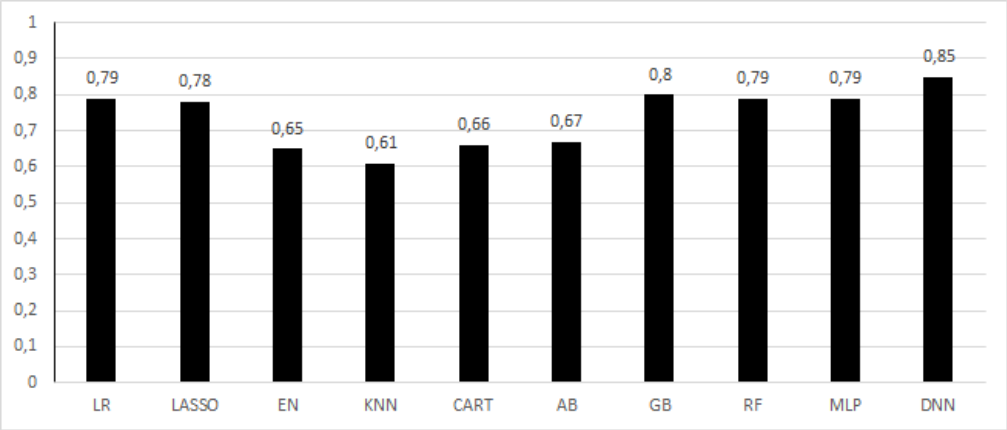


Fig. 2.5 Comparative Results: DNN against all others machine learning algorithms

2.6.2 Experiments with Pareto Distributed Categorical Data

In a further experiment, the company asked us to re-use our methodology on a different set of water meters and relative readings. The dataset included 17,714 devices, where 15,652 were non-defective ones, and the remaining 2,062 were defective. For each water meter, the features provided are the same as Table 2.6, there are three consecutive readings, enjoying the enhanced semantics of validity.

We follow the same methodology and deep learning algorithm described in Section 2.4. However, we observe a drop in the performance of the classifier. The AUC during the cross-validation passed from 85% to 78% while on the testing set it goes down from 89% to 83%.

Analysis of Categorical Data

Our intuition is that the problem lies in categorical data, that are:

- The type of material (attribute #6): it can take on 98 different values. From now on, we will refer to it as the categorical variable A.

- The specific type of the device (attribute #7): it can take on 45 different values. From now on, we will refer to it as the categorical variable B.
- The manufacturer of the meter (attribute #5): it can take on 48 different values. From now on, we will refer to it as the categorical variable C.
- The type of usage of the meter (attribute #9): it can take on 14 different values. From now on, we will refer to it as the categorical variable D.

Traditional Approaches to Dimensionality Reduction

To confirm our intuition and to overcome the problem with these data, we conduct three different experiments. In the first one, we used only numerical values to have proof of our intuition. Then, we reintroduce them employing two different techniques to tackle the problem of dimensionality reduction: the Principal Component Analysis (PCA) [99] and the Binning [109]. With the PCA, we set the sum of variances of all individual principal components approximately equal to 90%, decreasing the learning space dimensionality to 128. With the Binning, instead, it is reduced to 48, selecting the most frequent values and adding an "Other" to group all the other values.

The results of these experiments are reported in Table 2.7: with all the categorical variables (first row), without them (second row), using PCA (third row), and using Binning (fourth row). For each of them, we reported the AUC achieved during both the ten-fold cross-validation and testing (third and fourth columns).

As shown, the results confirm our intuition, since the removal of the categorical data allows us to return on performance similar to the ones obtained with the other dataset (experiment #2). Not only that, but also an attempt to reduce the dimensions of that space, using traditional techniques like PCA and Binning, either provided no real benefit (experiment #4, 85%, Binning) or caused a further deterioration of the prediction accuracy (experiment #3, 81%, PCA).

In essence, we have found a case where there is no way to obtain an improvement of the accuracy of the predictions if categorical data are added, whatever

Exp	Categorical Space Dimension	Validation	Testing	# Water Meters	
				Defective	Non-defective
#1	205	78% ± 8.5%	83%	2,062	15,652
#2	0	85% ± 1.7%	86%	2,062	15,652
#3	128	73% ± 12.5%	81%	2,062	15,652
#4	48	76% ± 13.4%	85%	2,062	15,652

Table 2.7 Prediction accuracy: w/ categorical variables, w/o categorical variables, PCA and Binning

technique is employed. At that precise point, we took the decision to adopt an alternative method.

Dimensionality reduction with a Pareto Analysis

We decide to better enquire on how the values taken by each categorical variable were distributed over our 17,714 water meter devices.

In Figure 2.6, histograms are plotted that begin to reveal an important fact. There are many devices that possess a given characteristic (or take on a specific value) of a certain categorical variable, while many other characteristics/values are scarcely relevant for those devices.

If we better analyze Figure 2.6, we see that we have four plots, each one for each analysed categorical variable. From top to bottom: A, B, C, and finally D. In each plot, one can read the number of devices that possess a certain characteristic (or value) for that categorical variable, using the scale set at the leftmost side of the y-axis of the Figure, while the n different characteristics (or values), for each categorical variable, are distributed over the x-axis. Obviously: n = 98 with A; n = 45 with B; n = 48 with C and n = 14 with D. It goes without saying that the higher is a histogram, the more are the devices possessing that given categorical characteristic.

There is another information portrayed in Figure 2.6: for each categorical variable, we have a dotted curve with the cumulative percentage distribution of those n values over our devices. As an additional note: Figure 2.6 has been

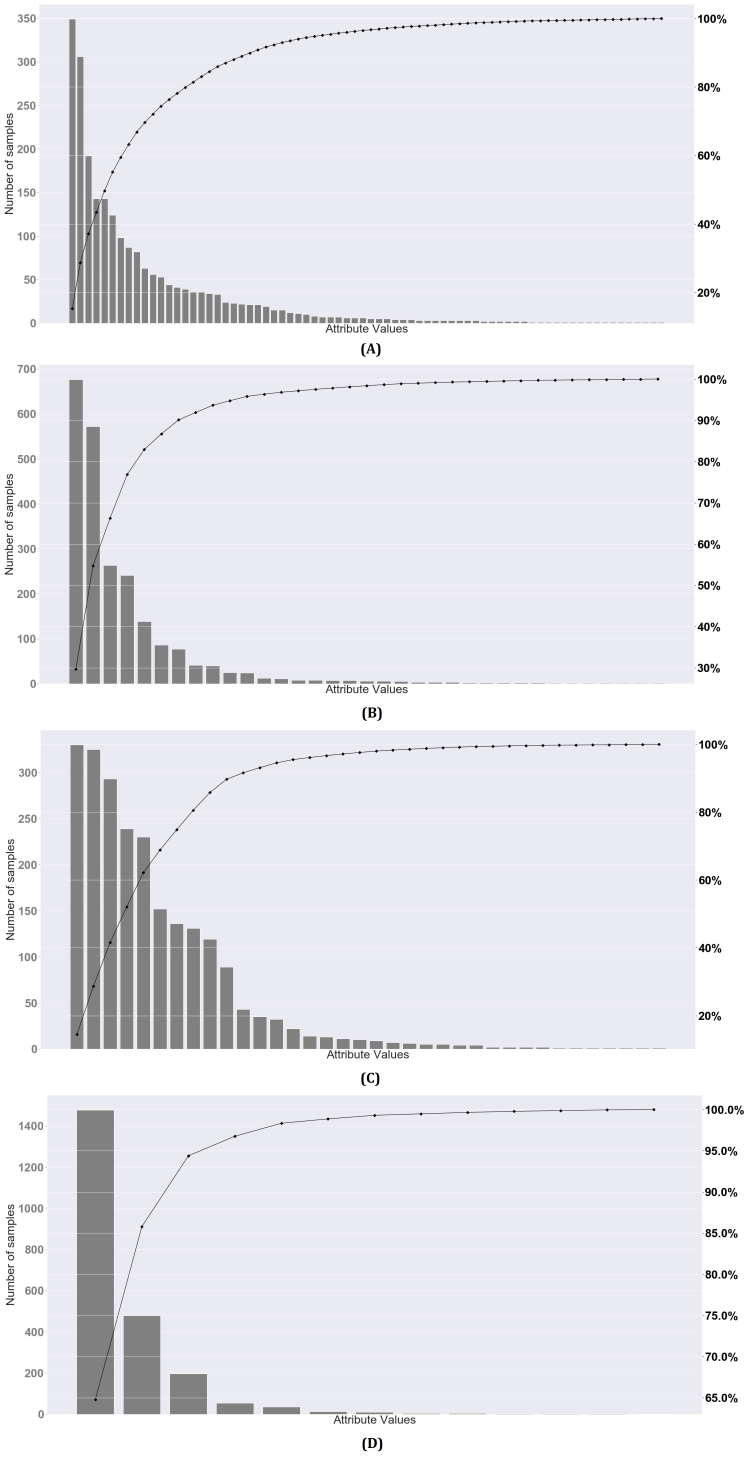


Fig. 2.6 Number of devices possessing a given categorical characteristic (From top to bottom: variables A, B, C, and D)

drawn only for the defective meter devices. For the sake of conciseness, we have omitted to report an additional figure for non-defective devices, as it would show very similar results.

In the end, a careful analysis of Figure 2.6 reveals that the distribution of the categorical characteristics possessed by our 17,714 devices is shaped like a quasi-Pareto function [94].

As to this choice of identifying with a Pareto distribution, the curves according to which the categorical values possessed by our devices is shaped, we could notice that there is a wide hierarchy of several other power-law or Pareto distributions (known, for example, as Pareto type I, II, III, IV, and Feller–Pareto distributions). However, our intent, here, is simply to emphasize that our empirical observation of the curves of Figure 2.6 shows that the typical 80-20 Pareto rule, stating that 80% of outcomes are due to 20% of causes, precisely reflect the situation under investigation. Only the adoption of a quasi-Pareto function fits well the trend of our four categorical variables where just a few of the most frequent values would provide a contribution in terms of knowledge representation of this phenomenon. In other words, what happens is that, given a categorical variable, just a small subset of its characteristics (or values) is possessed by most part of the water meter devices. On the contrary, a lot of values (or characteristics) that a categorical variable can take on are not representative of any device.

Table 2.8 better summarizes this aspect numerically. For each categorical variable, it reports: i) the number n of all the values a given variable can take on, ii) the number of the most frequently used values, and finally, iii) the number of devices (both defective and non-defective) that possess those most frequent characteristics. In essence, what emerges from Table 2.8 is that, on average, around the 90% of meter devices possess about the 20% of values of a given categorical variable.

By virtue of this analysis, we reconsider our approach to manage categorical data, not as input data on which to train the model, but as relevant information to take into account to reshape the training dataset. In essence, the idea at the

Variable	# Values	# Frequent Values	# of Meter Devices		
			Faulty	Non-Faulty	Total
A	98	23 (23%)	1,855 (90%)	13,474 (86%)	15,329
B	45	7 (16%)	1,854 (90%)	13,707 (88%)	15,561
C	48	11 (23%)	1,889 (92%)	13,369 (85%)	15,258
D	14	3 (21%)	1,945 (94%)	13,963 (89%)	15,908

Table 2.8 A quasi-Pareto distribution of the categorical characteristics

basis of our approach is that of employing the most frequent values (of a given categorical variable) to select the water meter devices on which a deep learning model should be trained. At the end of this first step, there are four deep learning models, one for each categorical variable in use. We can call them, respectively: DLM_A, DLM_B, DLM_C, and DLM_D. Obviously, each model can return a different prediction, each one with its associated accuracy. The predictions of the four models can then be merged using bagging [41].

The deep learning models have the same architecture depicted in Figure 2.1, but they use only the time series input and not the one for categorical data.

Results with the proposed approach

The results of the ten-fold cross-validation (with also the computation of the standard deviation values) and on the test sets of the four deep learning models are reported in Table 2.9.

Model	Cross-Validation	Testing
DLM_A	87% ± 1.4%	88%
DLM_B	87% ± 1.4%	87%
DLM_C	86% ± 1.1%	88%
DLM_D	87% ± 1.5%	87%

Table 2.9 Results using our approach to manage Pareto distributed categorical data

We would like to conclude by highlighting the fact that the prediction accuracy returned by our models, whose examples are selected with the Pareto rule, ranges from 87% to 88%. If we compare this result with the AUC value of 83% (obtained with a model trained with both the numerical and categorical variables, experiment #1, Table 2.7), we can observe a not negligible improvement. Nonetheless, this improvement can appear more limited if we look at other alternatives that either do not make use of categorical variables (86%, experiment #2, Table 2.7) or do exploit some dimension reduction procedure, like Binning, for example (85%, experiment #4, Table 2.7). Nonetheless, even in this case, we deem as important to have proposed a new and original method able to increase the prediction accuracy in the presence of categorical variables.

Finally, we have applied a Bagging strategy on that subset of our water meter devices, both defective and non-defective, in the testing set, that possessed the categorical characteristics of all the four models (DLM_A, DLM_B, DLM_C, and DLM_D). This intersection counted 2,304 non-defective devices and 313 defective ones. In this way, we have an improvement for what concerns the AUC which goes from 87%-88% to 90%.

2.7 Discussion and Conclusion

The answers to the research questions presented in Section **??** are presented and discussed in isolation in the following Subsections.

2.7.1 On the usefulness of the deep learning algorithm for predicting water meters failures

Once trained a deep learning algorithm able to predict the failures of water meters with good levels of accuracy, even if not perfect, a question arises spontaneously. How can such a model be effectively exploited even if it is not perfect?

To help with this question, it is important to describe first the current procedure that the company currently exploits to detect faulty meters. It goes as

follows. A software procedure identifies all the meters in which the last two consecutive valid (i.e., respecting the initial semantics of validity) are read with almost a null increment in water consumption. These meters make up the list of candidates to be faulty. At this point, an expensive human-based process starts to individuate, among the candidates, those meters that actually need a replacement. In some cases, controls are performed to make a report of suspected failure, by screening additional databases containing relevant information that could validate or not that suspect. For example, it could be the case when both water and gas are supplied by the same company. In this case, to confirm a suspicion of a failure occurring at a given water meter, the gas meter serving the same client should be recording a non-null gas consumption. Anyway, to be sure of a fault, verification interventions have to be scheduled and performed by human operators, who have to reach the place where the meter is installed and verify its operation. This is obviously unfeasible for all the faulty candidates. Finally, note also that not all those meters that are greatly suspected as faulty are finally changed, due to both operational and business motivations.

It is important, at this point, to talk about a few statistics regarding the company of interest. On average, per year: some 10,000 water meters are considered as faulty candidates, based on the estimates the company makes. Always on average, almost 5,500 are those meters that are greatly suspected as faulty after the execution of the procedures mentioned above, while some 1.500 meters are finally replaced in a year. The reader should put attention to this latter value of 1,500 replaced meters per year, as this is currently the maximum amount of faulty meters that the company can replace, based on its replacement policy.

With this data in mind, we can now discuss how the deep learning algorithm could be employed and integrated into the business processes.

Take for example the results we got with the testing phase conducted with almost 30,000 meters, with readings enjoying enhanced semantics of validity. Out of those 30,000 meters, 6,652 meters were suspected as faulty, while 22,634 were the non-faulty ones, as stated by the company. Just to remind it, our

classifier was able to make predictions in that context with an accuracy of 86%, in terms of the AUC-ROC metric, in the case of two readings.

We then could initially try to use our classifier, set with a decision threshold of 0,46, the one that minimizes the number of both the false negative (faulty meters predicted as non-faulty) and the false positive (non-faulty meters predicted as faulty), as per Figure 2.7. With that threshold, we would obtain a classification of faulty/non-faulty, like that represented in the confusion matrix of Figure 2.8.

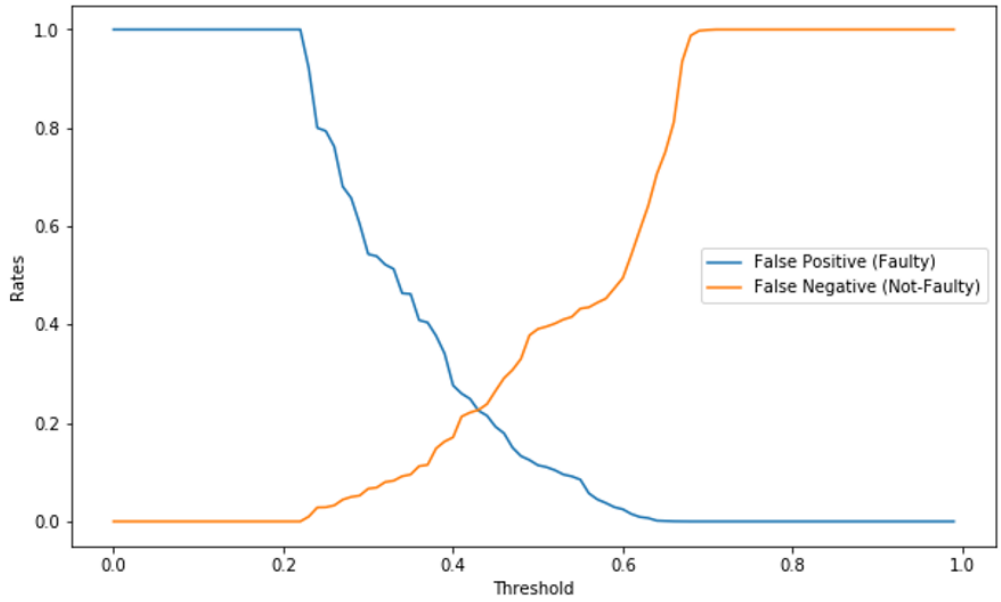


Fig. 2.7 False Positives vs False negatives rates

At this stage, we could propose two alternative operational approaches to replace faulty meters, that combine the results of our classifier with the traditional procedures already in use.

With the first one, we could suggest to the company not to scrutinize all the meters that our classifier has predicted as non-faulty (precisely 20,513 meters, computed as the sum of the top and the bottom quantities on the left side of Figure 2.8), and to concentrate their attention, as well as to deploy their traditional verification procedures, only to those meters that were predicted as

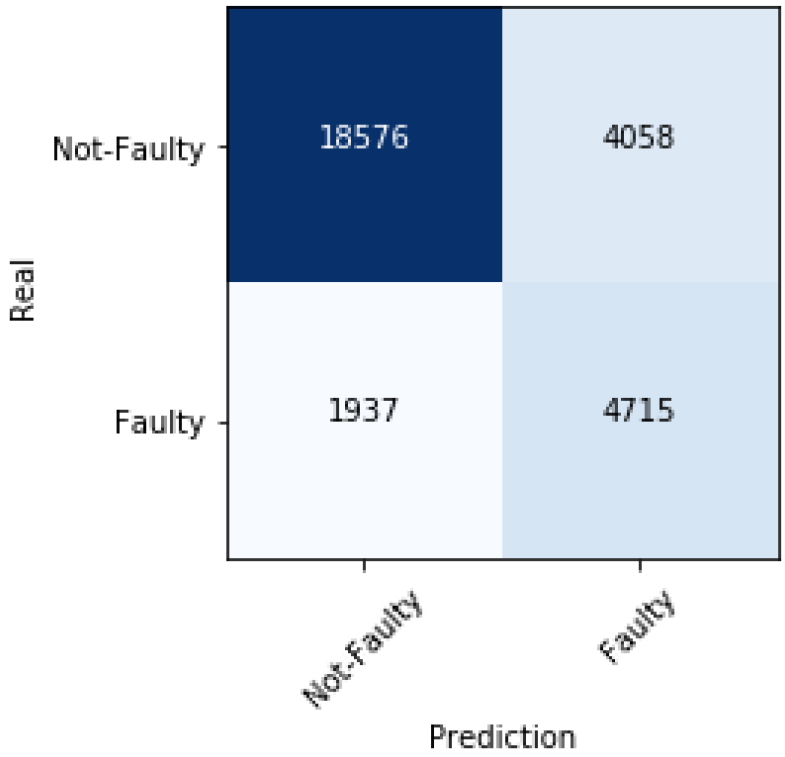


Fig. 2.8 Confusion matrix (decision threshold = 0.46)

faulty (i.e., greatly suspected; precisely 8,773 meters, obtained by summing all the quantities on the right side of Figure 2.8). Unfortunately, this approach does not work well in this case, due to the fact that the company should use its traditional and expensive verification procedures on a number of meters (8,773, for a six-months-long period) that almost doubles the average quantity of meters that are considered as faulty with the methods already in use (5,500, over a period of a year). Not only, with this approach, we know for sure that some 1,937 faulty meters will never be detected (left bottom sector in Figure 2.8).

To solve this latter problem, we could then move the decision threshold towards the direction of minimizing the number of false negatives; for example, setting the decision threshold at the value of 0.3, as per the confusion matrix of Figure 2.9. This would have the effect of decreasing the number of faulty

meters that are never detected down to 443 (left bottom sector in Figure 2.9). Unfortunately, this way, we have further exacerbated the problem of scrutinizing a huge number of meters that are suspected of being faulty, yielding almost 18,509 water meters to be verified with expensive procedures (computed as the sum of the quantities on the right side of Figure 2.9).

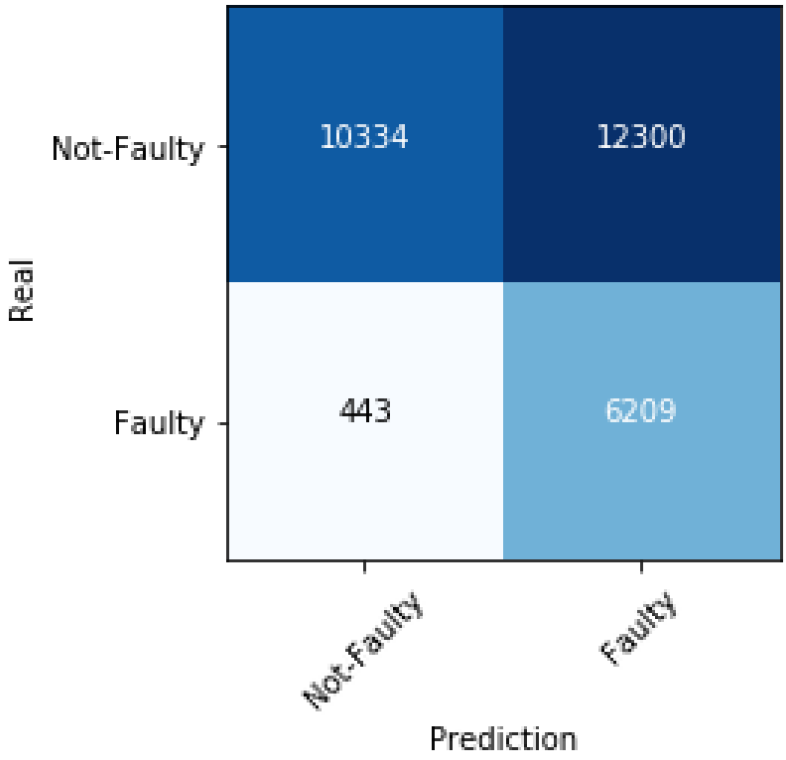


Fig. 2.9 Confusion matrix (decision threshold = 0.30)

Well promising, instead, is the second approach we propose.

The idea is that of minimizing the number of false positives, for example moving the decision threshold to a value of 0.65, as per Figure 2.10. With this approach, our suggestion to the company is to adopt a brand, new procedure to replace faulty meters, which is as follows: focus only on the meters that our classifier has predicted as faulty (1680 meters on the right side of Figure 2.10), and proceed directly with the replacement of all those meters.

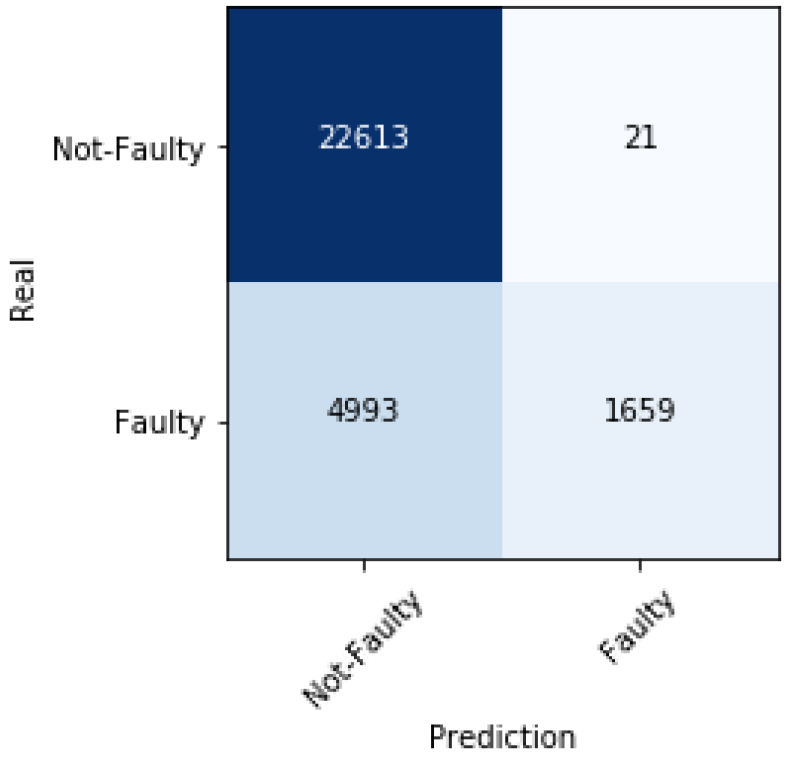


Fig. 2.10 Confusion matrix (decision threshold = 0.65)

Following this approach, the company will have to replace a number of faulty meters that is comparable to the maximum number of those it can replace based on its current replacement policy (1680 vs circa 1500), yet without the need to resort to complex and expensive procedures to individuate them.

Further, in this situation minimized is also the amount of those meters that go replaced even if they did not need any replacement. Indeed, only 21. In other words, in this case, meters have been predicted as faulty and then replaced with a precision of 98,75%. As a final consideration, this discussion demonstrates how savvy use of our intelligent classifier can help the company to detect faulty meters to be replaced without any interference on the business process currently in use [19].

The proposed approach highlights how deep learning algorithms, even if have not perfect prediction capabilities, can be effectively employed by the company. The use of machine learning algorithms does not have to be an all-or-nothing phenomenon in which machines have to be perfect. Such a claim of perfection is probably a cultural projection of the western culture and the results of the Christian apocalyptic thinking [27] in which there is a binary vision: technological apocalypse against techno-utopianism. On the one hand, there is the desire for perfect machines while on the other, a single case of failure is enough to put into question all of them. But when working in the real world, the perfect machine does not (yet) exist. Anyway, according to our vision, the question is not whether or not to rely on an imperfect machine, rather it is deciding how much autonomy we are willing to give to them. In fact, there is a wide spectrum of automation from fully normal to fully automatic performance. Hence, machine learning algorithms can substitute humans (task substitution) until complementing humans (task augmentation) [121]. In this case study, we showed how the deep learning algorithm can support humans, evaluating all those cases in which they are highly confident while humans will analyse the remaining cases, the most doubtful ones, in which the machine tends to make mistakes. In this way, we return to have the same concept of machines that humanity has always had, consisting of taking away the useless work.

2.7.2 Evaluating the Effect of human-in-the-loop Data Preparation

At the end of our experiments, we were able to successfully train a deep neural network that predicts water meters failures with good levels of accuracy, such that to allow its inclusion within company processes, as proposed in the previous Subsection. Then, we reflect on the implications of all the operations we had carried out so far on data. In fact, not all the data available that serves the interests of some specific business process can be considered adequate to instruct an intelligent machine that is intended to implement a new service if this process is

conducted without a deep reflection on the validity, sense, and subtle implications of the training data.

As described in Section 2.5, there has been a high level of interaction between humans (both domain experts and machine learning experts) and data, with the final aim of removing impurities and operational data, to be able to define a dataset in which the phenomenon of water consumption is adequately described. Hence, avoiding the negative consequences of a lack of attention on data.

In particular, we focus on the data relative to water consumption. We consider the consumption of the readings included into 1) the set of readings respecting the semantics of validity, 2) the set of readings enjoying the enhanced semantics of validity, and 3) the set of readings randomly, sampled from the ones of enhanced semantics of validity, used for training.

Take into consideration the plots of Figures 2.11 and 2.12. They both aim to measure the number of readings (y axis) whose average value equals a given value, say X (on the x axis). In Figure 2.11, we have the case of the dataset with the semantics of validity, while in Figure 2.12 we have the enhanced semantics of validity. As seen from a visual comparison of the two plots, we have a clear impression that the shapes of the two curves are not that different, even if the quantity of readings with an amount of consumed water equal to any given value X in the first dataset is larger than the correspondent quantity of readings with the X-Factor.

In order to have a deeper understanding, we subjected these values to statistical tests. Table 2.10 report for each of the three set the total amount of readings, the average (μ), and the standard deviation (σ) values of the consumed water per reading (cubic meters of consumed water).

Id	**Dataset**	**# of Readings**	μ	σ
1	Semantics of Validity	11,856,582	5,307	86,45
2	Enhanced Semantics of Validity	1,973,493	3,674	17,796
3	Sampled for training	135,018	3,647	11,852

Table 2.10 Water consumption statistics

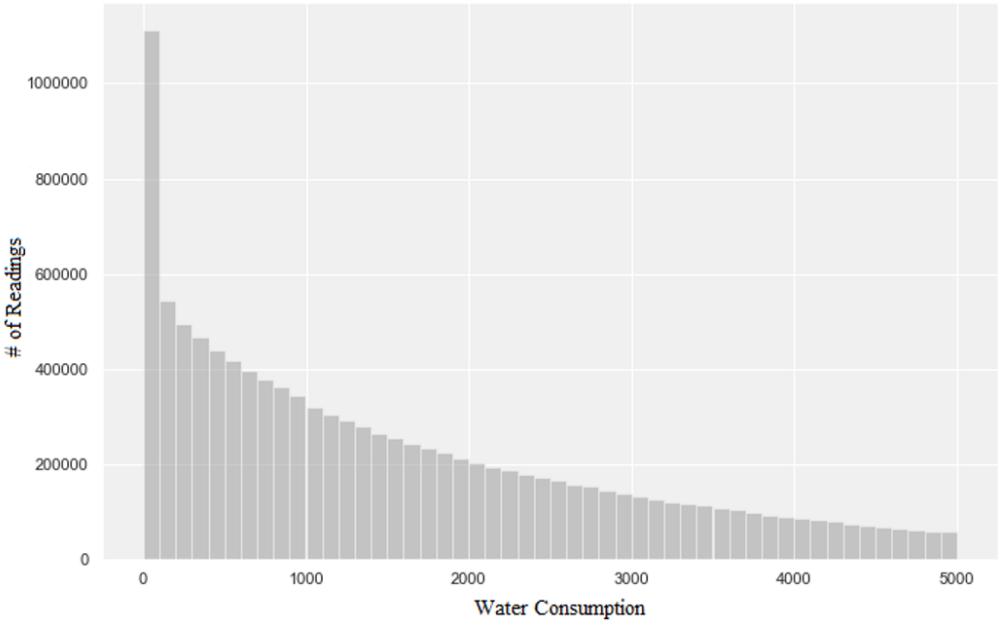


Fig. 2.11 Water consumption values using the semantics of validity

Since data are normally distributed with known values for both average and standard deviation, we use the Z test, whose results are reported in Table 2.11. The null hypothesis is that the average values of consumed water per reading are the same. We tested our null hypotheses with two different significance α values, 0.05 and 0.01. As seen from Table 2.11, the null hypothesis that the average values of consumed water per reading in the initial semantics of validity and in the subset of readings subjected to the enhanced semantics of validity are equal is to be rejected (first line). Instead, it cannot be rejected the null hypothesis that the whole subset of readings with the enhanced semantics of validity has an average value of consumed water per reading equal to the specific subset of those readings specifically used for training.

We can now go back to the initial research question: "*How much is the impact of human-in-the-loop approaches in data preparation?*". Starting from the analysis presented in this subsection, we can state that our human-in-the-loop approach to data preparation had a significant impact on the dataset and, conse-

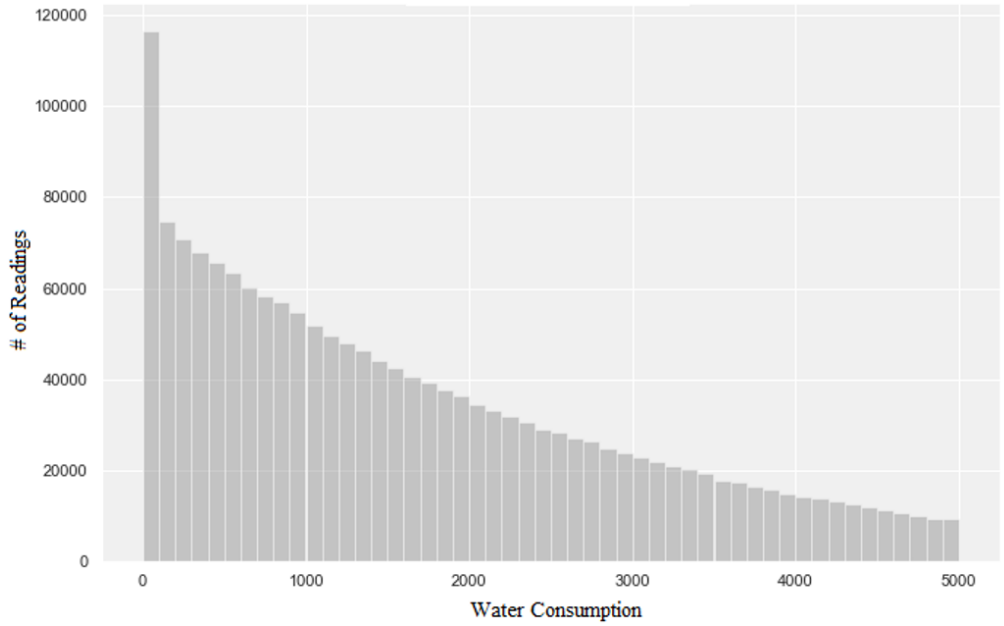


Fig. 2.12 Water consumption values using the enhanced semantics of validity

Test	**p-value**	α **= 0.05**	α **= 0.01**
$\mu_1 = \mu_2$	$< 10^{-5}$	Reject	Reject
$\mu_2 = \mu_3$	0.75	Fail to reject	Fail to reject
$\mu_1 = \mu_3$	$< 10^{-5}$	Reject	Reject

Table 2.11 Z Test - Results

quently, on the performance of the deep learning algorithm. In fact, the presented results highlight how the data preparation phase changed the characteristics of the dataset, from a statistical perspective. Such a result can be taken as a further confirmation that attention has to be paid with respect to the importance and the impact of data preparation from a statistical point of view.

2.7.3 Dimensionality Reduction for Pareto Distributed Data

We have proposed an approach to treat categorical, high dimensional data and we now emphasize both the advantages and the possible limitations of it. Our

approach has proven to be useful in all those cases with categorical variables when it can be shown that the training data are distributed following a (quasi) Pareto statistical distribution. This should not be considered as a limitation, because the field of application may extend very far from the field we have chosen for our study (i.e., the predictive maintenance of water meter devices) up to other research topics where this kind of unbalanced statistical data distributions often occur.

Second, another intriguing issue is that it could seem that, in our training process, we have mixed notions from two different genres (feature selection using the Pareto rule and deep learning). To this aim, we would like to emphasize the fact that while it is true that one of the strong advantages of a deep learning model is its inherent hierarchical feature selection along with the successive level of increasing abstraction in detecting patterns, many practical situations exist where the data have huge dimensions and are also very sparse. In those cases, it becomes difficult to use a pure deep learning approach. In those specific situations, a good practice can be that of using adequate projection algorithms that decrease the number of features to a reasonable number, which can be then effectively tackled by deep learning. When this happens, we should interpret such a procedure more as a feature extraction procedure, rather than a feature selection, which is more typical with classical machine learning algorithms. In simple words, the new features that are extracted are somewhat meaningless from the point of view of the deep learning method, yet their extraction can be useful to drive the learning process, in some specific cases. This has been exactly also our case. Not only, but a new type of literature is emerging that describes similar situations, like for example in [137, 106, 165].

This latest issue has another interesting implication which can be summarized with the question of whether more traditional machine learning classification algorithms, like Support Vector Machines, for example, could perform better with respect to the deep learning models we have utilized selected based on the Pareto rule. To investigate this subject, we have carried out an additional experiment, where more traditional machine learning algorithms were used. We employed

two classical machine learning algorithms like Support Vector Machine (SVM) [35] and Classification and Regression Trees (CART) [85], implemented using the Scikit-learn library [113]. Results from these experiments are shown in Table 2.12. In particular, in the second column of Table 2.12 the AUC values are reported, along with the correspondent standard deviation values, obtained with the ten-fold cross-validation procedure. In the third column, instead, we show the AUC values achieved during the final testing phase.

As it is evident from a comparison of these results with those from Table 2.9, traditional machine learning algorithms have provided, in our case, prediction accuracy performances that are worse than that obtained with the method proposed in this section, thus confirming the validity of our choices.

Model	**Cross-Validation**	**Testing**
SVM w/ Categorical	69% ± 10.4%	80%
SVM w/o Categorical	76% ± 2.6%	77%
CART w/ Categorical	65% ± 6.3%	73%
CART w/o Categorical	74% ± 2.7%	74%

Table 2.12 Comparative results using SVM and CART with and without categorical data

Third, it is important to provide an answer to a more practical question that could emerge at this point of the discussion: How can our method be serviceable if applied to any of the meter devices that are utilized to measure water consumption in our case? In other words, if we have to make a prediction on a new meter device, how can we proceed? The answer to this question relies on the simple application of the following procedure. We should, first, consider that device, and check if it possesses the categorical characteristics of either the variable A or B or C or D. If that device possesses any of the categorical characteristics of interest, we can use the corresponding model (either DLM_A, or DLM_B, or DLM_C, or DLM_D) to make our predictions. Instead, in the negative case, we should not use our models to make a reliable prediction for that device, and we should resort to a more traditional approach. However, it should be noticed that the likelihood that a device does not possess any of those

characteristics, at least in the context of the dataset we have studied, is quite low; i.e., below 10% on average, as our Pareto analysis has demonstrated.

Fourth and final. The approach we have proposed can be seen as a method that can be used in combination with additional techniques, useful to improve its predictive performances, like Bagging. This can be employed whenever there is a device that possesses the characteristics of all the four models together, in the following way. We could use each different model, in isolation, to return its prediction results for that given device, and then we could average over all those returned results, to produce a unique and comprehensive prediction. Obviously, we can expect that this Bagging strategy, at least with our dataset, can work only for a limited quantity of meter devices; yet it could provide finer predictions, whenever applicable.

Our experience can be considered as a further confirmation of the role of humans in the data science processes. In fact, they can provide insights and propose new methods to address possible problems. This is exactly our case, where we were able to devise a method to handle high-dimensional Pareto-distributed categorical data. The novelty of our approach rests upon the idea of exploiting those categorical values to better select the sets of devices on which the model goes trained. Avoiding the use of those categorical characteristics as a direct input to the model has removed the danger of an explosion of the dimensions of the learning space, and with this approach, we have reached predictive accuracies ranging from 87% to 90%, for an amount of 90% of the available devices. We have provided empirical evidence that this approach maintains its validity even if compared with more traditional space dimension reduction methodologies and classical machine learning algorithms.

Chapter 3

On combining statistical approaches ad machine learning to observe a phenomenon

> Can statistical approaches combined with machine learning be used on big data to observe different phenomena?
>
> **— RQ-4**

The big data era has sparked the debate which opposes data-driven to knowledge-driven science. In this chapter, we aim to contribute to such a discussion by presenting our idea that data combined with statistical methods and machine learning algorithms should be used to observe two phenomena, confirming or refuting hypotheses, as has always been the case using the scientific method. In particular, we studied the potential relationship between air pollution and CoVid-19 infections using statistical approaches and machine learning on data relative to the Emilia-Romagna region (Italy) and the New York State (USA). Findings show that such a relationship could exist.

3.1 Introduction

In the Big Data era, different positions are emerging according to which science should be driven by data [77]. This unprecedented availability of data can be exploited with several advanced data analysis techniques such as artificial intelligence approaches, in particular with machine learning ones, to extrapolate theories from the data. Contrary to these positions, scientists are claiming the importance of theory even in presence of such massive amounts of data, stating that "Big data need big theory too" [36]. Following this line of thought, data combined with statistical methods and machine learning algorithms should be used to confirm or refute hypotheses, as has always been the case using the scientific method. This is exactly the approach that we adopted in our case study, where we evaluated the possible relationship between air pollution and CoVid-19 infections.

Although CoVid-19 has originated in Wuhan, China in late 2019, several provinces of northern Italy have soon become among the hardest-hit regions in Europe. This virus outbreak spread with particular intensity to the Italian regions of Lombardy, Veneto, Emilia-Romagna, and Piedmont, in the period from late February to late April, with a severe toll in terms of human deaths. As simple evidence of this disaster, it suffices to remind that the Italian Institute of Statistics (ISTAT) has recently computed for Italy an average increase of 49,4% in the number of all the fatalities that occurred during the month of March 2020, as compared with the number of deaths of March 2019 [69]. Not to mention that, in the same month of March 2020, the official death toll, for some given provinces, like Bergamo and Brescia (in Lombardy), stands at more than five times the value recorded one year before, same period [69]. While it is true that Italy had the bad luck of being the first European country to be devastated by the outbreak, what has gone wrong has motivations in a combination of demographic, political, organizational, industrial, climatic factors, and low Intensive Care Unity (ICU) capacity as well, that need further specific investigations.

With regard to the New York State, the first case of CoVid-19 was recorded in New York City (NYC) on March 1st, 2020. A woman, travelling back from

Iran, was tested positive. This was only the first of a long series of infections, with the pandemic that swept through the whole State of New York, reaching in just one month the considerable amount of 25,665 infected people and more than 200 deaths [55]. We were just at the beginning of a sad story, since in the following three months NYC experienced a widespread diffusion of this contagion, recording over 200,000 infections and more than 21,000 confirmed deaths. After the words of the Governor, Andrew Cuomo: "The apex is higher than we thought and the apex is sooner than we thought", many measures were implemented to contain the spread of this virus, including public school closures on March 15th, and stay-at-home orders (for non-essential workers) on March 22nd [169]. While newly diagnosed infections, hospitalizations, and deaths peaked in April, this lockdown regime led to a substantial drop in May, and to a subsequent re-opening phase for industries, and other business activities, starting on June 7th, 2020 [163]. Since then, NYC has experienced a relatively long period with the number of the new daily CoVid-19 cases that have continued to fall, while they were climbing in the rest of the United States.

All this said many studies have been developed by scholars who have investigated how the CoVid-19 spreads and decays [161, 129]. While of great interest are those scientific investigations that look for the most effective non-pharmaceutical containment countermeasures, that could help to keep a lid on the epidemic (including contact tracing and testing) [40, 128, 125, 84, 3, 61, 79, 62], much attention has been also paid to inquire into the factors that can favour the contagion [38, 143]. In this broad spectrum of research, studies are emerging that have tried to understand if a possible association exists between the exposure to air pollution and CoVid-19 infections and deaths.

Our contribution, here, is to provide a further investigation on the possibility that a causal correlation exists between the two cited phenomena (i.e., air pollution and spread of the infections). We, as investigators, have to admit that we do not possess any prior knowledge of the researched correlation at a level appropriate to the scale of this phenomenon, e.g., biological, chemical, and physical, and we want to limit our study to an examination of the plausibility of the

existence of that correlation at a statistical level. In particular, we are interested in verifying if that correlation comes either confirmed (or rejected) with two different approaches in two different geographical areas: the Emilia-Romagna region (Italy) and the New York state (USA).

First, we subjected those temporal series of data to a Granger causality statistical hypothesis test. Testing for Granger causality means verifying the statistical hypothesis that a time series X Granger-causes another time series Y: in other words, X values would allow to better predict future values of Y, beyond the information contained in past values of Y alone. To do that, two different regression models are, typically, tested on the Y values. The first one uses only previous values of Y, while the second exploits both previous Y values, plus lagged values of X. If those tests are successful, it can be concluded that X values provide a statistically significant predictive information about future values of Y. In simple words, X Granger-causes Y. To this aim, it is worth mentioning that we analysed the daily values of the following air pollutants: $PM_{2.5}$, PM_{10}, and NO_2, treated as time series occurring in a given temporal period that has preceded the series of the CoVid-19 infections, in all the provinces of the Emilia-Romagna region and counties of the New York State. It is also worth mentioning that we know very well the limitations of the Granger-causality tests, that well be adequately treated and discussed [100].

Second, in order to provide further evidence in favor of this correlation, we conducted an additional series of experiments, using some Machine Learning (ML) algorithms. Specifically, four different ML algorithms were exploited in the following (non-traditional) way. At each step of this procedure, they were trained with the data (CoVid-19 infections vs. pollution) relative to all the scrutinized counties or provinces, except for the one for which we asked the algorithms to predict the number of the daily infections, given the concentrations of the pollutant occurred in the previous days. This procedure was repeated for all the counties and provinces of interest, resembling a kind of county/province cross-validation methodology.

Nevertheless, our study does not have to be treated as the final proof of a true causality nexus between the two phenomena, but as an additional clue on a case, that does not deserve to be already archived.

Finally, with our experiments, we want to contribute to the discussion about knowledge-driven as opposed to data-driven science, highlighting the importance of theory even in the big data era. Surely, this massive amount of data can be successfully exploited using artificial intelligence and statistical methods and present unique opportunities, providing further ways to confirm or refute hypotheses. However, theories are important as well as conceptual insights from the humans involved in the experiments.

The remainder of the chapter is structured as follows. Section 3.2 presents some related works. In the next Section, we detail the research questions that drove this study. Section 3.4 describes the methodology behind our approach while Section 3.5 presents the results we yielded. Finally, Section 3.6 concludes the chapter, critically discussing the answer to the research question.

3.2 Background and Related Work

CoVid-19 manifests as a severe respiratory disease, mostly pneumonia. This has led many researchers to focus their attention and study the potential relationship between exposure to particulate pollution and the rapid contagion brought by this virus. With this in view, recently, many international scientific studies were developed to investigate the relationship between particulate of various types and the CoVid-19 incidence.

Exemplar is the work by Jiang, Wu, and Guan that addresses two relevant issues, with reference to the association between particulate and CoVid-19 [72]. They start from the very general consideration that air pollutants raise concerns over their association with infectious diseases, being often the cause of local epidemics [164, 65]. This is typical with influenza since the airborne air pollutants perform as condensation nuclei for the virus to attach, as also confirmed by several other studies [145, 86, 93, 48]. Owing to this consideration,

Jiang, Wu, and Guan proceed with the following reasoning: since CoVid-19 is known to cause human-to-human transmission by infectious secretions [91], these secretions could be transferred in many different ways, including ambient air pollutants. Not only, Jiang, Wu, and Guan also observe that is not by chance that $PM_{2.5}$ is the air pollutant constantly associated with an increased CoVid-19 incidence in all the Chinese cities of their study, namely: Wuhan, Xiaogan, and Huanggang. Besides the fact that particulate could provide condensation nuclei for viral attachment, Jiang, Wu, and Guan add a second biomedical argument which is as follows. It has been discovered that the receptor for CoVid binding is the angiotensin-converting enzyme 2, which concentrates on the type II alveolar cells [168]. Since, type II alveolar cells are located in the alveoli, which are only reachable to particles with diameters less than 5 micrometers, it becomes evident that very small airborne pollutants, such as $PM_{2.5}$, have the potential to penetrate, unfiltered, the respiratory tract, down to the alveolar region [18, 146, 63, 141].

Similarly, interesting results were found also by Pansini and Fornacca who investigated the incidence of CoVid-19 mortality rate in highly polluted areas. They focused their attention on selected areas from different countries (including, among others, China, Italy, and US), and considered also CO and NO_2, in addition to particulates. In particular, they collected data about air quality from two kinds of sources: ground monitoring stations and satellites. According to the analysis they performed, they found significant positive correlations between CoVid-19 infections and air quality variables. Yet, while in China the strongest correlation was found with the (satellite-derived) CO values, in Italy and in the US the highest correlation values, with the incidence of CoVid-19, were those of NO_2, derived respectively from satellite (Italy) and ground measurements (US). One of their final observation is that the CoVid-19 mortality ratio is higher, regardless of the higher number of infections, in all those areas with poor air quality, that is, where values of CO, NO_2 and PM are constantly higher than the acceptable limits [111].

Nevertheless, besides this set of international studies developed in this field (the interested reader can refer also to [160, 108]), we scrutinized, with spe-

cial interest, just those recently conducted by members of the Italian scientific community, for two main reasons. First, the impact of particulate pollution was already being severely felt like a huge health problem in Northern Italy, well before the advent of CoVid-19, and second, those studies have been put at the centre of a heated debate in Italy, and considered not convincing under different perspectives.

To be precise, (almost) all those papers at the centre of this controversy have followed two concurrent lines of reasoning, that are typical when one wishes to infer causal relations from data. On one side, they have tried to acquire (through experimentation) the knowledge of the biological/chemical/physical mechanisms at the basis of the possible correlation between the particulate and the virus spread. On the other side, they have tried to confirm the existence of a true causal relationship between the two aforementioned phenomena, using some kind of statistical hypothesis testing.

The works conducted by Setti et al., for example, provided a quite convincing contribution to this discussion, by both revealing that traces were found of the CoVid-19 RNA in PM_{10} samples in Bergamo [134], and also testing the hypothesis of such a correlation between the daily surplus of that particulate and the consequent contagion between humans by exploiting the statistical model of the coefficient of determination [135]. Daily infections were recorded in the period from February 24th to March 13th, while the surplus of PM_{10} values was considered, on a daily basis, in the period from February, 9th to February 29th.

Conticini, Frediani, and Caro, instead, without any statistical testing activity in support of their hypothesis, argued about the fact that poor air quality can lead to a state of permanent body inflammation and chronic respiratory difficulties, along with a hyper-activation of the immune system; being these all circumstances that make human lungs prone to be attacked by the virus. This is their hypothesis explaining the high mortality rate, recorded in Emilia- Romagna and Lombardy, owing to the virus outbreak [34].

Finally, Becchetti et al. analysed both the PM_{10} and $PM_{2.5}$ values, although recorded on an annual basis, and correlated them to CoVid-19 infections and

mortality, using a cross-sectional regression statistical method. Theirs is a vast study, where scrutinized are also other factors, including temperature, population density, income, number of lung ventilators, and public transport usage. Nonetheless, the conclusion is that air pollution can be considered as a strong predictor for both virus contagions and mortality [11]. In that paper, again, cited as mechanisms at the basis of the correlation between the particulate pollution and the contagion are, respectively, the hypotheses that: i) humans living in highly polluted areas have a reduced respiratory capacity to react to the virus, and ii) the particulate may act as a carrier for the virus.

It is important to remind that all these papers present evidence of the correlation between air pollution and CoVid-19 infections based on statistical analysis. Hence, they should not be treated as final proof, as criticized by some scholars who underlined that all those discoveries boil down to vague clues, completely preliminary, with some of them not yet subjected to peer-review by experts in the field [26].

3.3 Research Questions

Starting from the background analyzed in the previous sections, in this chapter our goal is to respond to the following research questions:

> **RQ-4** *Can statistical approaches combined with machine learning be used on big data to observe different phenomena?*

3.4 Methods

We now present some preliminary information relevant to our study and a description of the data we have used, along with some reflections on the statistical methodology we have employed.

3.4.1 Data Description

Emilia-Romagna

As already anticipated, in this study we are interested in reasoning around the plausibility of a correlation between air pollution and the spread of CoVid-19 infections in the Emilia-Romagna region.

Prior to beginning, it is important to make clear that we have taken into considerations all the provinces of the Emilia-Romagna region, in the period of interest, namely: Bologna, Ferrara, Forlì-Cesena, Modena, Parma, Piacenza, Reggio nell'Emilia, Rimini, and Ravenna. It is worth mentioning that this Italian region is populated by almost 4,500,000 citizens and has been one of the more seriously affected by this virus, with a total number of infections of 26,719, and as many as 3,827 fatalities, as of May 9th, 2020.

Relevant for considering this process from the right temporal perspective is also the chronology according to which restrictions were first imposed on human activities in those provinces and then released after a substantial decay of the virus incidence. In particular, we can count four subsequent phases:

- Prior to 8 March 2020, no specific restriction was imposed, which was valid for all the nine provinces of Emilia-Romagna, except for some local control measures (for example, for schools and universities);
- A full lockdown was first imposed to the provinces of Modena, Parma, Piacenza, Reggio nell'Emilia, and Rimini, as of 8 March 2020 [51], and then extended to the remaining provinces of Bologna, Ferrara, Forlì-Cesena, and Ravenna on 10 March, 2020 [52];

The data on which we exerted our testing activity were essentially of two types: i) the time series relative to the new daily CoVid-19 infections, and ii) the air pollution In Emilia-Romagna, under the form of the measurements of the following pollutants: $PM_{2.5}$, PM_{10}, and N02, taken on a daily basis at all the aforementioned provinces (Bologna, Ferrara, Forlì-Cesena, Modena, Parma, Piacenza, Reggio nell'Emilia, Rimini and Ravenna).

The number of daily infections was collected using the GitHub repository of the Italian Civil Protection, for the entire period starting on February, 24th and closing on April, 17th, 2020 [37].

The daily values of the pollutants mentioned before, instead, were collected using the website of the Regional Environmental Protection Agencies (ARPA) of the Emilia-Romagna region, for all the nine provinces we have cited before [10]. Since there were multiple monitoring stations distributed over each province, an average of the values returned by each station was computed, on a daily, provincial basis.

More important is what follows. We have all learnt that this CoVid-19 infection can be subjected to an incubation period, whose duration can range from a few days to almost 14 before an infected human begins to manifest some given symptoms. More precisely, authors of [83] maintain that the median incubation period can be estimated to be 5.1 days (with a confidence interval of 95%, it takes from 4.5 to 5.8 days) and that the 97.5% of those who develop symptoms will do so within 11.5 days (with a confidence interval of 95%, it takes from 8.2 to 15.6 days). These estimates imply, in the end, that 99% of the infected population will develop symptoms within 14 days. Further, other authors also emphasize that a spare delay of 3.6 days can be experienced from the moment in time the result of a virological test is performed, and the time when it is recorded in the correspondent database [29].

These are the reasons why we designed the two different time series:

- the one with the average daily pollution values (say X), and the one with the number of the new daily infections (say Y),
- where X was anticipated in time with respect to Y of 14 days.

We decided not to use an offset of sixteen days (as resulting from the sum of 12.5 with 3.6) between a) and b), simply because this minimum time difference lag was absorbed by the specific statistical methodology we have employed (i.e., the Granger causality), where we have varied the so-called lag length parameter

in a range from 3 to 8 days (as better explained in the Subsection 2.4 below) [56].

Following this reasoning, the period when we measured the particulate (specifically, $PM_{2.5}$, PM_{10}, and NO_2) started on February, 10th and closed on April, 3rd, 2020. As already told, instead, the period for measuring the infections was: February, 24th – April, 17th, 2020.

Hence, in the end, it should be clear that an offset has been put that temporally separates these two time series, due to the consideration that all that can happen on a given day, say x, may have its effect in terms of manifestations of the infection after a period in time which can be as long as x + 14 days.

Figures 3.1, 3.2, and 3.3 report the twenty-seven graphs showing how our time-series ($PM_{2.5}$, PM_{10}, and NO_2 vs. infections) evolve in time.

As already anticipated, we are interested in studying the relationship between air pollution and CoVid-19 infections also with machine learning. The idea is to count the number of daily infections registered per each province, in all the nine provinces of Emilia-Romagna, during the four days that preceded the lockdown decision taken by the Italian Government on 8–10 March 2020 (the specific day depends on the specific province).

Once those infections counts were obtained, we computed an average value of those daily numbers on a per-province basis for those 4 days. We then got nine numbers that were finally aggregated on a regional basis, under the form of a further average count, thus yielding the average number of infections per province on a regional basis in Emilia-Romagna. The result was 17 (from now on, the so-called threshold). Told differently, the daily number of infected people, in Emilia-Romagna, averaged over those four days, amounts to 17 times 9 = 153.

Now, please follow the reasoning. If the Italian Government, using its own decisional models, opted for a lockdown decision, as soon as the average regional number of daily infections on a per-province basis in Emilia-Romagna had surpassed the threshold of 17, then we could use that number as a key to design the predictions scheme of our ML model. Not to forget also the fact that Emilia-Romagna was, at that time, the region with the largest number of

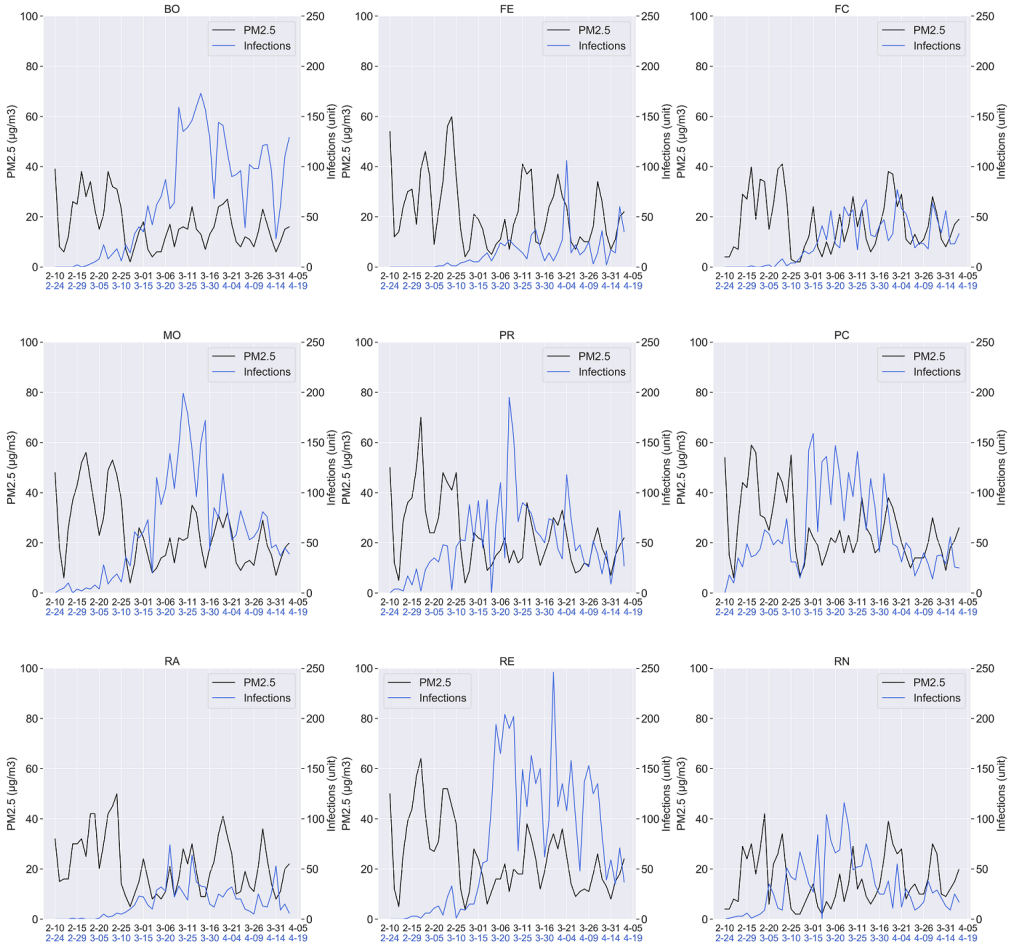


Fig. 3.1 Particulate matter ($PM_{2.5}$) and CoVid-19 infections (all the examined periods in Emilia-Romagna)

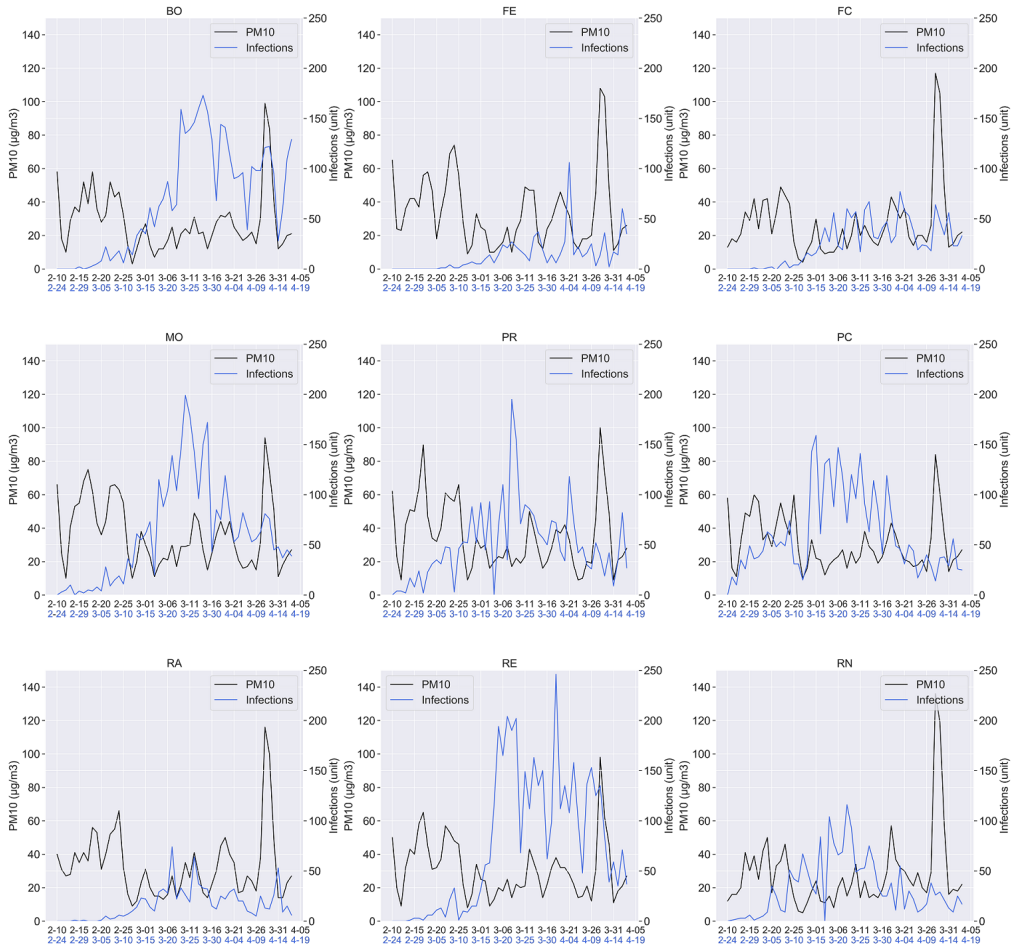


Fig. 3.2 Particulate matter (PM_{10}) and CoVid-19 infections (all the examined periods in Emilia-Romagna)

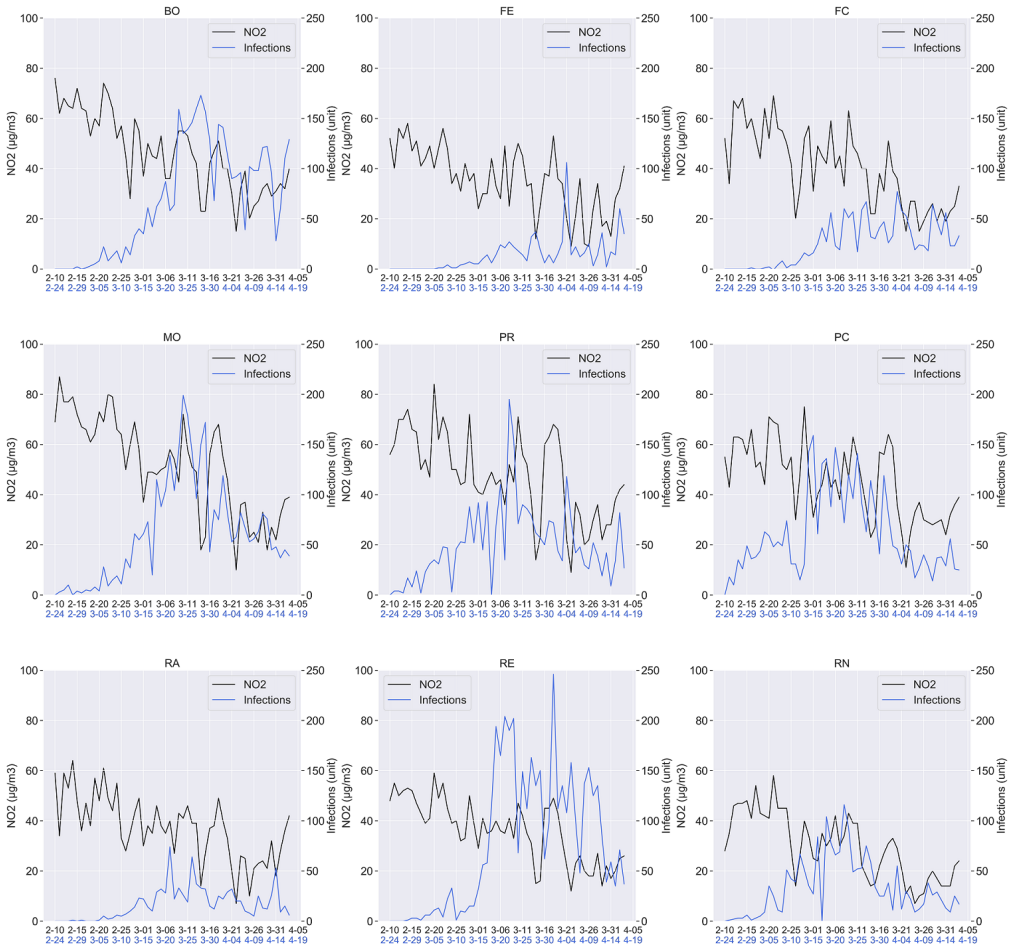


Fig. 3.3 Nitrogen dioxide (NO_2) and CoVid-19 infections (all the examined periods in Emilia-Romagna)

infections after Lombardy. Hence, the number of infections that occurred in this region has had an important role in that lockdown decision.

To conclude this reasoning, our intention is to replace the initial idea to predict if, in a given day, the number of infected people will surpass that threshold of 17, on that day, on a per-province, regional basis. More precisely, we ask our ML model to compute the probability that, in a given future day, each province in Emilia-Romagna will count a number of infections larger than 17—and, then, we look at the regional picture with all its nine provinces, and the probability that the number of infections for each exceeds 17.

For the sake of completeness, in Figure 3.4, we provide a graph with the cumulative quantities of infected people, per day, for all the nine provinces of interest, plus the cumulative values of the regional and the national averages, registered during the four days prior to 8–10 March 2020.

In Figure 3.4, one can read: Bologna, bo; Ferrara, fe; Forlì-Cesena, fc; Modena, mo; Parma, pr; Piacenza, pc; Ravenna, ra; Reggio nell'Emilia, re; Rimini, rn; Emilia-Romagna, er; and Italy, ita.

Important to note is the fact that, in Figure 3.4, our regional infection average, being cumulative over those four days, amounts to 17 times 4 = 68 (as read at the rightmost end of the figure).

By contrast, if one takes into consideration the national average, one can notice that the following value of 8 times 4 = 32 can be computed (as read at the rightmost end of the figure). This smaller quantity at a national level is due to the fortunate fact that many regions in Southern Italy were not severely affected by the virus, thus providing a smaller contribution to the national average.

Interesting to remind is also the average number of infections per day in Lombardy, computed in a similar way (i.e., the average per-province number of infections, on a regional basis), which was as high as 38. This latter number is important. It is well known that in that period, Lombardy was really the hardest-hit Italian area, thus becoming a sort of hotspot for CoVid-19 diffusion in Italy. This is why we decided not to choose this number (i.e., 38) in our

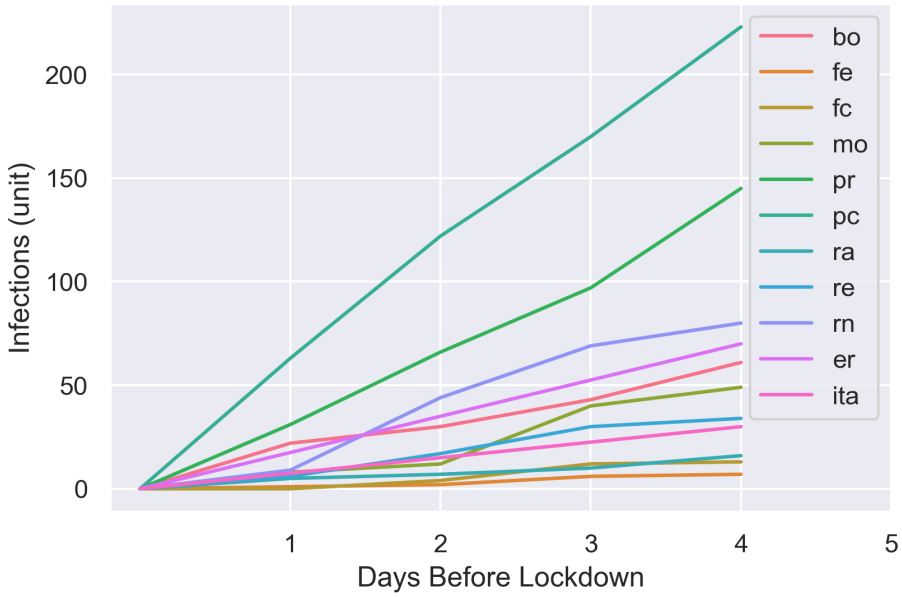


Fig. 3.4 4–7 March 2020—cumulative number of infections: Modena, Parma, Piacenza, Reggio nell'Emilia, and Rimini; 7–10 March 2020—cumulative number of infections: Bologna, Ferrara, Forlì-Cesena, and Ravenna; cumulative regional and national infections averages

scheme. It would have been somewhat misleading, especially in consideration of the fact that we want predictions that are valid for the Emilia-Romagna region.

At this point, it is important to mention that, using the value of 17, we have essentially split our initial dataset into two separate portions:

- The former, with all those days with a number y of daily infections, equal or smaller than 17; and
- The latter, with those days registering a number of newly infected people larger than 17.

Not only this but also, to properly manage the hypothesis of a relationship between pollutants and infection spread, crucial is also the concept of lag. In particular, with lag, we account for the following fact: On a given day, say z, we

may have registered a certain number of infections, say y. Those y infections could have manifested themselves after exactly fourteen days since the original contagion happened exactly on the day: z - 14. Nonetheless, we also know that there is a degree of uncertainty, affecting the exact number of days that should be taken into account for this count.

To take this fact into account, with a lag equal to 4, for example, we reason as if all the y infections, which occurred on day z, originated from the contribution of pollutants that were in the air during a longer time interval of length 4 (starting from day: z - 14). In this specific case, our interval would go from day z - 14 up to day z - (14 - 4 + 1), that is, day: z - 11.

This is an important fact, giving rise to an important implication: With the concept of lag, which can range from 1 to 8 in our model, we try to mitigate the uncertainty concerning the exact day when people get infected.

State of New York

Also in this case, the data at the base of our study was essentially corresponding to two types of time series.

The former was relative to the new daily CoVid-19 infections registered in all the counties of interest in the period March 4th - 22nd 2020, while the latter was concerned with the air pollution, in particular, the particulate matter $PM_{2.5}$ registered on a daily basis, in the period February 19th - March 8th, 2020, in all the counties of interest.

The information relative to the daily CoVid-19 infections was retrieved from the website of the New York State Department of Health - CoVid-19 Tracker [104], while the daily pollution $PM_{2.5}$ levels were collected from the website of the United States Environmental Protection Agency, under the Outdoor Air Quality Data section [152]. Since, in each county there were different sensor stations returning pollution values, at various times during the same day, the correspondent data was aggregated using an average daily value, for each county.

The first issue to explain regarding these two time series of data is concerned with the different periods that were analyzed, that is February 19th - March 8th ($PM_{2.5}$) vs. March 4th – March 22nd (CoVid-19 infections).

It is also important to note that our investigation period ends on March 22nd, in correspondence with the announcement of the Governor of New York State, Andrew Cuomo, who placed the statewide stay-at-home order, starting from 8 p.m. on March 22nd [105]. Hence, the reason to limit our study to the pre-lockdown period is that the lockdown measures might have significantly altered the general situation, with a slowdown of human activities and a consequent change in the pollution levels.

The two corresponding curves of our interest (pollutant and infections), staggered by 14 days, are shown in the plots of Figure 3.5, where those curves are reported for all the counties we have examined (precisely: New York, Kings, Bronx, Richmond, Queens, Nassau, Suffolk, Rockland, Westchester, Onondaga, Oneida, and Monroe).

In Figure 3.5, black lines are the $PM_{2.5}$ pollutant values and the corresponding time period, while red lines are the new daily infections and the relative time period of interest. For each plot, one can see (leftmost) the measurement unit of the $PM_{2.5}$ pollutant (measured in micrograms/m3) and (rightmost) the number of the new daily infected people.

Now is the time to provide the motivations behind our choice of investigating those precise 12 counties in the New York state, represented in grey in the map of Figure 3.6. The rationale was essentially the following: to evaluate if the (potential) relationship between the $PM_{2.5}$ pollutant and CoVid-19 diffusion remained valid through very different scenarios, yet all geographically relative to NYC.

This was translated into the following two conditions. First, we have wanted to investigate both densely populated districts, like those comprised in the city of New York, and also in less populated areas. Second, we have wanted also to extend our analysis both to those counties that are comprised in NYC (or are very close to NYC) and to those counties that are further away from NYC. In

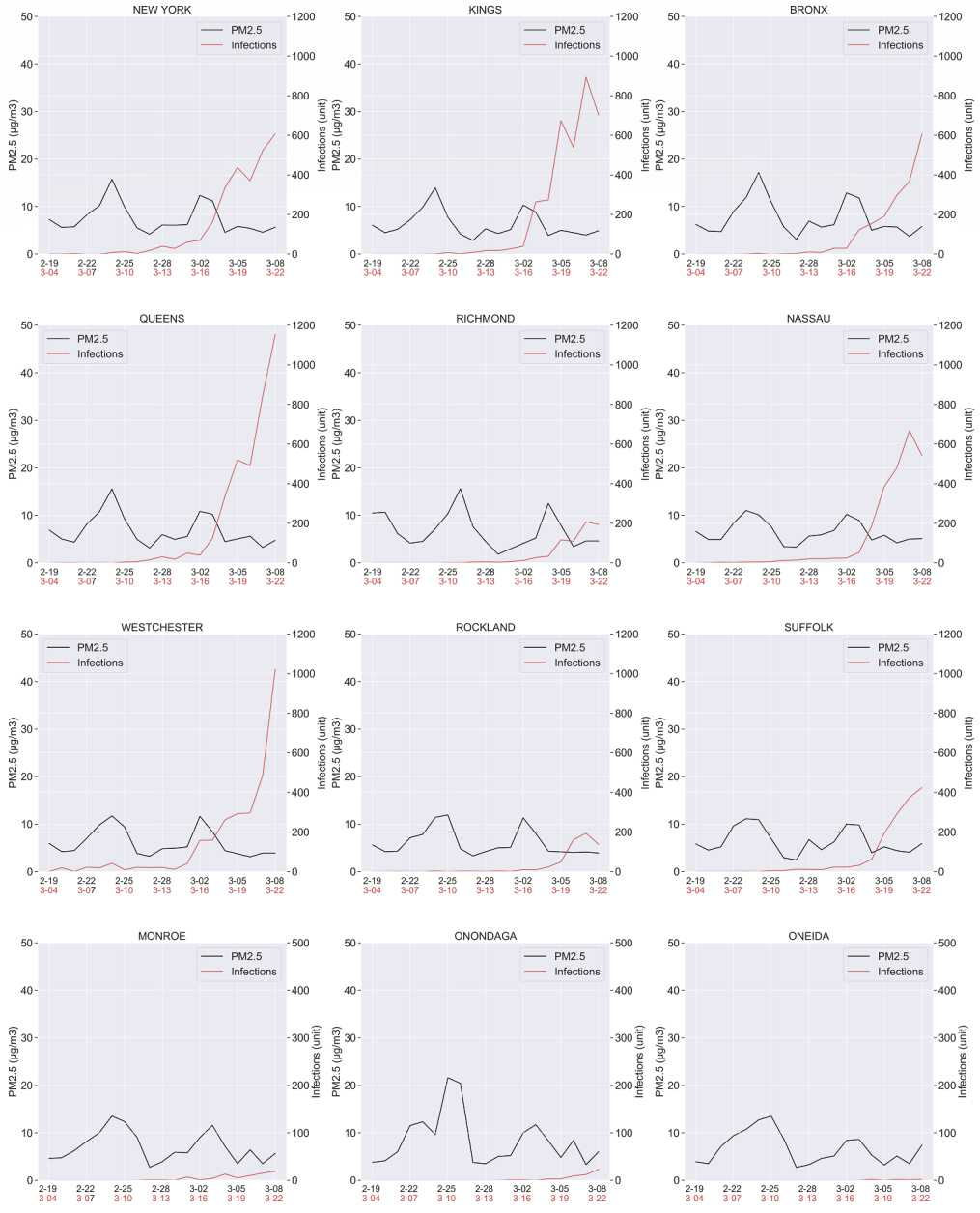


Fig. 3.5 $PM_{2.5}$ and relative period (Black) vs. CoVid-19 infections and relative period (Red).

the end, we decided to choose 12 different counties, that can be considered as clustered in three different groups. First, the counties/boroughs of NYC: New York (Manhattan), Kings (Brooklyn), Bronx (The Bronx), Richmond (Staten Island), and Queens (Queens). Second, the counties that are very/quite close to NYC: Nassau, Suffolk, Rockland, and Westchester. Third, a group of suburban, less populated counties that are far away from NYC: Onondaga, Oneida, and Monroe.

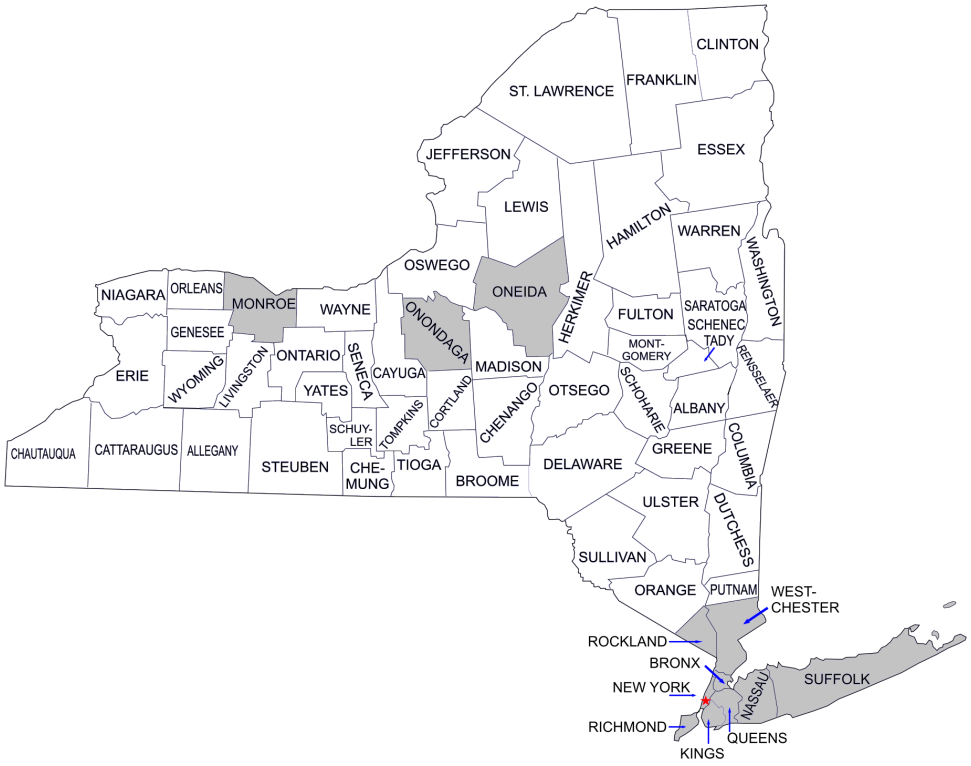


Fig. 3.6 New York State map (scrutinized counties: in grey).

Also for the New York State chosen counties, we computed the threshold of infections, the led to the decision of a full lockdown. We followed the same approach used for the Emilia Romagna region. We counted the number of daily infections registered per each county, in all the twelve counties of interest, during the four days that preceded the lockdown decision. Once obtained those daily

infections counts, we computed an average over all those 4 days, on a per-county basis. We, then, got 12 numbers that were definitely aggregated under the form of a final average count, thus yielding an infections threshold equal to 122.8. All the values used to compute our average are reported in Table 3.1.

County	**Infections**			
	03/17	**03/18**	**03/19**	**03/20**
New York	69	161	335	437
Kings	39	264	273	674
Bronx	29	123	154	191
Queens	38	123	336	519
Richmond	11	26	33	116
Nassau	24	52	186	385
Westchester	157	158	261	292
Rockland	9	8	23	48
Suffolk	22	31	62	193
Monroe	1	4	13	5
Onondaga	1	0	3	3
Oneida	0	0	2	0
Daily Average	122,7916667			

Table 3.1 Number of CoVid-19 infections over four different days per each county.

3.4.2 Granger Causality Approach

As already mentioned, we have employed a Granger causality testing model to study if a causal correlation may exist between particulate matter and the spread of new CoVid-19 infections in Emilia-Romagna [111].

This is a statistical hypothesis testing model typically used to determine if there is a causal relationship between two time series. In particular, a time series X is said to Granger-causes a time series Y if the prediction of the nth value of Y, using both the past values of X and Y, provides more information rather than the prediction based only on past values of Y [57].

This model typically rests upon two axioms. The former is that past and present may cause the future, but the future cannot cause the past. The latter is that the cause contains unique information about its effects. Usually, the null hypothesis of such a test is set to the fact that the time series X does not Granger-cause the time series Y, while, consequently, the unique alternative hypothesis is that the time series X Granger-causes the time series Y.

In our study, the alternative hypothesis was that the pollutants' time series Granger-causes the time-series of the infections. Hence, our aim has been to verify if we could reject the opposite null hypothesis (i.e., pollution does not Granger-cause infections), based on the available data.

To this aim, we set the level of significance at 5%, hence preparing to reject the null hypothesis, only in the case that the corresponding p-values came less than 0.05. Further, as the test assumes that both the time-series under investigation should be stationary, we check and found this condition satisfied using the well-known augmented Dickey-Fuller method [39].

Not only. Since we have designed two time series where the former (X = pollution) temporally precedes the latter (Y = infection), we did not need to check if the infection Granger causes the pollution, given that the time precedence of Y by X comes naturally.

Nonetheless, it is important to repeat, here again, a concept we have already anticipated in the Introduction. Neither the Granger causality method, nor any other statistical test can provide a final and convincing evidence that two phenomena are correlated, from an epistemological viewpoint, if one has neither a clear knowledge of the motivation that causes that relationship nor has developed sufficient experiments at a scale that should be appropriate to the observed phenomena.

With this regard, the Granger causality approach suffers from an additional problem. In fact, if both X and Y are driven by a common third process, say W, one might still accept the alternative hypothesis of Granger causality (X Granger causes Y), even though it is evident that both X and Y have a common cause (i.e., W), that determines their mutual correlation [9].

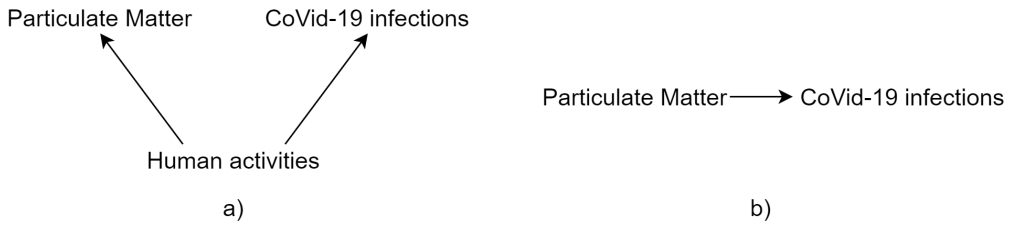


Fig. 3.7 Causality structure

Moving this argument at the center of our specific case, one could even argue that the human activities (playing the role of W, here) have been the common basis for the correlation between pollution and infections (as portrayed in example a) of Figure 3.7) and, hence, a true causality relation between pollution and infection could not be demonstrated, even when our alternative hypothesis is accepted. Nonetheless, in the next section we will show results, taken both before and after the lockdown decisions (when almost all human activities were at a minimum), that do seem to confirm the existence of a causal structure similar, instead, to that shown in example b) of Figure 3.7.

To better understand how a Granger causality testing model works from a computational perspective, fundamental is the following explanation.

We start from two time series X and Y (i.e., pollution and infections), whose causal relationship is to be either demonstrated or rejected. In other words, X and Y are the time series under investigation that can be modeled with the following Granger causality equation:

$$Y_t = \sum_{i=1}^{L} \alpha_i Y_{t-i} + \sum_{i=1}^{L} \beta_i X_{t-i} + \varepsilon_t$$

Specifically, Y_t and X_t are the single elements of the two series Y and X, and, in our case, they correspond to the values that Y and X can take on, on a daily basis. In essence, with the formula above we can compute current values of Y, based on previous values of both X and Y. How far back one can go with previous values of X and Y, to perform the computation of the current value of

Y, is given by the value of L, the so-called lag. To complete the formula, ε_t is a white-noise-random vector.

This said, now comes the turn of explaining how to use this formula for performing a Granger causality hypothesis testing. To this aim, crucial is the role of the β coefficients. In fact, we can say that X Granger-causes Y only if the β coefficients are not zero, since only in this case past values of X (and Y) become useful to compute current values of Y. On the contrary, β coefficients equal to zero make a null contribution to the final sum. It is now easy to understand that modelling a causal relationship with the Granger formula amounts to perform a statistical hypothesis test, where the null hypothesis is that all the β coefficients are zero:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_L = 0$$

The alternative hypothesis is, instead, that at least one of the β coefficients is different from zero.

From a computational perspective, at this point, in a case like that of our study, assigned all the actual values for Y and Y, a vector autoregressive procedure (VAR) is to be run to derive the β coefficients. Upon computation of those β coefficients, the F test procedure must be performed to check if those computed values fit with all zero distributions of the null hypothesis. This statistical test will return p-values. The higher the returned p-values, the more plausible is the null hypothesis. The lower the p-values, the more plausible is the alternative hypothesis: that is, X Granger-causes Y.

Said about the general Granger computational process, now comes the motivation why we have chosen this procedure for our study, rather than other more traditional statistical approaches, like, for the example, the one adopted in [145].

To better understand, consider the following example: Suppose we want to evaluate if a relationship exists between the number of viral infections that happened on a specific day (e.g., February 28th) and the amount of pollution in the air. To do that, traditional approaches would compute values, based on

measurements taken on just two days: the day of the infections vs. the day assumed to be the one when the pollution occurred that was considered at the basis of those infections, say for example February 14th. Exactly like in the example a) of Figure 3.8.

With the approach based on the Granger formula, instead, we can take into simultaneous consideration multiple days, each with its amount of measured pollution. This is by virtue of the lag factor (i.e., the L value in the Granger formula above) that allows one to go back as many days as one wants in the computation. For example three days, like in the case b) of Figure 3.8 (or from 3 to 8, like in the case of our study, see Subsection 2.2).

This is a prominent computational aspect that should not go neglected since the information on when a given infection precisely occurs comes with a large amount of uncertainty. Still more remarkably, since CoVid-19 is manifesting with variable temporal dynamics, we should adopt flexible computational methods to study it. From this point of view, as the series shown in Figure 3.9 comparatively demonstrate, methods à la Granger should be preferred, since they hold the promise to analyze simultaneous contributions to the cause of a unique effect.

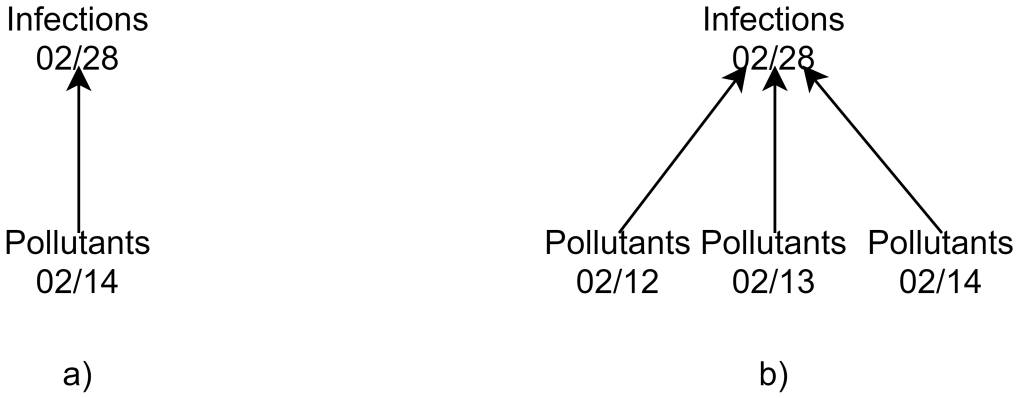


Fig. 3.8 The role of the lag factor in the Granger formula

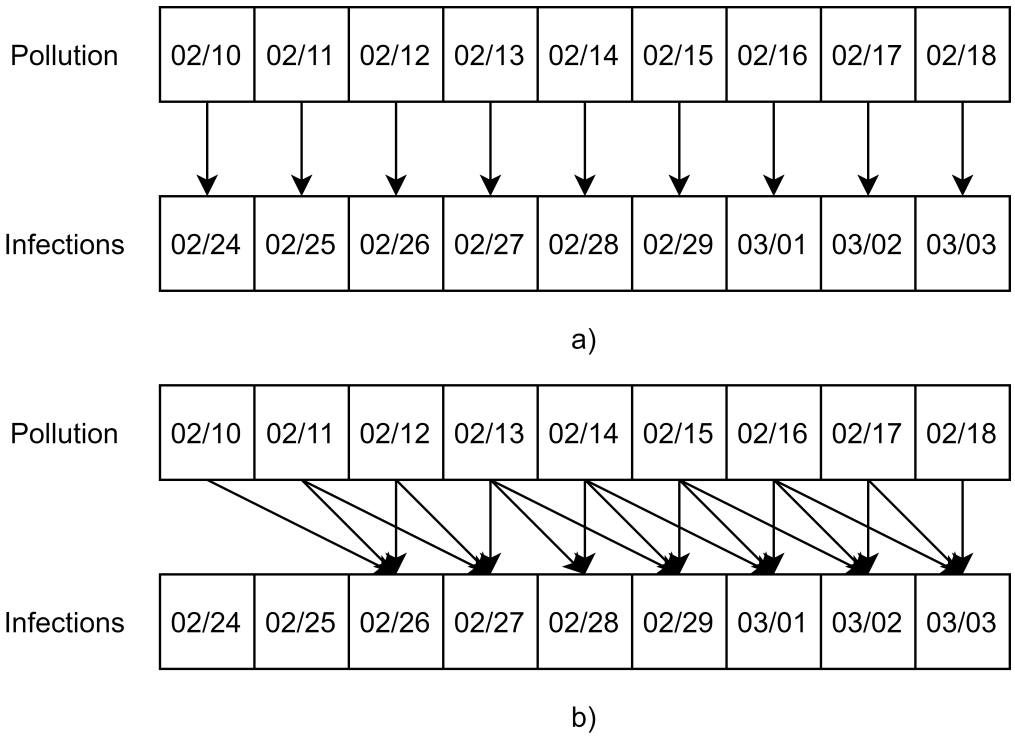


Fig. 3.9 Comparing temporal series: traditional methods (a); à la Granger methods (b)

3.4.3 Machine Learning Models

Essentially, we devised a non-traditional procedure, resembling a kind of an ML (province/county) cross-validation methodology which went as follows: during the training activity, we let some ML algorithms be instructed with the daily values of the $PM_{2.5}$ (input) and the CoVid-19 infections (output). The periods of these two series of daily data were different for the Emilia-Romagna region and the New York counties. For the former, the $PM_{2.5}$ ranged from February 10th up to June 30th, 2020 while the CoVid-19 infections spanned from February 24th to July 7th. For the latter, instead, we consider from February 19th to March 8th for $PM_{2.5}$ and from March 4th to March 22th for CoVid-19 infections.

More precisely, the number of the CoVid-19 infections for each given precise day, say X, were put in relation with the amount of the values of the $PM_{2.5}$,

registered in all those days included in the following time interval: [X – 7, X – 14]. The choice of these eight days, prior to X, was taken depending on two different factors: i) the need to be as close as possible to the correspondent lag value used in the Granger analysis (which was equal to 5), and ii) as a result of the ML hyper-parameters optimization process. After this learning phase, this procedure went through a kind of testing validation where the instructed algorithms had to predict if in a given day, in a specific province/county, the number of infections either exceeded a predefined infections threshold or they did not.

To summarizing, the entire process worked as follows. With each round of our procedure, our ML algorithms were trained with the data ($PM_{2.5}$ vs. CoVid-19 infections) relative to all nine provinces of Emilia-Romagna regions and to all the twelve scrutinized New York counties, except for the ones for which we asked our algorithms to predict the number of daily infections, given the concentrations of the $PM_{2.5}$ particulate occurred in previous days. This procedure was repeated, in turn, for all the provinces/counties under investigation. Obviously, the more accurate were the predictions on the infections threshold exceedances, for the counties subjected to our investigation, the more was confirmed the hypothesis of a correlation between $PM_{2.5}$ and the CoVid-19 spread, in those areas.

We now come to the employed ML algorithms. We used the following ones:

- K-Nearest Neighbors [75],
- Support Vector Machine [35],
- Multi-Layer Perceptron [96],
- Extra Tree [53].

With regard to their hyper-parameters, they are reported in Table 3.2.

Finally, as concerns the evaluation metrics, we employ the F1 score. In a classic classification problem (comprising true and false positives, and true and false negatives), it is intended to be the harmonic mean of the precision and recall values, where such a score reaches its best at one. In turn, precision is

Algorithm	Hyper-parameters	Value
KNN	N Neighbors	5
	Weights	uniform
SVC	C	1
	Kernel	RBF
	Degree	3
	Gamma	1/8
MLP	Hidden Layer	1
	Hidden Layer size	100
	Max Epochs	500
	Activation Function	ReLU
	Optimization Algorithm	Adam
	Batch Size	16
	Learning Rate	0,001
ET	N Estimators	50
	Criterion	Gini
	Min Samples Split	2
	Min Samples Leaf	1
	Max Features	Sqrt(8)
	Bootstrap	False

Table 3.2 ML algorithms hyper-parameters

the number of true positives divided by the number of true positives plus the number of false positives, while recall is the number of true positives divided by the number of true positives plus the number of false negatives (i.e., all the samples that should have been identified as positive).

3.5 Results

3.5.1 Granger Causality: results on Emilia Romagna region

We present the results returned by the Granger causality test on the data of the Emilia-Romagna region, differentiating between those illustrating the situation before the lockdown measures were adopted that contained the infection surge, and those showing the ex-post situation.

Before the lockdown

Figure 3.10 reports the results of our Granger causality testing campaign, conducted for all the nine aforementioned provinces of the Emilia-Romagna (Bologna-BO, Ferrara-FE, Forlì-Cesena-FC, Modena-MO, Parma-PR, Piacenza-PC, Ravenna-RA, Reggio nell'Emilia-RE and Rimini-RM).

As already anticipated, we tried to verify if the series X, comprised of all the average daily values of a given pollutant (e.g., $PM_{2.5}$), measured in terms of micrograms per cubic meter, starting on day x1 and closing on day x2, Granger causes the series Y of the new daily infections, measured in terms of infected human beings, starting on day y1 and closing on day y2, where obviously: y1 = x1 + 14 and y2 = x2 + 14, for all the days between x1 and x2.

For each of the possible combination pollutant ($PM_{2.5}$, PM_{10}, and NO_2) infections, our Figure 3.10 shows in the correspondent cell the p-value obtained through a pairwise series computational comparison, using Granger. All this yields a total amount of 189 pairwise series comparisons. In particular, to read well the results: if a cell in Figure 3.10 reports a p-value less than 0,005, we have a confirmation of the causality relation between pollutant and infections (finally,

note that if a cell in the Figure reports the value of 0, this means that a p-value less than 10-4 was computed).

For an easier comprehension of the Figure, one should also notice that the time scale values reported at the left of Figure 4 are the closing days of the two series (respectively, for pollutants and infections), namely the values termed: x2 and y2.

Precisely, x2 ranges in the Figure from March, 1st to 7th or from March, 3rd to 9th, depending on the specific province under consideration with its correspondent lockdown date (March, 8th, or 10th), while y2 may range from March, 15th to 21st or from March, 17th to 23rd, due to the 14 days-long temporal shift with which we distanced the two series (pollution precedes infections).

To note, finally, is the fact of prominent importance that all the pairwise series comparisons whose results are reported in Figure 3.10 were conducted during a period when the lockdown measures were still inactive since the specific series supposed as the cause of this relation (that is X, the pollutants) starts on February, 10th and closes on March, 7th or 9th, depending on the province.

All this said, what is clear from an analysis of Figure 3.10 is that we have got a total amount of 175 (out of 189) statistical confirmations (almost 93%) that X Granger causes Y; that is. that the pollutants under consideration have some effect on the number of new infections, from a Granger-causality perspective. In particular, this correlation is slightly more evident with $PM_{2.5}$ (yielding 94%), rather than with PM_{10} and NO_2 (92%). Further, to be specified is the fact that 189 are the different pairwise temporal series comparisons, each performed with the Granger method.

Nonetheless, before one can come to some final conclusion, we have to remember, here again, the reflection we have anticipated in the previous Section, and that we can repeat, under the alternative form of a question: What about if the human activities carried out in the period from February, 10th to March 7th or 9th, were the only common cause for both pollution and infections, exactly like in the causality scheme portrayed in the example a) of Figure 3.7?

End Period		PM10->Contagi					End Period		PM10 -> Contagi			
PM	Infections	MO	PR	PC	RE	RM	PM	Infections	RA	BO	FE	FC
01-mar	15-mar	0.0000	0.0014	0.0862	0.4079	0.1665	03-mar	17-mar	0.0000	0.0000	0.0065	0.0000
02-mar	16-mar	0.0000	0.0000	0.0000	0.0000	0.0703	04-mar	18-mar	0.0000	0.0000	0.0003	0.0000
03-mar	17-mar	0.0042	0.0003	0.0000	0.0004	0.0001	05-mar	19-mar	0.0000	0.0002	0.0006	0.0000
04-mar	18-mar	0.0000	0.0183	0.0000	0.0000	0.0000	06-mar	20-mar	0.0000	0.2441	0.0009	0.0000
05-mar	19-mar	0.0031	0.0040	0.0000	0.0000	0.0000	07-mar	21-mar	0.0000	0.0253	0.0085	0.0001
06-mar	20-mar	0.0000	0.0000	0.0000	0.0041	0.0000	08-mar	22-mar	0.0000	0.0002	0.0020	0.0317
07-mar	21-mar	0.0000	0.0000	0.0000	0.0022	0.0000	09-mar	23-mar	0.0000	0.0001	0.0000	0.0425

End Period		PM2.5->Contagi					End Period		PM2.5->Contagi			
PM	Infections	MO	PR	PC	RE	RM	PM	Infections	RA	BO	FE	FC
01-mar	15-mar	0.0000	0.0046	0.0487	0.3044	0.0079	03-mar	17-mar	0.0000	0.0000	0.0956	0.0006
02-mar	16-mar	0.0000	0.0000	0.0000	0.0141	0.2789	04-mar	18-mar	0.0000	0.0000	0.0035	0.0001
03-mar	17-mar	0.0000	0.0003	0.0000	0.0194	0.0057	05-mar	19-mar	0.0000	0.0000	0.0028	0.0000
04-mar	18-mar	0.0000	0.0216	0.0000	0.0055	0.0008	06-mar	20-mar	0.0000	0.0009	0.0026	0.0000
05-mar	19-mar	0.0000	0.0153	0.0000	0.0000	0.0000	07-mar	21-mar	0.0000	0.0004	0.0068	0.0001
06-mar	20-mar	0.0000	0.0000	0.0000	0.0001	0.0000	08-mar	22-mar	0.0000	0.0000	0.0504	0.0282
07-mar	21-mar	0.0000	0.0000	0.0000	0.0001	0.0000	09-mar	23-mar	0.0000	0.0000	0.0066	0.0186

End Period		NO2->Contagi					End Period		NO2->Contagi			
PM	Infections	MO	PR	PC	RE	RM	PM	Infections	RA	BO	FE	FC
01-mar	15-mar	0.0000	0.0000	0.0000	0.0000	0.0000	03-mar	17-mar	0.0016	0.0000	0.0068	0.0000
02-mar	16-mar	0.0000	0.0000	0.0000	0.0000	0.0000	04-mar	18-mar	0.0008	0.0000	0.0017	0.0000
03-mar	17-mar	0.0031	0.0000	0.0000	0.0000	0.0000	05-mar	19-mar	0.0000	0.0000	0.0115	0.0000
04-mar	18-mar	0.0004	0.0000	0.0000	0.0001	0.0000	06-mar	20-mar	0.0000	0.0001	0.0602	0.0000
05-mar	19-mar	0.0000	0.0000	0.0000	0.0369	0.0000	07-mar	21-mar	0.0000	0.0000	0.1248	0.0000
06-mar	20-mar	0.0000	0.0000	0.0000	0.0000	0.0000	08-mar	22-mar	0.1249	0.0000	0.0852	0.0000
07-mar	21-mar	0.0000	0.0000	0.0000	0.0000	0.0000	09-mar	23-mar	0.0803	0.0000	0.0162	0.0000

Fig. 3.10 Particulate matter and CoVid-19 infections (before lockdown): Granger-causality and p-values

If so, the value of the analysis we have conducted so far would be almost controversial. To respond to this doubt, we ask the reader to refer to the next Subsection.

After the lockdown

As already told, the causal modeling method proposed by Granger was designed to handle pairs of variables, and consequently, it may suffer from a typical limitation when a third variable is engaged in the relation, as explained in a previous Section. In our specific case, this third variable could be identified with all the variety of human activities that could be the common cause for both the air pollution and the spread of infections in Emilia-Romagna.

Nonetheless, an important factor has come to the scene through which we will try to argue that the relationships identified in the previous Subsection still

hold. This factor amounts to the lockdown decisions taken either on March 8th or 10th, depending on the specific province under investigation.

As a result of these decisions, human activities had fallen down to a minimum starting again on either March 8th or 10th, depending on the province under consideration. This has a precise meaning with an impact on the rationale behind our analysis, which is as follows: All that happens after those dates can no longer be ascribed to the activity carried out by humans (if not minimally).

Nonetheless, looking at this from an opposite perspective, one should also argue that this new factor (i.e., the lockdown) can also have a confusing effect on the researched phenomena since the absence of humans in the scene could open the way to new unexpected implications, and hence to a variety of different possible interpretations.

To avoid this possible pitfall, we have redesigned our experiments with specific care to select for our analysis only those provinces whose general characteristics could be considered to be more easily observable, with less external interferences. Two design principles drove us for this new set of experiments. The first was to exclude from our analysis all those provinces with a too high number of infected individuals per population, with respect to the average value of the region under investigation. This way, Piacenza, Reggio nell'Emilia, Parma, and Rimini were excluded, yielding the highest percentages of infected individuals per population, namely: 1.509%, 0.906%, 0.723% and 0.606% (as recorded on May 9th, 2020). For an analogous reason, we excluded the largest province in the region, precisely Bologna, since it is suffering a very high number of infected individuals, which are currently as many as 4,751. For an opposite motivation, we cut off from the second part of our study also the province of Ferrara, which for a long time, fortunately, had hit the lowest rate of infected individuals per population (even though it has recently recorded higher values, thus reaching currently the percentage of 0.281%).

Finally, excluded went also the province of Forlì-Cesena, in this case, due to the fact that we measured a marked decrease in the amount of the values of the particulate measured during the new period of investigation.

To this aim, it is interesting to notice that the difference between the amount of particulate matter taken both before and after the lockdown, computed as an average of the daily measurements of the two two-weeks long periods that preceded and followed the lockdown date, ranged in an interval from + 8.65 micrograms per cubic meter (Parma) to – 3.33 micrograms per cubic meter (Rimini) for the PM2.2.5 pollutant, and from + 6.30 micrograms per cubic meter (Parma) to – 10.47 micrograms per cubic meter (Rimini) for the PM_{10} pollutant. (At this point, it is also interesting to remind the reader that the acceptable daily limit considered for PM_{10} pollutant is set to be 50 micrograms per cubic meter).

All this considered, both the province of Modena and Ravenna were rather stable under this perspective, with incremental values amounting to: + 7.86 ($PM_{2.5}$) and + 6.11 ($PM_{2.5}$) micrograms per cubic meter for Modena, and + 2.09 (PM_{10}) and - 3.37 (PM_{10}) micrograms per cubic meter for Ravenna.

In essence, our post-lockdown analysis was confined to just the two provinces of Modena and Ravenna, because they both satisfy all the following requirements:

- a rate of infected individuals ranging from moderate to mild (Modena, 0.538%-3,792; Ravenna, 0.281%-995);
- the number of infected individuals not hitting the highest values in absolute, like instead Reggio nell'Emilia (4,835) and Bologna (4,751), for example;
- relative stability in the in/decrease of the particulate matter after the restrictions imposed by the lockdown.

Summing up, our choice towards these two provinces has been orientated by the fact that they looked like to us as the only provinces on which the changes induced by the lockdown had a minimal external impact, even though the human activities were prohibited. In some sense, they were those provinces less affected by interferences whose causal factors rest not observable and unknowable.

All this said Figure 3.11 reports the results of our Granger causality analysis conducted for the provinces of both Modena (MO) and Ravenna (RA).

End Period		MO		End Period		RA	
PM	Infections	PM10->Contagi	PM2.5->Contagi	PM	Infections	PM10->Contagi	PM2.5->Contagi
08-mar	22-mar	0.0000	0.0000	10-mar	24-mar	0.0000	0.0000
09-mar	23-mar	0.0000	0.0000	11-mar	25-mar	0.0000	0.0000
10-mar	24-mar	0.0000	0.0000	12-mar	26-mar	0.0001	0.0000
11-mar	25-mar	0.0000	0.0001	13-mar	27-mar	0.0144	0.0039
12-mar	26-mar	0.0000	0.0001	14-mar	28-mar	0.0077	0.0006
13-mar	27-mar	0.0000	0.0000	15-mar	29-mar	0.0089	0.0178
14-mar	28-mar	0.0267	0.0500	16-mar	30-mar	0.0073	0.0113
15-mar	29-mar	0.0305	0.0668	17-mar	31-mar	0.0126	0.0120
16-mar	30-mar	0.0272	0.0603	18-mar	01-apr	0.0134	0.0213
17-mar	31-mar	0.0438	0.0472	19-mar	02-apr	0.0126	0.0351
18-mar	01-apr	0.0133	0.0158	20-mar	03-apr	0.0139	0.0382
19-mar	02-apr	0.0165	0.0180	21-mar	04-apr	0.0283	0.0330
20-mar	03-apr	0.0113	0.0116	22-mar	05-apr	0.0282	0.0323
21-mar	04-apr	0.0112	0.0109	23-mar	06-apr	0.0236	0.0273
22-mar	05-apr	0.0254	0.0242	24-mar	07-apr	0.0182	0.0205
23-mar	06-apr	0.0238	0.0197	25-mar	08-apr	0.0142	0.0157
24-mar	07-apr	0.0320	0.0230	26-mar	09-apr	0.0127	0.0154
25-mar	08-apr	0.0358	0.0265	27-mar	10-apr	0.0152	0.0244
26-mar	09-apr	0.0387	0.0285	28-mar	11-apr	0.0073	0.0132
27-mar	10-apr	0.0332	0.0238	29-mar	12-apr	0.0063	0.0143
28-mar	11-apr	0.0402	0.0247	30-mar	13-apr	0.0043	0.0256
29-mar	12-apr	0.0436	0.0243	31-mar	14-apr	0.0027	0.0150
30-mar	13-apr	0.0334	0.0203	01-apr	15-apr	0.0002	0.0085
31-mar	14-apr	0.0260	0.0328	02-apr	16-apr	0.0000	0.0137
01-apr	15-apr	0.0177	0.0431	03-apr	17-apr	0.0000	0.0110

Fig. 3.11 Particulate matter and CoVid-19 infections (after lockdown): Granger-causality and p-values

For a full comprehension of the Figure, one should notice that all have remained unchanged here, with respect to Figure 3.10, as to how the experiments were developed, with just these three natural considerations:

- each observed series closes in a period ranging, respectively, from March 8th (Modena) and March 10th (Ravenna) for the pollutants' series, and from March 22nd (Modena) and from March 24th (Ravenna) for the infections, up to April 1st (Modena) and to April 3rd (Ravenna) for the pollutants' series, and up to April 15th (Modena) and to April 17th (Ravenna) for the infections;

- the beginning day for both series (pollution and infections) remains the same as in the comments provided for Figure 3.10;
- the analysis, this time, was conducted just for the particulate matter of type: $PM_{2.5}$ and PM_{10}, not being available at that time stable measurements for NO_2.

In essence, our scientific target, here, was to verify if the pairwise series correlation observed before was still confirmed, even if we have been adding some more 25 days at each series, with all the 25 days that happened after that the lockdown took place.

To this aim, an analysis of the p-values of Figure 3.11 shows that we have got a total amount of 97 (out of 100) statistical confirmations (yielding a 97% value) that X Granger causes Y; that is, that some given pollutants have some effect on the number of infections, from a Granger-causality perspective.

To be precise, interesting is the fact that a similar analysis conducted for all the other provinces (Bologna, Ferrara, Forlì-Cesena, Parma, Piacenza, Reggio nell'Emilia and Rimini) provides a more controversial result, with a lower number of statistical confirmations (approximately around 50%), probably depending on all those interferences, happened as a consequence of the lockdown, which we mentioned before as the motivation of our decision for the exclusion.

3.5.2 Machine Learning: results on Emilia Romagna region

The results of the province-cross-validation are reported in Table 3.3. In the Province column, we reported the province used as validation. We tested separately each Machine Learning model: K-Nearest Neighbors (KNN), Support Vector Machine (SVC), Multi-Layer Perceptron (MLP), and Extra Tree (ET). We also add a column and a row with the average values for each province and algorithm. All the results are in terms of the F1 score.

Important to note is the fact that we allowed the models to learn our function with all the pollutants (i.e., $PM_{2.5}$, PM_{10}, and NO_2) considered together.

Province	KNN	SVC	MLP	ET	Avg per province
MO	0.87	0.9	0.88	0.91	0.89
PR	0.83	0.8	0.78	0.83	0.81
PC	0.86	0.83	0.85	0.87	0.85
RE	0.82	0.89	0.8	0.91	0.86
RN	0.72	0.73	0.72	0.77	0.74
RA	0.82	0.75	0.82	0.79	0.80
BO	0.79	0.85	0.8	0.8	0.81
FE	0.79	0.74	0.82	0.8	0.79
FC	0.84	0.79	0.79	0.85	0.82
Avg per Algorithm	0.82	0.81	0.81	0.84	

Table 3.3 ML results for Emilia-Romagna region: F1-score obtained with data from to each province as validation set

In essence, each cell in Table 3.3 tells us how accurate, on average, the prediction was that a given model has made that the threshold of 17 infections was either surpassed or not, for a given day, with a certain amount of pollutants in the air.

If one accurately analyzes Table 3.3, they can find that almost all the ML models have comparable performances. In fact, all the four models under consideration yielded a reasonably good performance; nonetheless, the one with the best F1-score was ET, Extra Tree, which achieved an average F1-score of 0.84 compared to 0.81 and 0.82 values, reached by the other models. In other words, ET is the model that better learned the function of pollution/infections on which our hypothesis is based. This can be considered as a further clue of the relationship between air pollution and CoVid-19 infections.

3.5.3 Granger Causality: results on New York counties

The results returned by the Granger procedure are reported in Table 3.4. As previously mentioned, we have subjected to our Granger tests the data relative to the following counties of New York State: New York, Kings, Bronx, Queens, Richmond, Nassau, Westchester, Rockland, Suffolk, Monroe, Onondaga, and

Oneida. For each county, we have evaluated if the time series of the average daily values of the $PM_{2.5}$ particulate (X) Granger-causes the time series of the new daily CoVid-19 infections (Y). The two time series were staggered by fourteen days. This means that, for each Yi, the following temporal relation held Yi - 14 = Xi. The time series of the new daily infections started on March 4th. The air pollution time series started consequently fourteen days before, precisely on February 19th. With regard to the end of the infections time series, we considered different alternatives, using March 20th, 21st, and 22nd as the final days. As a consequence, the end of the $PM_{2.5}$ time series was set, respectively, on March 6th, 7th, and 8th. Therefore, for each county, we subjected three different time series to our Granger procedure. Since we are considering twelve counties, the total number of tests we carried out was thirty-six. For each of these thirty-six tests, Table 3.4 shows the corresponding p-values.

As previously explained, the null hypothesis can be rejected only if the corresponding p-value, returned by the statistical test, is lower than 0.05. Only in that case, we can maintain that the $PM_{2.5}$ time series Granger-causes the time series of the CoVid-19 infections.

As shown in Table 3.4, out of thirty-six tests, the null hypothesis was rejected in as many as thirty-three cases, yielding a 92% of experiments in favor of a hypothesis of an association between the $PM_{2.5}$ particulate and the spread of the CoVid-19, in the various geographical areas relative to NYC. The few cases when the null hypothesis was not rejected are highlighted in red in Table 3.4.

3.5.4 Machine Learning: results on New York counties

We have now come to the end of this Section by showing Table 3.5, where all the forty-eight values of the accuracy of the predictions returned by our algorithms are reported, given in terms of the F1-score metric. Except for just one case (Nassau/SVC), we have obtained forty-seven good F1-score values, all exceeding the value of 0.7. This is both for those counties comprised in (or closer to) NYC, and also for those counties that are further away from the City. In Table 3.5, we have also reported the mean F1-score, respectively computed, averaging both

Start Date (Infections)	04/03		
End Date (Infections)	**03/20**	**03/21**	**03/22**
New York	$< 10^{-4}$	0.0518	0.0902
Kings	$< 10^{-4}$	0.0003	$< 10^{-4}$
Bronx	$< 10^{-4}$	0.0011	0.0003
Queens	$< 10^{-4}$	0.0002	0.1283
Richmond	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Nassau	$< 10^{-4}$	$< 10^{-4}$	0.0018
Westchester	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Rockland	0.0071	0.0	$< 10^{-4}$
Suffolk	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Monroe	$< 10^{-4}$	0.0058	0.001
Onondaga	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Oneida	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$

Table 3.4 Granger causality tests with New York counties

on the twelve counties and on the four algorithms. If we look at the average F1-scores for the counties comprised in (or near to) New York City, they range from 0.84 to 0.89, while average values from 0.87 to 0.95 were returned for those counties that are further away from NYC (i.e., Onondaga, Oneida, and Monroe).

3.6 Discussion and Conclusion

CoVid-19 has radically changed the habits of millions of people around the world, being a global threat that has pushed the healthcare sector all over the world to its limits, causing a dramatic death toll. Hence, factors increasing the risks for severe and fatal courses of such a disease have been deeply studied and include demographic, healthcare, political, business, organizational and climatic ones [159, 13]. Since it mainly manifests as respiratory disease, mostly pneumonia, many researchers are studying the potential relationship between exposure to air pollution, in particular under the form of particulate matter, and the rapid contagion brought by this virus.

County	KNN	SVC	MLP	ET	Avg per county
New York	1	1	0.95	0.82	0.94
Kings	0.95	0.8	1	0.79	0.89
Bronx	0.85	1	0.95	0.82	0.91
Queens	0.9	0.89	0.89	0.89	0.89
Richmond	0.87	0.87	0.87	0.91	0.88
Nassau	0.8	0.7	0.95	0.89	0.84
Westchester	0.95	0.83	0.76	0.76	0.83
Suffolk	0.9	0.85	0.85	0.9	0.88
Rockland	0.77	0.82	0.82	0.82	0.81
Avg per algorithm	0.89	0.86	0.89	0.84	
Monroe	0.85	0.85	0.88	0.91	0.87
Onondaga	0.85	0.88	0.91	1	0.91
Oneida	0.91	0.94	1	0.94	0.95
Avg per algorithm	0.87	0.89	0.93	0.95	

Table 3.5 ML results for New York counties: F1-score obtained with data from to each county as validation set

We presented a study with the aim of providing further proof of this possible relationship, contributing to this discussion. We considered different air pollutants, including $PM_{2.5}$, PM_{10}, and NO_2, under the form of time series of daily values and we evaluated their relationship with respect to the time series of CoVid-19 infections. We considered data relative to two different geographical areas. The former one is the nine provinces of the Emilia-Romagna region, which has been one of the hardest-hit regions in Italy and in Europe, especially at the beginning of the pandemic. The latter one, instead, is the New York State, of which we studied some of the counties, with a particular focus on the ones nearby New York City.

We have conducted a statistical analysis that confirms, under a Granger causality perspective, that a causal correlation may exist between the two researched phenomena of air pollution and CoVid-19 infections, in both the Emilia-Romagna region and in the New York State counties. Furthermore, using the

air pollution data, it was possible to predict if in a given province or state the number of infections exceeds, or not, a certain threshold.

With regard to the possible limitations of our study, we feel necessary to discuss, at least, the three following points: i) the robustness of the scientific methodology we adopted, ii) the choice of the Emilia-Romagna region and the New York State as the primary subjects of our study, and finally iii) the scientific validity of the data we used.

As far as the methods employed, we have already admitted that neither Granger causality, nor Machine Learning algorithms, nor any other statistical testing procedure, can provide final evidence that the two phenomena we have studied (i.e., air pollution vs. CoVid-19 infections) are definitely correlated in nature. In fact, to achieve an ultimate knowledge of this correlation, statistical evidence, like those demonstrated in this chapter, should be always accompanied by additional experiments at a scale that is appropriate to the observed phenomena; that is, in this case, at a biomedical, chemical or even physical level. Apart from this issue, our study has demonstrated that using Granger and Machine Learning algorithms may be valid solutions, over alternative computational methodologies, to infer statistical evidence from sets of data subjected to high levels of temporal uncertainty.

To move on to the second issue, we understand very well that the choice to limit our study to only two different geographical areas, the Italian region of Emilia-Romagna and some counties of the New York State, can be a source of controversy, and a limitation, as well. Anyway, the choice of using two distant and different areas whose populations have populations with very different cultures mitigate such a drawback.

Thirdly, it is the turn of the data. First, we want to emphasize that all the data and statistics used were publicly available, at the time of our investigation, on Italian governmental sites [69, 37, 10], the New York State Department of Health [104], and the United States Environmental Protection Agency, under the Outdoor Air Quality Data section [152]. It is also worth noticing that all our experiments are reproducible using the data available in the public repositories

we have mentioned. Nevertheless, it is also a fact that CoVid-19 infections are by now assumed to be more widespread than initially expected, thus making many of the studies conducted so far (including ours) a poor proxy for understanding the extension of this infection, with all the relative implications [140].

It is important to remind that final proof of the connection between the spread of this fatal virus and air pollutants is not possible employing statistical methods but it requires deep investigations at different levels such as the biological, chemical, and physical ones. However, with our experience, we want to contribute to the discussion about the paradigm shift from knowledge-driven to data-driven science. While we are well aware that big data have been one of enabling factors for the resurgence of artificial intelligence, and in particular of machine learning, in our opinion, the theory has still a central role. In particular, we believe that artificial intelligence and statistical methods can be successfully exploited to confirm or refute hypotheses, based on conceptual insights from the humans involved in the experiments.

Chapter 4

On the interaction with machines for codifying and transferring knowledge

> How humans should interact with algorithms to codify and transfer knowledge to them?
>
> **— RQ-5**

With the development and diffusion of artificial intelligence, organizations and machines increasingly share existing knowledge and generate new knowledge. In this chapter, we contribute to the understanding of the role of human-machine interaction as instances of knowledge codification and transfer. Once modeled different training strategies based on organizational learning theories, we use them to structure the interaction between expert geologists (the trainers) and a set of twelve neural networks (the trainees) to identify the underwater path of an optical cable. Then, we vary: i) the amount of information provided, ii) the level of interaction between the trainers and the trainees, and iii) the empirical setting. Finally, we discuss the implications of the use of the different training strategies and their impact.

4.1 Introduction

Knowledge and competencies are part of companies' intangible assets [95, 70, 33, 58]. Their importance is crucial in an economy where investments in human resources, information technology, and research and development have become critical to prosper and grow [20, 32, 45, 89]. For companies, intangible assets represent a crucial resource for their value creation and competitive advantage [54, 60, 88, 154, 155].

The widespread use of artificial intelligence is affecting our society, including the way companies develop, manage and take advantage of their knowledge base. The 2016 report of the Stanford Committee of the so-called One Hundred Year Study on Artificial Intelligence [144] highlighted the positive impact in many industries and applications, concluding that there is no reason to fear evolution in the field. The proper framing of the current situation is deemed key to face the different decisions that governments, institutions, and firms are called to make on many grounds; from regulation to value building propositions. More skeptical conclusions have been instead presented in the OECD Next Production Revolution Report released in June 2017 [107] based on the analysis of a broader set of technological trajectories and country-level productivity data. Anyway, advancing our understanding of how the introduction of AI could impact an individual organization can, therefore, provide a significant contribution to many companies and their profitability [144]. Rahwan et al. recently framed a new field of study called machine behavior consisting of "... the scientific study of intelligent machines, not as engineering artifacts, but as a class of actors with particular behavioral patterns and ecology." [120]. In this chapter, we aim to provide a contribution to this emerging field, combining computer science and strategic management perspectives. In fact, this work has born from a collaboration with colleagues from the Department of Management. We present the design and deployment of AI-based decision-making systems and their interaction with routines and competencies at both individual and organizational levels, with the aim of providing empirical evidence to advance the debate on

the role of human-machine interaction as instances of knowledge codification and transfer, and, therefore, critical to sustain competitive advantage.

The adoption of AI systems involves several steps. Some of them are characterized by objective elements, such as the dataset, while others consist of subjective choices, like the framing of the problem to be solved and how the system is designed and trained [101]. Therefore, such systems are the results of a human-machine interaction [139], that requires several decisions directly related to organizational and individual knowledge-based assets. A structured solution might improve the codification of previously tacit routines and competencies. At the same time, embedding both individual and organizational knowledge into a specific algorithm might reduce its relevance within a company and alter the level of expertise required in the organization. The study of the interaction between humans and machines when implementing an AI-based system is important to understand the implications on the evolution of knowledge-based assets and how this could change and affect their competitive advantage.

Crucial decisions in the introduction of AI systems, in particular those based on Machine Learning, regard the training process [47]. The nature of data, the level of supervision during training, the choice of the hyper-parameters, and the overall feedback provided to improve the learning process are just examples of such decisions. Implicitly, these processes lead to the development of learning-to-learn routines at the machine level, exploiting individual and organizational knowledge combined with the technical expertise of the developers. As much as training is essential to understand how knowledge and competencies are developed within organizations, machine training becomes essential to understand where these competencies will reside and how they could be leveraged.

In this chapter, we present our experience in the implementation of an AI system based on a company-specific knowledge base. The company operates in the field of geological services. It mainly conducted underwater explorations, collecting, analyzing, and packaging underwater soil characteristics to provide drilling maps and underwater routes to clients in different sectors, including oil and gas, power cables, and telecommunications industries. The distinctive

competencies of the company are the knowledge and experience of its geologists. Hence, the way such knowledge is codified and transferred becomes crucial for the company competitive advantage. We conducted five experiments testing different training strategies to evaluate the feasibility of supporting the geologists in their work of defining the routes or redefining their role in the company.

At the end of our experiments, we were able to successfully develop a system able to support geologists in tracing underwater routes for the installation of submarines cables. Our experience offers several insights. First of all, it is clear that different interactions between the trainer and the trainee has a significant impact on the transferred knowledge and, therefore, on the overall learning process. Then, once the learning has been completed, the routines developed generate consistent behavior in similar scenarios. Hence, the introduction of AI-based approaches into existing processes might lead to the development of new routines and competencies related to the training and deployment of AI.

The remainder of the chapter is structured as follows. Section 4.2 employs organizational learning literature to model different learning strategies and their impact on organizational routines and competencies. Then, Section 4.3 illustrates the research questions at the base of this study. Section 4.4 presents the context and the problem tackled together with the experiments conducted, the results of which are described in Section 4.5. Finally, Section 4.6 concludes the chapter, summarizing our findings, their limitations, and implications for research and practice.

4.2 Background and Related Work

This Section first presents some work to highlight the impact of AI on organizational learning and then discusses the importance of the training phase.

4.2.1 The impact of AI on organizational learning

Two factors have contributed to the possibility of leveraging on AI-based systems, a higher amount of data and the decreasing cost of computational power [144]. Many experts from the academia and from the industry converged on the idea that AI will soon be the core of a company's operating model and an integral part of organizational learning, thus facilitating the reshaping of businesses through the affirmation of forms of collaborative intelligence [157, 82, 123].

However, given the different types of learning processes highlighted above, there are multiple ways in which the interaction between organizational and machine learning can occur. While some applications are opening up radically new markets, many are increasingly used by companies to speed up decision processes, improve their market understanding, increase manufacturing efficiency, and re-engineer their current processes [144]. When the focus is on information processing and coding and on aligning the current routines and processes to the opportunities offered by the novel solutions [2], companies engage in what we have previously described as single-loop/first-order learning. Rather than reconfiguring the organization significantly to face an unexpected exogenous change, they adapt to the changes required by the new opportunities. Through organizational memory, an integral element of the learning process, organizations encode, store, and retrieve relevant knowledge. AI solutions can help preserve, refine, and update an organization's memory, making the availability of routines independent of individuals and capable of surviving considerable turnover over time, complementing organizational learning in the exploitation of historical knowledge [112].

AI technologies could be beneficial not only for exploiting their current knowledge base, but they could also contribute to challenging existing routines and learning to unlearn [90]. The different techniques identified in the literature – e.g., learning by rote, learning by deduction, learning by analogy, and learning by induction, further categorized into supervised and unsupervised learning - are analogous to specific processes deployed at an organizational level [14]. When trained to identify inconsistent assumptions, they can extend their impact

to 'double-loop learning' processes and act as a strategic complement of the organizational learning process, increasing the complexity of the interaction between the organization and the machine.

AI systems could help overcome two specific types of difficulties. First, they could reduce the relevance of computational hurdles and increase the ability to combine large datasets to identify new opportunities through different aggregations or the isolation of deviations. Second, they could go beyond existing assumptions by combining information usually treated separately due to previous experience and practice.

4.2.2 Machine training as a distinctive organizational capability

As well as what happens for human capital in organizations, training is crucial for machines. Large amounts of data, the so-called big data, and high computational power, alone, are not enough. In fact, AI-based systems performance strictly depends on the quality and relevance of the information used for training, and how the training process is carried out. These aspects have caught the eye of scholars from different disciplines, that have focused on the interaction between humans and machines during training [120].

Studies on different types of applications ranging from algorithmic trading to parole decisions [98, 43] show that, if human-specific cognitive biases characterize the information used to train the machines, so will be the prediction generated. Similarly, studies ranging from autonomous driving to online pricing [16, 7] show the relevance for the predicted outcome of the assumptions used to design how to train the machines. Within a 'single-loop learning' perspective, training is relevant to determine the level of efficiency of the new technology in the selected applications. Within a 'double-loop learning' perspective, the training process generates the opportunity to develop a new set of routines associated with the deployment of the new technology within the organization.

Compared to other technologies, AI-based systems have specific novelties. In particular, a higher level of interaction with the existing organizational knowledge is required to design and carry out the training process [149]. The different training choices not only have an influence on the quality of the output of the process but have also an impact on how the system will be incorporated in the organization processes and how it will be used in the future [68]. Focusing on the training processes of machines is, therefore, relevant to shed light on how the interaction between humans and machines could impact organizational knowledge codification and transfer and its strategic implications.

4.3 Research Questions

In this chapter, given the context presented in the previous Sections, we focused our attention on the following research question:

> **RQ-5** *How humans should interact with algorithms to codify and transfer knowledge to them?*

4.4 Methods

In this Section, we describe the context of our case study, the formulation of the problem that we tackled, the dataset provided by the company, the machine learning algorithm employed, and all the experiments conducted.

4.4.1 The Context

We performed our experiments in collaboration with a company that provides geological services supporting underwater explorations. Such a company is particularly concerned with collecting, analyzing, and packaging underwater soil characteristics with the final aim of providing drilling maps and underwater routes to companies working in various sectors, including oil and gas, power cables, and telecommunications ones.

The key competencies of the company derive from the combination of knowledge and advanced tools. With the term knowledge, we mean all the technical notions and experience of the work team, which includes different figures such as surveyors, geologists, computer scientists, and geophysicists. The advanced tools consist of highly technical seabed sampling and testing devices. Hence, we can state that the way geologists codify and share their knowledge is, therefore, an important differentiating factor for companies within this context.

Our case study, in particular, regards the telecommunications industry, with the company that had to identify an underwater route for the installation of cables. Essentially, the company has to find a path, given a set of constraints and parameters, including the risk of cable breaking and route length. The geologists operate using numerous parameters inferred from the reconstruction of sonar-based maps to determine a point-to-point route within a corridor, usually not larger than half a sea mile. As the sea-bed could not be fully observed, though, the geologists' experience in cable route surveying and charting of the seas is critical. The company provided us with the data of two similar projects in different settings with significant geomorphological differences. For confidentiality reasons, we call these two settings Atlantis and Heracleion.

4.4.2 Problem Formulation

At first glance, it might seem like a typical path-finding problem. Such a type of problem is usually tackled and modelled using traditional optimization techniques. In fact, such problems can be traced back to the shortest path problem, using a graph and employing Dijkstra's algorithm to find the shortest path between two vertices. Or, considering machine learning, one could employ Reinforcement Learning, in which an agent takes actions in an environment with the aim of maximizing a cumulative reward.

Yet, we are not here to question those approaches, as their efficacy has been often demonstrated in several contexts. However, those approaches typically leverage on either basic or complex decision logic, i.e. a set of local/general rules to be satisfied within a specific context, in the presence of all the data that need

to fully describe it, and clear rules that determine the costs (and, considering this case study, the risks). Unfortunately, the situation we are encountering in this case study, in an underwater scenario, does not present these peculiarities. In fact, discussing with several geologists of the company, it was not possible to determine the complete set of parameters evaluated by them in determining the best path and their respective weights. In some sense, it was not possible to formalize the tacit knowledge.

For these reasons, we treated the problem as a supervised one. The idea is to train a machine learning model to understand, given the information relating to the current point of the route and to those of its neighbors, what the next point should be. The available information consists of: i) the starting and the ending points, ii) all the points where it is possible to pass the cable with the relative characteristics, and iii) the extent to which the cable could bend. Such data are the same used by the team of geologists when a contract is acquired. When new colleagues join the team, they are provided with these data and work closely with their older teammates to learn from their experience. Our model, therefore, could be trained accordingly. However, we could not account for possible biases related to the geologists' previous experience, as well as any form of tacit knowledge.

4.4.3 Dataset Description

For both Heracleion and Atlantis settings, the two datasets consist of: i) seabed points with latitude and longitude values, expressed as coordinates x and y, and the sea depth as their z coordinate; ii) the soil types description with the lists of polygons that delimit areas characterized by such soil types, and iii) the optimal routes identified by the geologists.

In the Atlantis case, x values vary between 0m and 117,183m., while y between 0m and 71,160m. The total amount of points included in the dataset is 2,302,913, characterized by eleven soil typologies, reported in Table 4.1. The points are spaced 50m from each other.

#	Soil Type
1	coarse sediment
2	fine sediment
3	rock
4	subcropping rock
5	seagrass
6	gas charged sediment
7	depressions
8	slumps
9	scars
10	pipe

Table 4.1 Atlantis soil types

Instead, in the Heracleion case, the x coordinates vary between 0 and 256,735m, while the y between 0m and 143,722m. In this setting, the total number of points included in the dataset is 8,802,133. With respect to soil typologies, in the Heracleion setting, there are twenty-one different ones, eight of which are shared with Atlantis, and thirteen new ones.

#	Soil Type	#	Soil Type
1	coarse sediment	12	escarpment
2	fine sediment	13	hardened sea floor
3	rock	14	linear sonar contact
4	subcropping rock	15	magneto
5	gas charged sediment	16	pock
6	slumps	17	sonar contact
7	scars	18	sond
8	pipe	19	stiffclay
9	cable	20	wreck
10	conefacies	21	vcs
11	debris		

Table 4.2 Heracleion soil types

In both settings, the dataset used to train the machine learning algorithm is created as follows. For each point of the route traced by the geologists, we collect its typology and those of neighboring points. These are used as the input for the algorithm. The output, instead, is represented by the movements, expressed as classes, to reach the subsequent points of the route.

4.4.4 Machine Learning algorithm

We chose to use a Multi-Layer Perceptron (MLP) with only one hidden layer composed of fifty neurons. As already mentioned, the MLP, given a position in a grid with the soil types of the neighbors, has to learn at which point it has to move in. We employ the Rectified Linear Unit (ReLU) as the activation function, using Adam as the solver with a learning rate equal to 0.001. We set the maximum of iterations to 200 even if the training process stops if the validation score has not improved in the last ten epochs.

We divided the dataset into training and test sets, using two-third of the data for the training and the remaining one-third for the testing, in a stratified way. The validation set is composed by randomly selecting the 10% of the training set. To avoid unfortunate cases, in each experiment, we trained twelve different MLPs, randomly changing the composition of the training and test sets. To evaluate the performance of the MLPs, we employ the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC).

Once the training is completed, we use the MLPs to generate the routes. The generated routes are evaluated considering two main aspects. The former one is that the points of the routes have to be within the inspected seabed corridors. Hence, we count how many times the routes get outside the seabed corridor. The latter is the route length, which has to be comparable to the one traced by the geologists.

4.4.5 The Experiments

We conduct several experiments by manipulating the way geologists supervise the MLPs to model four different types of training approaches and the corresponding learning processes, characterized by an increasing level of human-machine interaction. In Experiment 1, we simply trained the MLPs using data relative to the soil type of the surrounding points. The human-machine interaction is minimal and limited to the provisioning of a portion of the relevant information available. The machine, therefore, learns in a pure experiential way. Basically, our MLPs learn how to move on a bi-dimensional grid, like that in Table 4.3. In particular, given the current point (x,y), at the center of the grid, our MLPs have a freedom of choice in terms of movements which has no limits. In essence, it can go straight up to reach point six, left-up to reach point five, right-up to reach point seven, and so on to reach each and any different point within the grid. No other contextual information was initially formalized.

5	6	7
4	**x,y**	0
3	2	1

Table 4.3 Experiment 1, possible movements

In Experiment 2, we complete the information context by adding cable bending constraints. This is done by simply reducing the possible movements, which now depends also in the direction of which the cable is pulled. The movements allowed now are simply three: 1) Turn right forty-five degrees, 2) go straight, and 3) turn left forty-five degrees. Learning is still purely experiential as in Experiment 1, but here the MLPs are given all the codified knowledge present in the organization and which was available to its geologists when faced with the original problem.

In Experiment 3, we complement the general scenario by backtracking the MLPs every time they move outside the portion of the seabed covered by the sonar maps. This is equivalent to having the trainers correct the trainee anytime it makes a wrong choice, based on the information available and their experience,

without explaining why the trainee made a mistake. The trainers, therefore, interact with the machine favoring vicarious learning.

In Experiment 4 we further increase the level of human-machine interaction between the human and the machine. More specifically, the MLPs learn from errors identified by the geologists through a vector recording incorrect moves chosen at any round. This error vector becomes additional training information available at the beginning of the subsequent round. At the end of each round, the vector is updated to include new errors, and this procedure is repeated until no further error occurs. The trainer, therefore, identifies the errors made by the trainee and shares them for future improvements through generative learning.

In Experiment 5, we repeat Experiment 4 to replicate the most sophisticated type of human-machine interaction in a different seabed. The same 12 MLPs used in all previous experiments in Atlantis are deployed in Heracleion.

Table 4.4 summarizes the 5 experiments and their main components: the trainer-trainee relationship, the type of learning process observed, and the settings.

# Exp	**Trainer – Trainee interaction**	**Learning**	**Setting**
1	Basic information transfer	Experiential with incomplete information	Atlantis
2	Complete information transfer	Experiential with complete information	Atlantis
3	Backtracking	Vicarious	Atlantis
4	Feedback on mistakes	Generative	Atlantis
5	Feedback on mistakes	Transductive	Heracleion

Table 4.4 A summary of the experiments

4.5 Results

The results of the experiments are discussed in isolation in the following Subsections.

4.5.1 Basic Information Transfer

The results of Experiment 1 are reported in Table 4.5. The columns are the classes of moves coherent with the alternatives available. Each row reports the AUC value after the training for each set of moves for all MLPs, and the average AUC of each MLP across all meaningful classes. In fact, it is important to notice that there were no examples for classes five and six. Hence, it is not possible to evaluate the accuracy of the classifiers on those classes.

In Experiment 1, experiential learning with incomplete information (i.e. based only on seabed characteristics alone) shows very different levels of AUC performance, with a minimum of 0,61 and a maximum of 0,96, and the average AUC value varying between 0,78 and 0,84.

Istance	**Classes and relative AUC**								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**avg**
MLP 1	0.83	0.75	0.90	0.77	0.85	-	-	0.75	0.81
MLP 2	0.86	0.74	0.92	0.73	0.83	-	-	0.81	0.81
MLP 3	0.83	0.74	0.92	0.70	0.82	-	-	0.81	0.82
MLP 4	0.70	0.70	0.93	0.75	0.80	-	-	0.89	0.80
MLP 5	0.89	0.74	0.90	0.70	0.86	-	-	0.81	0.82
MLP 6	0.88	0.73	0.91	0.81	0.84	-	-	0.87	0.84
MLP 7	0.89	0.62	0.95	0.73	0.89	-	-	0.87	0.83
MLP 8	0.85	0.70	0.96	0.70	0.82	-	-	0.87	0.82
MLP 9	0.87	0.72	0.93	0.65	0.83	-	-	0.75	0.79
MLP 10	0.89	0.73	0.93	0.61	0.82	-	-	0.69	0.78
MLP 11	0.87	0.75	0.93	0.76	0.83	-	-	0.81	0.83
MLP 12	0.88	0.70	0.93	0.76	0.84	-	-	0.87	0.83

Table 4.5 Results for Experiment 1 (Experiential learning with incomplete information)

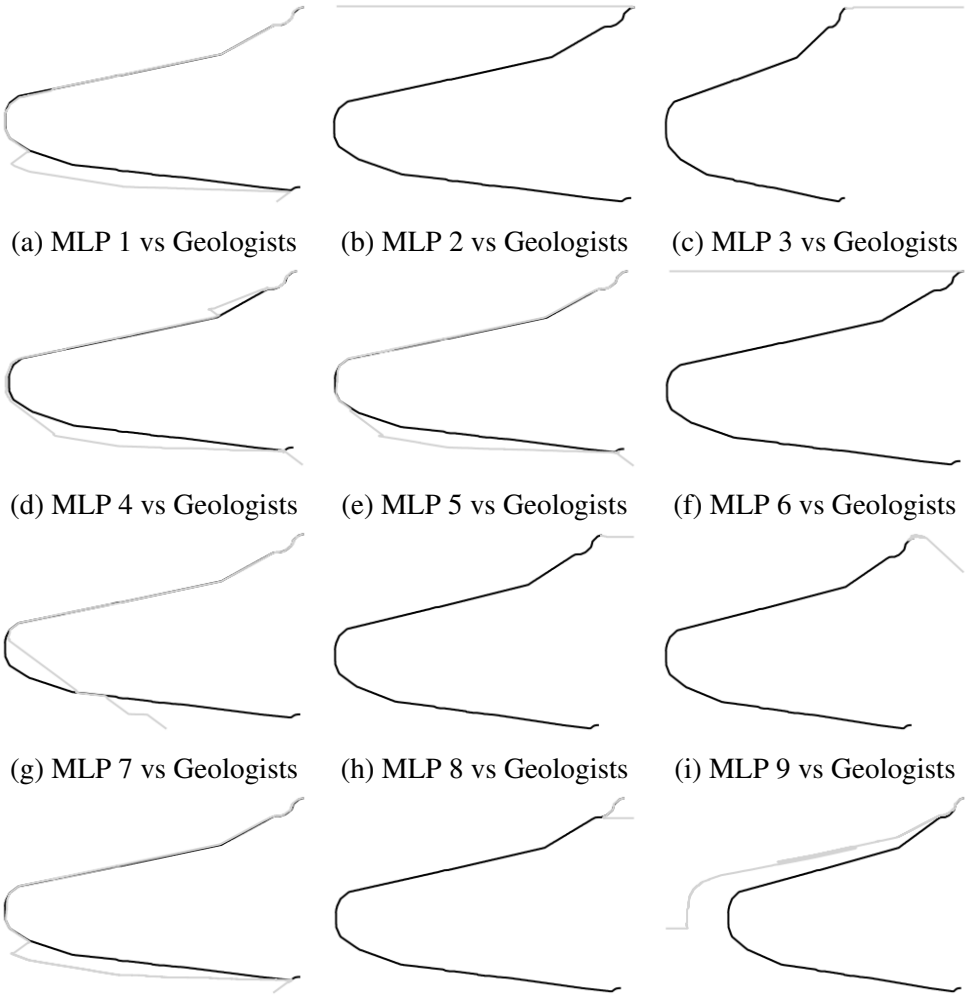
Then, we used the twelve trained MLPs to generate twelve different paths, portrayed in Figure 4.1. As shown, in many cases, the MLPs completely fail in generating the route.

4.5.2 Complete Information Transfer

In Experiment 2, when we include the additional information about the impossibility to bend the cable more than forty-five degrees, the number of classes useful to guide the MLP after the training is further reduced to classes 1, 2, and 3, as explained in the previous Section. Table 4.6 report the results of the Experiment 2. Also in this case, the columns are the classes of moves coherent with the alternatives available. When the characteristics of the cable and the corresponding directionality constraints complete the information provided to guide experiential learning, AUC values improve. The overall minimum raises to 0,71 and the maximum to 0,97, while the average values vary between 0,8 and 0,88.

Istance	**Classes and relative AUC**			
	1	**2**	**3**	**avg**
MLP 1	0.86	0.78	0.97	0.87
MLP 2	0.85	0.78	0.96	0.86
MLP 3	0.85	0.80	0.97	0.87
MLP 4	0.84	0.80	0.97	0.87
MLP 5	0.84	0.77	0.96	0.86
MLP 6	0.82	0.78	0.96	0.86
MLP 7	0.85	0.78	0.96	0.86
MLP 8	0.83	0.71	0.96	0.83
MLP 9	0.86	0.77	0.92	0.85
MLP 10	0.85	0.80	0.97	0.87
MLP 11	0.82	0.79	0.97	0.86
MLP 12	0.86	0.80	0.97	0.88

Table 4.6 Results for Experiment 2 (Experiential learning with complete information)



(a) MLP 1 vs Geologists (b) MLP 2 vs Geologists (c) MLP 3 vs Geologists

(d) MLP 4 vs Geologists (e) MLP 5 vs Geologists (f) MLP 6 vs Geologists

(g) MLP 7 vs Geologists (h) MLP 8 vs Geologists (i) MLP 9 vs Geologists

(j) MLP 10 vs Geologists (k) MLP 11 vs Geologists (l) MLP 12 vs Geologists

Fig. 4.1 Experiment 1: MLPs (Light Gray) vs Geologists (Black)

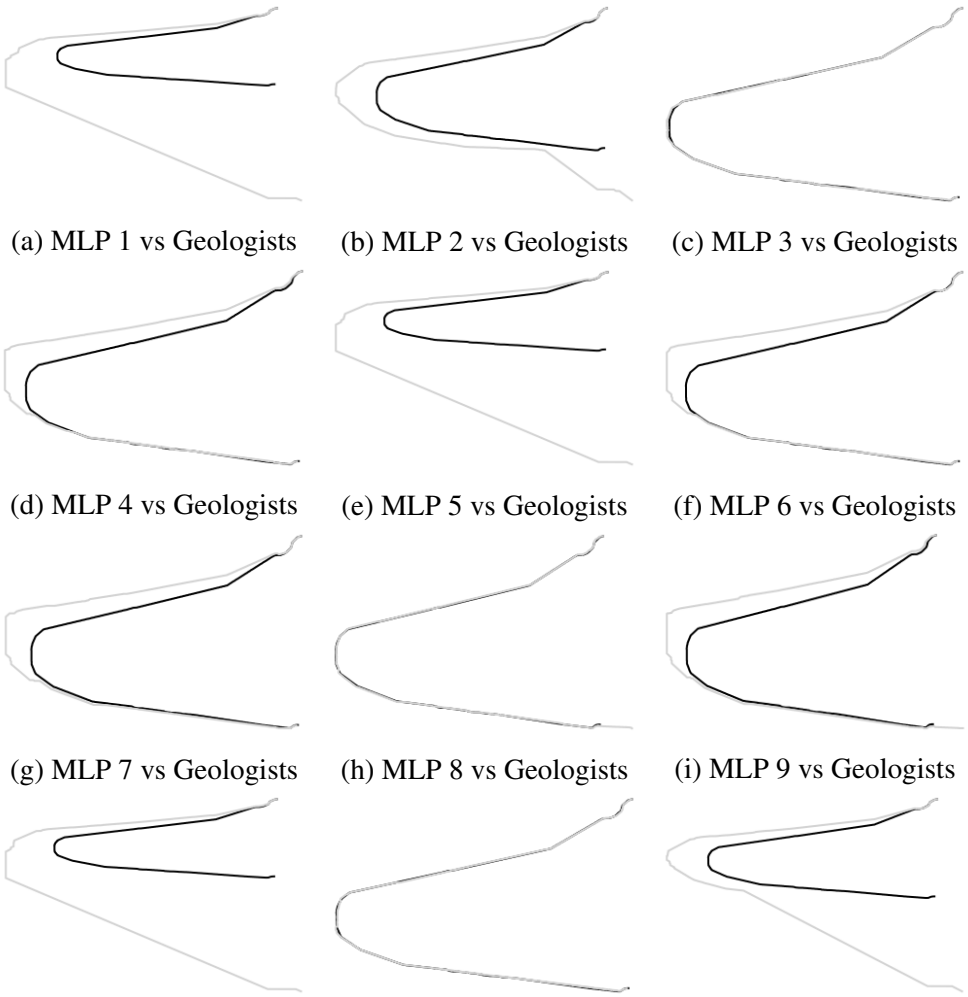
Despite this improvement, however, the overall performance remains unsatisfactory. In fact, after the completion of the training process, we generated twelve routes, using the twelve MLPs. They are depicted in Figure 4.2. Considering the costs associated with laying an underwater cable, making the machine learn only based on the codified information available will generate considerable inefficiencies, even after an extensive training period conducted on different subsets of the benchmark available. Therefore, in a high number of cases, the MLPs fail to reproduce the benchmark.

4.5.3 Backtracking

From Experiment 3, we increase the level of interaction between the trainer and the trainee. In the beginning, the MLPs have the same knowledge acquired in Experiment 2 (i.e., they are the same twelve MLPs trained during Experiment 2). In this experiment, a minimum amount of feedback is provided by sending the MLPs back in the mapped sea-bed corridor anytime it chooses a move that would place them outside. The feedback is minimum because no other information is given, nor is this information used for training. The MLPs, therefore, bounce within the corridor and perform a single run of the estimates for the full route, always completing the path. Hence, as already explained in Subsection 4.4.4, we report also two different measures of performance: the number of errors, and the length ratio, contrasting the MLPs final path with the benchmark. Such measures are summarised in Table 4.7.

The average length ratio varies between a very low 1.01 and a very high 5.72, and the number of errors with the corresponding corrections between 6 and 831. For some, the improvement from Experiment 2 is considerable, while others perform very poorly. The results, therefore, depend both on the new training strategy, and on the subset used in Experiment 2 for training.

For completeness, we have reported the various routes generated during Experiment 3 in Figure 4.3.



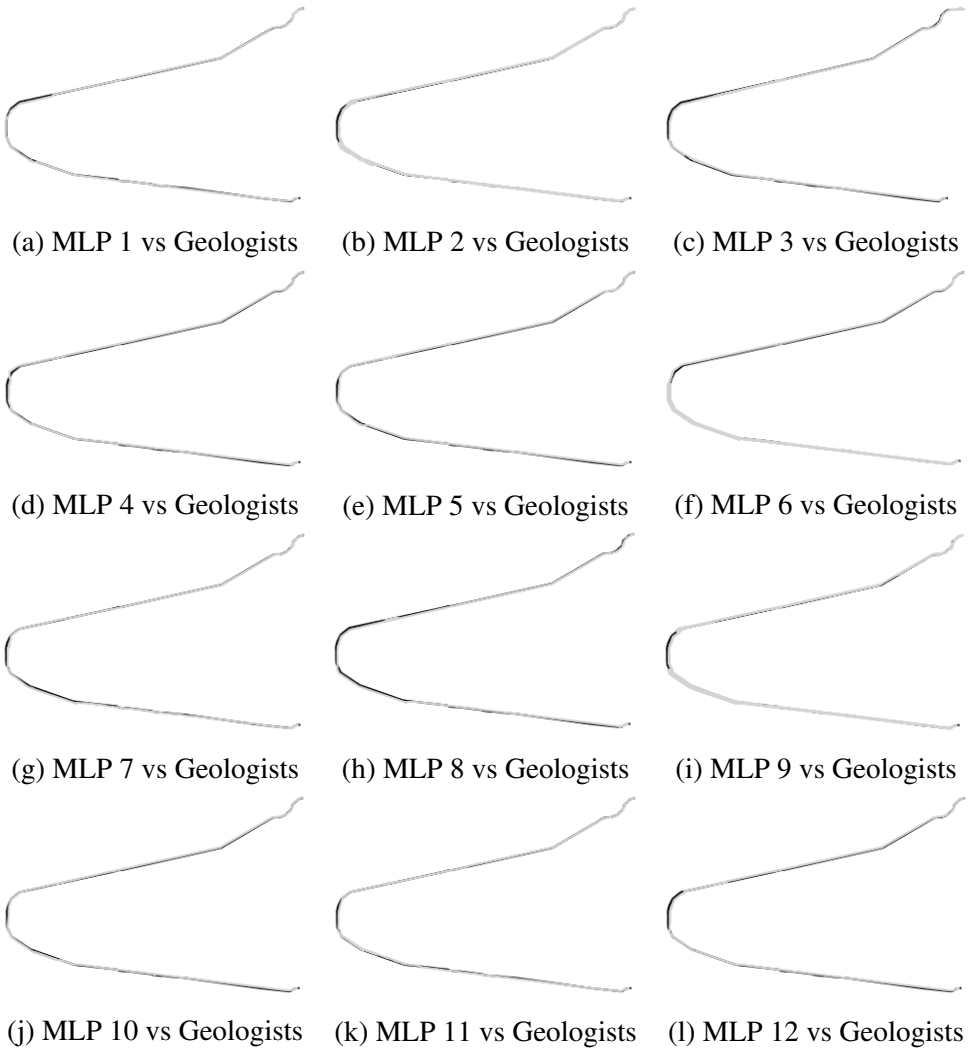
(a) MLP 1 vs Geologists (b) MLP 2 vs Geologists (c) MLP 3 vs Geologists

(d) MLP 4 vs Geologists (e) MLP 5 vs Geologists (f) MLP 6 vs Geologists

(g) MLP 7 vs Geologists (h) MLP 8 vs Geologists (i) MLP 9 vs Geologists

(j) MLP 10 vs Geologists (k) MLP 11 vs Geologists (l) MLP 12 vs Geologists

Fig. 4.2 Experiment 2: MLPs (Light Gray) vs Geologists (Black)



(a) MLP 1 vs Geologists (b) MLP 2 vs Geologists (c) MLP 3 vs Geologists

(d) MLP 4 vs Geologists (e) MLP 5 vs Geologists (f) MLP 6 vs Geologists

(g) MLP 7 vs Geologists (h) MLP 8 vs Geologists (i) MLP 9 vs Geologists

(j) MLP 10 vs Geologists (k) MLP 11 vs Geologists (l) MLP 12 vs Geologists

Fig. 4.3 Experiment 3: MLPs (Light Gray) vs Geologists (Black)

Istance	N° of Errors	Length Ratio
MLP 1	23	1.02
MLP 2	688	4.52
MLP 3	22	1.41
MLP 4	28	1.02
MLP 5	6	1.04
MLP 6	743	4.96
MLP 7	11	1.08
MLP 8	37	1.17
MLP 9	831	5.72
MLP 10	34	1.04
MLP 11	25	1.16
MLP 12	45	1.01

Table 4.7 N° of errors and Length Ratio for Experiment 3 (Vicarious learning)

4.5.4 Feedback on Mistakes - Atlantis

Then, in Experiment 4, we further increase the level of interaction between the trainer and the trainee. Also in this case, in the beginning, the MLPs are the same twelve ones from Experiment 2 but we then introduce a training strategy based on error-based learning with codified feedback. As in Experiment 3, each MLP is sent back in the mapped sea-bed corridor anytime it chooses a move that would place it outside. This time, however, the errors are saved into an additional information set used to fine-tune each MLP in the following rounds. In fact, at the end of each round, the MLPs are re-trained using these additional data sets.

The results are shown in Table 4.8. As shown, the overall performance improves significantly across all MLPs. The length ratio varies between 1,00 and 1,04, 7 out of 12 MLPs make no error in round 2 after learning from those made in round 1, and all the others take only one more round to report no error (Table 4.8, Round 3). The new training strategy, therefore, reduces performance variance among networks, neutralizing the effects of the training choices that generated a high level of performance variance in Experiment 3.

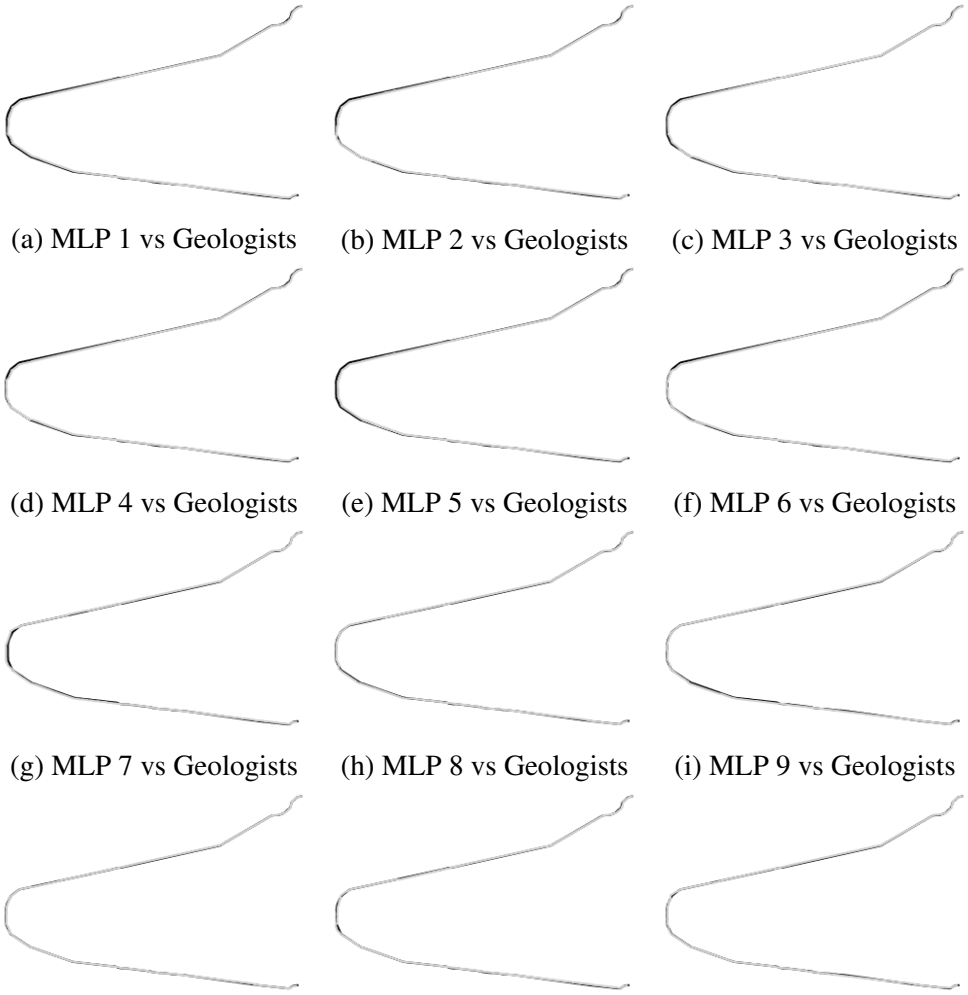
Istance	N° of Errors			Length Ratio
	Round 1	Round 2	Round 3	
MLP 1	20	14	0	1.02
MLP 2	285	0	-	4.52
MLP 3	10	0	-	1.41
MLP 4	30	8	0	1.02
MLP 5	1	0	-	1.04
MLP 6	8	0	-	4.96
MLP 7	493	0	-	1.08
MLP 8	41	18	0	1.17
MLP 9	457	12	0	5.72
MLP 10	37	0	-	1.04
MLP 11	4	0	-	1.16
MLP 12	115	39	0	1.01

Table 4.8 N° of errors and Length Ratio for Experiment 4 (Generative Learning)

The routes generated by each MLP during the final round of Experiment 4 are depicted in Figure 4.3.

4.5.5 Feedback on Mistakes - Heraclion

Finally, in the fifth and last experiment, we apply the same strategy used in Experiment 4, using the same set of the twelve MLPs employed in the final round. But we do not generate routes in the Atlantis setting but in the Heraclion one. This process falls within the category of transductive learning, where the problem is the same, and the corresponding knowledge base, in this case, what we have learned with respect to training strategies, is applied to a different domain. The results of Experiment 5 can be found in Table 4.9. Error-based learning with codified feedback adapts to the new scenario with good levels of efficacy and efficiency, although the MLPs had never been trained before with any data based on Heracleion. The length ratio varies between 1.04 and 1.18, and all 12 MLPs make no error after round 2. Overall the length ratio values are



(a) MLP 1 vs Geologists (b) MLP 2 vs Geologists (c) MLP 3 vs Geologists

(d) MLP 4 vs Geologists (e) MLP 5 vs Geologists (f) MLP 6 vs Geologists

(g) MLP 7 vs Geologists (h) MLP 8 vs Geologists (i) MLP 9 vs Geologists

(j) MLP 10 vs Geologists (k) MLP 11 vs Geologists (l) MLP 12 vs Geologists

Fig. 4.4 Experiment 4: MLPs (Light Gray) vs Geologists (Black)

Instance	N° of Errors			Length Ratio
	Round 1	Round 2	Round 3	
MLP 1	723	10	0	1.07
ANN 2	654	4	0	1.08
ANN 3	464	3	0	1.09
ANN 4	269	79	0	1.07
ANN 5	707	27	0	1.06
ANN 6	834	10	0	1.06
ANN 7	1003	252	0	1.04
ANN 8	992	231	0	1.05
ANN 9	815	204	0	1.05
ANN 10	518	64	0	1.18
ANN 11	475	6	0	1.08
ANN 12	1684	189	0	1.14

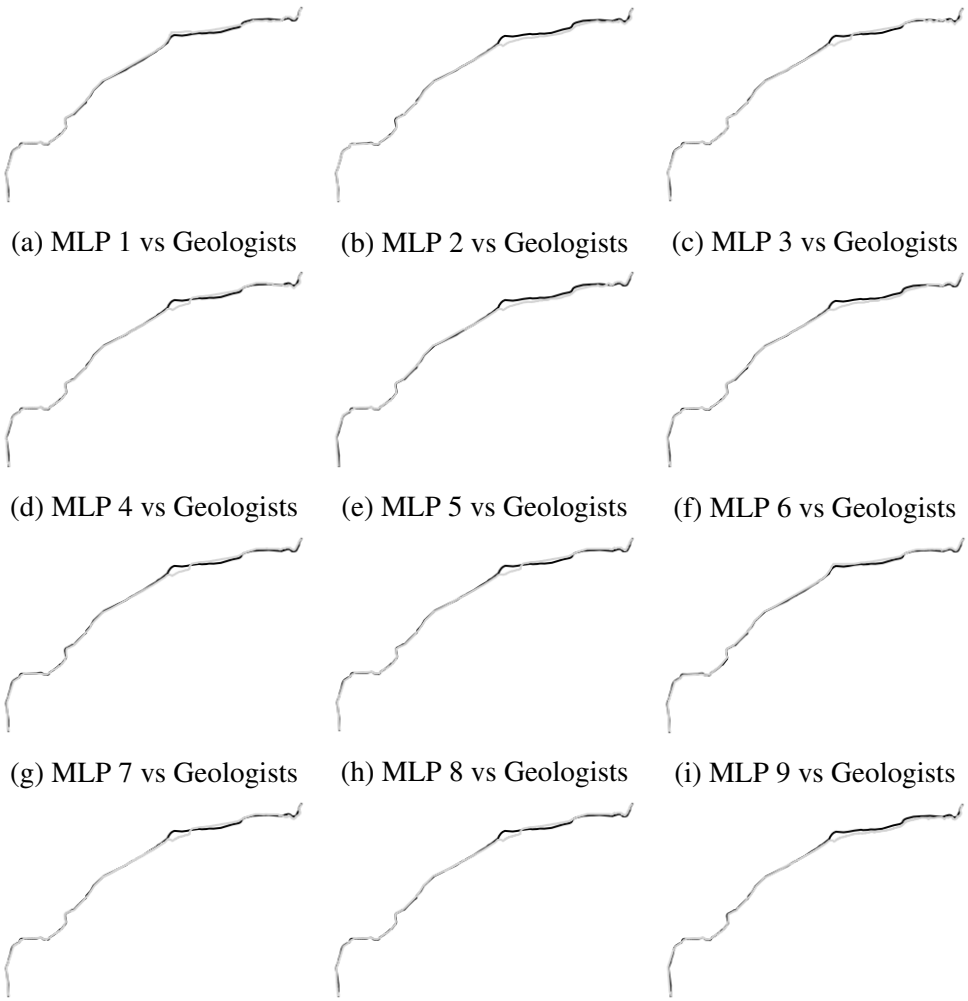
Table 4.9 N° of errors and Length Ratio for Experiment 5 (Generative learning in a new scenario)

higher than in Experiment 4, but in 9 out of 12 cases are within an 8% range, which is considered acceptable given the specific context of the application.

Finally, the routes generated in the third round of Experiments 5 are reported in Figure 4.5.

4.6 Discussion and Conclusion

In this chapter, we presented our experience in the implementation of an AI system based to support geologists of a company in finding the best underwater path for the installation of cables in two different sea-beds. We designed five different experiments in which we manipulated the trainer-trainee relationship, that determines the type of learning process observed. Following Rahwan et al. [120] adaptation of Tinbergen's categories [150], we focus on how a particular set of machines (in our experiments MLPs) acquire their behavior at a hybrid human-machine scale of inquiry. Our findings show that experiential learning with no interaction is less effective even if the codified information set is complete.



(a) MLP 1 vs Geologists (b) MLP 2 vs Geologists (c) MLP 3 vs Geologists

(d) MLP 4 vs Geologists (e) MLP 5 vs Geologists (f) MLP 6 vs Geologists

(g) MLP 7 vs Geologists (h) MLP 8 vs Geologists (i) MLP 9 vs Geologists

(j) MLP 10 vs Geologists (k) MLP 11 vs Geologists (l) MLP 12 vs Geologists

Fig. 4.5 Experiment 5: MLPs (Light Gray) vs Geologists (Black)

Moreover, the most effective training strategy is based on generative learning. Not only it improves the average accuracy of all MLPs, but it significantly reduces their variance to values that make each one fully competitive against the benchmark. Furthermore, when moved in a different scenario, it still offers similar performance, without the need for additional training.

Codifying and transferring information between humans and machines can critically affect knowledge-based assets, that are strategical for an organization. More sophisticated training strategies reduce the impact of potentially subjective choices in the design of the training but they require a non-trivial understanding of both the technical components of AI and of the problem to be solved. Even if most of the debate is currently concentrated around the shortage of supply of the former [147], the latter is often taken for granted as if it could be easily made available or codified to facilitate its transfer and use. It is through the combination of technical and domain-specific knowledge, though, that organizations create, inform and model the behavior of the machines. Several studies focused on the choice of the algorithms and the data as critical aspects to be closely monitored and considered [149, 78]. Our results highlight the relevance of training strategies as an additional component to carefully design and control.

The interaction between humans and machines in terms of AI-based systems could affect the way companies exploit existing knowledge-based assets and develop new ones [153]. All the approaches that fall within the AI umbrella require the transfer of knowledge from organizations to the machine, that is then able to perform a complex, yet narrow, task instead of simply following a set of instructions [68].

To reap the benefits of this new approach, it is important to pay attention to both the technological and the human sides. The behavior of machines instructed with information and feedback can bring about a change in the behavior of the company with respect to problem-solving and experimentation of new approaches to work. They require data, knowledge, and expertise coming from different parts of an organization. Corporate functions that are not used to talk to each other, filled with resistance in sharing results and progress, will found

uncomfortable struggling with the need to start collaborative sessions. When each silo in a firm has its own data, it's impossible to build connections across the silos or with external networks or ecosystems [68]. Differently, companies that transfer their encoded experiences on a regular basis will leverage these new opportunities to enhance communication and collaboration, as well as to reinforce and spread their knowledge base. Our results show that the more they rely on the latter to develop and sustain their competitive advantage, the greater the strategic importance of the design, planning, and deployment of their human-machine interactions.

Based on our results, we also expect that the development of collaborative intelligence forms will increase the strategic role of human resource management [21, 157]. The new required abilities will include teaching intelligent machines new skills or thinking of new ways to overhaul processes to gain improvements. New organizational roles will emerge, such as trainers (of the machines), explainers (of the results stemming out from AI predictions), and sustainers (for the responsible use of machines). New roles will require specific career paths and incentive systems, a redefinition of organizational critical competencies, the way their knowledge-base is aligned with its vision and goals, the way it is stored, developed, protected, and deployed. Our results suggest that expecting a plug-and-play approach to the incorporation of AI-based systems would significantly discount the relevance of all these complementary investments.

In this chapter, we have experimented with a scenario where humans develop, train, and manage AI applications, allowing machines to work as collaborative partners. Our results are limited to a specific set of cases, where existing knowledge and routines can be extended by the use of new technologies. It is an incremental approach relevant for many organizations that are incorporating new methodologies in their routines. Moreover, we incorporated the human-machine interaction in our manipulation of the information set to generate alternative learning opportunities and we made general use of errors in the learning process.

These design choices limit the generalizability of our results in different ways. First, future research should investigate their applicability to more complex

situations such as creative thinking and experimentation, improvisation, and strategy development. Second, we have observed and compared the behavior of the machine based on a change in the behavior of the human component, but not vice-versa. Future research should benefit from field-level studies monitoring the interaction from both sides and explicitly considering the reaction of humans to changes in machine behavior. Third, we isolated our experiments from any exogenous influence or specific organizational constraints, which might, directly or indirectly, modify human-machine interactions and their evolution. Future research should consider explicitly monitoring the relevance and influence of additional organizational factors to help highlight the specificity, if any, of the adoption of AI-based systems. Finally, learning through failure can be costly and not all mistakes are alike, as we treated them in our studies. Future work should consider if a more parsimonious approach might further improve the efficiency of the training without hampering its effectiveness.

Chapter 5

Conclusions

The presence of big data and the consequent diffusion of machine learning algorithms are profoundly shaping social, economic, and political spheres, becoming part of the collective imagination. In this thesis, we have focused on how big data and machine learning algorithms influence the interaction between humans and computers. Starting from different case studies, specially chosen for their characteristics, we have presented some high-level reflections with the aim of contributing to the framing of such a phenomenon. This chapter sums up the various contributions of this thesis, starting from the research questions we presented in Chapter 1.

Chapter 2 describes a case study in which we have designed and implemented a deep learning algorithm with the aim of predicting water meter failures. We collaborated with a company that supplies water in Northern Italy. They provided us with a huge dataset comprised of almost fifteen million water meter readings, plus other contextual information, relative to more than one million water meters. The historical data relates to a period of four years, ranging from the beginning of 2014 to the end of 2018. This chapter answers the research questions **RQ-1**, **RQ-2**, and **RQ-3**. Once implemented the deep learning algorithm, we studied how this can be effectively exploited by the company and how this could be integrated into its processes, even if it is not *perfect* (RQ-1). At the end of the process, we also evaluated the impact of the human-in-the-loop data preparation

activities we carried out from a statistical point of view (RQ-2). Finally, we also proposed an approach to deal with high-dimensional Pareto-distributed categorical data (RQ-3).

Then, Chapter 3 aims to contribute to the discussion that contrasts data-driven with knowledge-driven science, answering the research question **RQ-4**. We conducted an observational study to evaluate the potential relationship between air pollution and CoVid-19 infections. In particular, we studied the time series of air pollution (in the form of both particulate matter 2.5 and 10 and nitrogen dioxide) and infections relative to two different geographical areas: the Emilia-Romagna region (Italy) and some counties of the New York State, with a particular interest on those of New York City. We employed both statistical approaches and machine learning algorithms. In particular, we took advantage of the Granger causality approach and we tested several machine learning algorithms including K-Nearest Neighbors, Support Vector Machine, Multi-Layer Perceptron, and Extra Tree. Our findings suggest that a possible correlation between the two aforementioned phenomena could exist, even if it is important to remind that a final proof of the connection requires deep investigations at completely different levels such as the biological, chemical, and physical ones.

Finally, Chapter 4 presents a work done in collaboration with colleagues from the department of Management, in which we focused on knowledge codification and transfer, from a human-machine interaction perspective. The chapter answer the research question **RQ-5**. We worked in collaboration with a company that design underwater paths for the installation of cables. We implemented different machine learning algorithms, modelling different training strategies based on organizational learning theories. Our findings show that codifying and transferring information between humans and machines can critically affect knowledge-based assets, that are strategical for an organization. Our findings highlight the relevance of training strategies as an additional component to carefully keep under consideration during the design and the implementation.

With regard to future contributions, there are plenty of other aspects about the interaction between human and intelligent algorithms, that could be investigated, keeping the user at the centre. Just to cite a few, it could be interesting evaluating in which contexts the users are more prone to trust algorithms predictions or studying how the accuracy of those algorithms affects the user trust.

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[2] Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

[3] Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., Seneviratne, A., Hu, W., Janicke, H., and Jha, S. K. (2020). A survey of covid-19 contact tracing apps. *IEEE access*, 8:134577–134601.

[4] Akram, T., Lodhi, H. M. J., Naqvi, S. R., Naeem, S., Alhaisoni, M., Ali, M., Haider, S. A., and Qadri, N. N. (2020). A multilevel features selection framework for skin lesion classification. *Human-centric Computing and Information Sciences*, 10(1):1–26.

[5] Allen, G. I. (2017). Statistical data integration: Challenges and opportunities. *Statistical Modelling*, 17(4-5):332–337.

[6] Alvisi, S., Casellato, F., Franchini, M., Govoni, M., Luciani, C., Poltronieri, F., Riberto, G., Stefanelli, C., and Tortonesi, M. (2019). Wireless middleware solutions for smart water metering. *Sensors*, 19(8):1853.

[7] Anbarci, N. and Feltovich, N. (2018). Pricing in competitive search markets: the roles of price information and fairness perceptions. *Management Science*, 64(3):1101–1120.

[8] Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.

[9] Arntzenius, F. (2008). Reichenbach's common cause principle.

[10] Arpae (2020). Emilia-romagna. https://arpae.it/mappa_qa.asp?idlivello=1682&tema=stazioni. Accessed: 2020-04-18.

[11] Becchetti, L., Conzo, G., Conzo, P., and Salustri, F. (2020). Understanding the heterogeneity of adverse covid-19 outcomes: the role of poor quality of air and lockdown decisions. *Available at SSRN 3572548*.

[12] Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A., and Yehudayoff, A. (2019). Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44.

[13] Bhargava, A., Fukushima, E. A., Levine, M., Zhao, W., Tanveer, F., Szpunar, S. M., and Saravolatz, L. (2020). Predictors for severe covid-19 infection. *Clinical Infectious Diseases*, 71(8):1962–1968.

[14] Bhatt, G. D. and Zaveri, J. (2002). The enabling role of decision support systems in organizational learning. *Decision Support Systems*, 32(3):297–309.

[15] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

[16] Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576.

[17] Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

[18] Brankston, G., Gitterman, L., Hirji, Z., Lemieux, C., and Gardam, M. (2007). Transmission of influenza a in human beings. *The Lancet infectious diseases*, 7(4):257–265.

[19] Brock, V. and Khan, H. U. (2017). Big data analytics: does organizational factor matters impact technology acceptance? *Journal of Big Data*, 4(1):1–28.

[20] Brown, S., Lo, K., and Lys, T. (1999). Use of r2 in accounting research: measuring changes in value relevance over the last four decades. *Journal of Accounting and Economics*, 28(2):83–115.

[21] Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534.

[22] Buchanan, M. (2019). *The limits of machine prediction*. PhD thesis, Nature Publishing Group.

[23] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

[24] Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518.

[25] Calude, C. S. and Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of science*, 22(3):595–612.

[26] Caserini, S., Perrino, C., Forastiere, F., Poli, G., Vicenzi, E., and Carra, L. (2020). Pollution and covid. two vague clues don't make an evidence. http://www.scienceonthenet.eu/articles/pollution-and-CoVid-two-vague-clues-dont-make-evidence/stefano-caserini-cinzia-perrino. Accessed: 2020-04-29.

[27] Cassauwers, T. (2019). How confucianism could put fears about artificial intelligence to bed. https://www.ozy.com/the-new-and-the-next/how-confucianism-could-put-fears-about-artificial-intelligence-to-bed/93206/. Accessed: June 21, 2019.

[28] Cerda, P., Varoquaux, G., and Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8):1477–1494.

[29] Cereda, D., Tirani, M., Rovida, F., Demicheli, V., Ajelli, M., Poletti, P., Trentini, F., Guzzetta, G., Marziano, V., Barone, A., et al. (2020). The early phase of the covid-19 outbreak in lombardy, italy. *arXiv preprint arXiv:2003.09320*.

[30] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[31] Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., and Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, 151:147–160.

[32] Collis, D. J. (1994). Research note: how valuable are organizational capabilities? *Strategic management journal*, 15(S1):143–152.

[33] Conner, K. R. (1991). A historical comparison of resource-based theory and five schools of thought within industrial organization economics: Do we have a new theory of the firm? *Journal of Management*, 17(1):121–154.

[34] Conticini, E., Frediani, B., and Caro, D. (2020). Can atmospheric pollution be considered a co-factor in extremely high level of sars-cov-2 lethality in northern italy? *Environmental pollution*, 261:114465.

[35] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[36] Coveney, P. V., Dougherty, E. R., and Highfield, R. R. (2016). Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2080):20160153.

[37] CoVid-19 Italia (2020). Monitoraggio situazione. https://github.com/pcm-dpc/CoVid-19. Accessed: 2020-04-18.

[38] Di Crosta, A., Palumbo, R., Marchetti, D., Ceccato, I., La Malva, P., Maiella, R., Cipi, M., Roma, P., Mammarella, N., Verrocchio, M. C., et al. (2020). Individual differences, economic stability, and fear of contagion as risk factors for ptsd symptoms in the covid-19 emergency. *Frontiers in psychology*, 11:2329.

[39] Dickey, D. and Fuller, W. A. (1979). Distribution of the estimators for time series regressions with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

[40] Dinnon, K. H., Leist, S. R., Schäfer, A., Edwards, C. E., Martinez, D. R., Montgomery, S. A., West, A., Yount, B. L., Hou, Y. J., Adams, L. E., et al. (2020). A mouse-adapted model of sars-cov-2 to test covid-19 countermeasures. *Nature*, 586(7830):560–566.

[41] Ditzler, G., LaBarck, J., Ritchie, J., Rosen, G., and Polikar, R. (2017). Extensions to online feature selection using bagging and boosting. *IEEE transactions on neural networks and learning systems*, 29(9):4504–4509.

[42] Doan, A. (2018). Human-in-the-loop data analysis: a personal perspective. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–6.

[43] Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

[44] Elish, M. C. and Boyd, D. (2018). Situating methods in the magic of big data and ai. *Communication monographs*, 85(1):57–80.

[45] Ely, K. and Waymire, G. (1999). Intangible assets and stock prices in the pre-sec era. *Journal of Accounting Research*, 37:17–44.

[46] Emani, C. K., Cullot, N., and Nicolle, C. (2015). Understandable big data: a survey. *Computer science review*, 17:70–81.

[47] Epstein, Z., Payne, B. H., Shen, J. H., Dubey, A., Felbo, B., Groh, M., Obradovich, N., Cebrian, M., and Rahwan, I. (2018). Closing the ai knowledge gap. *arXiv preprint arXiv:1803.07233*.

[48] Feng, C., Li, J., Sun, W., Zhang, Y., and Wang, Q. (2016). Impact of ambient fine particulate matter (pm 2.5) exposure on the risk of influenza-like-illness: a time-series analysis in beijing, china. *Environmental health*, 15(1):1–12.

[49] Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, 38(1):87–111.

[50] Gazalba, I., Reza, N. G. I., et al. (2017). Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. In *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 294–298. IEEE.

[51] Gazzetta Ufficiale della Repubblica Italiana (2020a). Decreto del presidente del consiglio dei ministri. https://www.gazzettaufficiale.it/eli/id/2020/03/08/20A01522/sg. Accessed: 2020-04-20.

[52] Gazzetta Ufficiale della Repubblica Italiana (2020b). Decreto del presidente del consiglio dei ministri. https://www.gazzettaufficiale.it/eli/id/2020/03/09/20A01558/sg. Accessed: 2020-04-20.

[53] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

[54] Ghemawat, P. (1991). Commitment: the dynamic of strategy free press. *New York*.

[55] Goldstein, J. and McKinley, J. (2020). Coronavirus in n.y.: Manhattan woman is first confirmed case in state. https://www.nytimes.com/2020/03/01/nyregion/new-york-coronvirus-confirmed.html. Accessed: 2020-11-24.

[56] Granger, C. W. (1969). Some recent developments in a concept of causality," journal of econometrics, vol. 39, 1988, pp. 199—211; and. *Investigating causal relations by econometric models and cross-spectral methods," Econometrica*, 37:424–38.

[57] Granger, C. W. (1986). Investigating causal relations by econometric models and cross-spectral methods," econometrica, 37, 424 438.-(1980). *Testing for Causality: A Personal Viewpoint," Journal of Economic Dynamics and Control*, 2:329–352.

[58] Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic management journal*, 17(S2):109–122.

[59] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

[60] Hall, R. (1993). A framework linking intangible resources and capabilities to sustainable competitive advantage. *Strategic management journal*, 14(8):607–618.

[61] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Sun, F., et al. (2020). Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496.

[62] Hernández-Orallo, E., Calafate, C. T., Cano, J.-C., and Manzoni, P. (2020). Evaluating the effectiveness of covid-19 bluetooth-based smartphone contact tracing applications. *Applied Sciences*, 10(20):7113.

[63] Hinds, W. C. (1999). *Aerosol technology: properties, behavior, and measurement of airborne particles*. John Wiley & Sons.

[64] Holzinger, A., Haibe-Kains, B., and Jurisica, I. (2019). Why imaging data alone is not enough: Ai-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(13):2722–2730.

[65] Horne, B. D., Joy, E. A., Hofmann, M. G., Gesteland, P. H., Cannon, J. B., Lefler, J. S., Blagev, D. P., Korgenski, E. K., Torosyan, N., Hansen, G. I., et al. (2018). Short-term elevation of fine particulate matter air pollution and acute lower respiratory infection. *American journal of respiratory and critical care medicine*, 198(6):759–766.

[66] Hosni, H. and Vulpiani, A. (2018). Forecasting in light of big data. *Philosophy & Technology*, 31(4):557–569.

[67] Hutson, M. (2018). Artificial intelligence faces reproducibility crisis.

[68] Iansiti, M. and Lakhani, K. R. (2020). *Competing in the age of AI: strategy and leadership when algorithms and networks run the world*. Harvard Business Press.

[69] Istituto Nazionale di Statistica (2020). Impatto dell'epidemia covid-19 sulla mortalità totale della popolazione residente primo trimestre 2020. https://www.istat.it/en/files//2020/05/Rapporto_Istat_ISS.pdf. Accessed: 2020-05-06.

[70] Itami, H. and Roehl, T. W. (1991). *Mobilizing invisible assets*. Harvard University Press.

[71] James, A. P. and Dimitrijev, S. (2012). Ranked selection of nearest discriminating features. *Human-Centric Computing and Information Sciences*, 2(1):1–14.

[72] Jiang, Y., Wu, X.-J., and Guan, Y.-J. (2020). Effect of ambient air pollutants and meteorological variables on covid-19 incidence. *Infection Control & Hospital Epidemiology*, 41(9):1011–1015.

[73] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

[74] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

[75] Keller, J. M., Gray, M. R., and Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4):580–585.

[76] Ketkar, N. (2017). Introduction to keras. In *Deep learning with Python*, pages 97–111. Springer.

[77] Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1):2053951714528481.

[78] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

[79] Kretzschmar, M. E., Rozhnova, G., Bootsma, M. C., van Boven, M., van de Wijgert, J. H., and Bonten, M. J. (2020). Impact of delays on effectiveness of contact tracing strategies for covid-19: a modelling study. *The Lancet Public Health*, 5(8):e452–e459.

[80] Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 156(3):379–392.

[81] Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2008). Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904.

[82] Lant, T. K. and Shapira, Z. (2000). *Organizational cognition: computation and interpretation*. Psychology Press.

[83] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582.

[84] Lauritano, D., Moreo, G., Limongelli, L., Nardone, M., and Carinci, F. (2020). Environmental disinfection strategies to prevent indirect transmission of sars-cov2 in healthcare settings. *Applied Sciences*, 10(18):6291.

[85] Lawrence, R. L. and Wright, A. (2001). Rule-based classification systems using classification and regression tree (cart) analysis. *Photogrammetric engineering and remote sensing*, 67(10):1137–1142.

[86] Lee, G. I., Saravia, J., You, D., Shrestha, B., Jaligama, S., Hebert, V. Y., Dugas, T. R., and Cormier, S. A. (2014). Exposure to combustion generated environmentally persistent free radicals enhances severity of influenza virus infection. *Particle and fibre toxicology*, 11(1):1–10.

[87] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30.

[88] Lev, B. (2001). Intangibles: management. *Measurement, and Reporting, Brookings Institution Press*.

[89] Lev, B. and Zarowin, P. (1999). The boundaries of financial reporting and how to extend them. *Journal of Accounting research*, 37(2):353–385.

[90] Levitt, B. and March, J. G. (1988). Organizational learning. *Annual review of sociology*, 14(1):319–338.

[91] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England journal of medicine*.

[92] Li, Z. and Wang, Y. (2018). Domain knowledge in predictive maintenance for water pipe failures. In *Human and machine learning*, pages 437–457. Springer.

[93] Liang, Y., Fang, L., Pan, H., Zhang, K., Kan, H., Brook, J. R., and Sun, Q. (2014). Pm 2.5 in beijing–temporal pattern and its association with influenza. *Environmental Health*, 13(1):1–8.

[94] Lipovetsky, S. (2009). Pareto 80/20 law: derivation via random partitioning. *International Journal of Mathematical Education in Science and Technology*, 40(2):271–277.

[95] Lippman, S. A. and Rumelt, R. P. (1982). Uncertain imitability: An analysis of interfirm differences in efficiency under competition. *The bell journal of Economics*, pages 418–438.

[96] Lorencin, I., Anđelić, N., Španjol, J., and Car, Z. (2020). Using multi-layer perceptron with laplacian edge detector for bladder cancer diagnosis. *Artificial Intelligence in Medicine*, 102:101746.

[97] Loslever, P., Laassel, E. M., and Angue, J. (1994). Combined statistical study of joint angles and ground reaction forces using component and multiple correspondence analysis. *IEEE Transactions on biomedical engineering*, 41(12):1160–1167.

[98] Man, K., Wang, J., and Wu, C. (2013). Price discovery in the us treasury market: Automation vs. intermediation. *Management Science*, 59(3):695–714.

[99] Markopoulos, P. P., Kundu, S., Chamadia, S., and Pados, D. A. (2017). Efficient l1-norm principal-component analysis via bit flipping. *IEEE Transactions on Signal Processing*, 65(16):4252–4264.

[100] Maziarz, M. (2015). A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8(2):86–105.

[101] McAfee, A. and Brynjolfsson, E. (2017). *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company.

[102] Monedero, I., Biscarri, F., Guerrero, J. I., Roldán, M., and León, C. (2015). An approach to detection of tampering in water meters. *Procedia Computer Science*, 60:413–421.

[103] Moore, A. (2019). When ai becomes an everyday technology. *Harvard Business*.

[104] New York State Department of Health (2020a). Covid-19 tracker. https://covid19tracker.health.ny.gov/views/NYS-COVID19-Tracker/NYSDOHCOVID-19Tracker-DailyTracker. Accessed: 2020-11-24.

[105] New York State Department of Health (2020b). New york state on pause. https://coronavirus.health.ny.gov/new-york-state-pause. Accessed: 2020-11-24.

[106] Ntakaris, A., Mirone, G., Kanniainen, J., Gabbouj, M., and Iosifidis, A. (2019). Feature engineering for mid-price prediction with deep learning. *Ieee Access*, 7:82390–82412.

[107] OECD (2017). Next production revolution report.

[108] Ogen, Y. (2020). Assessing nitrogen dioxide (no2) levels as a contributing factor to coronavirus (covid-19) fatality. *Science of the Total Environment*, 726:138605.

[109] Omura, K., Kudo, M., Endo, T., and Murai, T. (2012). Weighted naïve bayes classifier on categorical features. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 865–870. IEEE.

[110] Palaniappan, R. and Mandic, D. P. (2007). Biometrics from brain electrical activity: A machine learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):738–742.

[111] Pansini, R. and Fornacca, D. (2020). Initial evidence of higher morbidity and mortality due to sars-cov-2 in regions with lower air quality. *MedRxiv*.

[112] Parkes, D. C. and Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272.

[113] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

[114] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[115] Pettersen, L. (2019). Why artificial intelligence will not outsmart complex knowledge work. *Work, Employment and Society*, 33(6):1058–1067.

[116] Pietrucha-Urbanik, K. (2015). Failure prediction in water supply system–current issues. In *International Conference on Dependability and Complex Systems*, pages 351–358. Springer.

[117] Pisner, D. A. and Schnyer, D. M. (2020). Support vector machine. In *Machine Learning*, pages 101–121. Elsevier.

[118] Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422.

[119] Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301.

[120] Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, 568(7753):477–486.

[121] Rai, A., Constantinides, P., and Sarker, S. (2019). Next generation digital platforms:: Toward human-ai hybrids. *Mis Quarterly*, 43(1):iii–ix.

[122] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., and Nandi, A. K. (2018). Credit card fraud detection using adaboost and majority voting. *IEEE access*, 6:14277–14284.

[123] Ransbotham, S., Kiron, D., Gerbert, P., and Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*, 59(1).

[124] Ranstam, J. and Cook, J. (2018). Lasso regression. *Journal of British Surgery*, 105(10):1348–1348.

[125] Rezaei, M. and Azarmi, M. (2020). Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic. *Applied Sciences*, 10(21):7514.

[126] Rezig, E. K., Ouzzani, M., Elmagarmid, A. K., Aref, W. G., and Stonebraker, M. (2019). Towards an end-to-end human-centric data cleaning framework. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–7.

[127] Roberts, S. and Monks, I. (2015). Fault detection of non-residential water meters. In *MODSIM2015, 21st international congress on modelling and simulation. Modelling and simulation society of Australia and New Zealand*, pages 2228–33.

[128] Rocklöv, J. (2020). Sjödin h, wilder-smith a covid-19 outbreak on the diamond princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *J Travel Med*, 27(3):taaa030.

[129] Roda, W. C., Varughese, M. B., Han, D., and Li, M. Y. (2020). Why is it difficult to accurately predict the covid-19 epidemic? *Infectious Disease Modelling*, 5:271–281.

[130] Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, 100(5):401–409.

[131] Saukani, N. and Ismail, N. A. (2019). Identifying the components of social capital by categorical principal component analysis (catpca). *Social Indicators Research*, 141(2).

[132] Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85.

[133] Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 329. John Wiley & Sons.

[134] Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Perrone, M. G., Borelli, M., Palmisani, J., Di Gilio, A., Torboli, V., Fontana, F., et al. (2020a). Sars-cov-2rna found on particulate matter of bergamo in northern italy: first evidence. *Environmental research*, 188:109754.

[135] Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Perrone, M. G., Piazzalunga, A., Borelli, M., Palmisani, J., Di Gilio, A., Piscitelli, P., et al. (2020b). The potential role of particulate matter in the spreading of covid-19 in northern italy: first evidence-based research hypotheses. *MedRxiv*.

[136] Shen, Y., Mardani, M., and Giannakis, G. B. (2017). Online categorical subspace learning for sketching big data with misses. *IEEE Transactions on Signal Processing*, 65(15):4004–4018.

[137] Shensheng Xu, S., Mak, M.-W., and Cheung, C.-C. (2017). Deep neural networks versus support vector machines for ecg arrhythmia classification. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 127–132. IEEE.

[138] Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267):467–479.

[139] Simon, H. A. (1969). The sciences of the artificial mit press. *Cambridge, MA*.

[140] Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., and Bhattacharya, J. (2020). Seroprevalence of sars-cov-2–specific antibodies among adults in los angeles county, california, on april 10-11, 2020. *Jama*, 323(23):2425–2427.

[141] Soza, L. N., Jordanova, P., Nicolis, O., Střelec, L., and Stehlík, M. (2019). Small sample robust approach to outliers and correlation of atmospheric pollution and health effects in santiago de chile. *Chemometrics and Intelligent Laboratory Systems*, 185:73–84.

[142] St. Clair, A. M. and Sinha, S. (2012). State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! *Urban Water Journal*, 9(2):85–112.

[143] Staszkiewicz, P., Chomiak-Orsa, I., and Staszkiewicz, I. (2020). Dynamics of the covid-19 contagion and mortality: Country factors, social media, and market response evidence from a global panel analysis. *Ieee Access*, 8:106009–106022.

[144] Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., et al. (2016). Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence.

[145] Su, W., Wu, X., Geng, X., Zhao, X., Liu, Q., and Liu, T. (2019). The short-term effects of air pollutants on influenza-like illness in jinan, china. *BMC public health*, 19(1):1–12.

[146] Tellier, R. (2009). Aerosol transmission of influenza a virus: a review of new studies. *Journal of the Royal Society Interface*, 6(suppl_6):S783–S790.

[147] Terence, C., Esposito, M., and Goh, D. (2019). *The AI republic: Building the nexus between humans and intelligent automation*. BookBaby.

[148] Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

[149] Thomaz, A. L. and Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737.

[150] Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, 20(4):410–433.

[151] Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on pattern analysis and machine intelligence*, (3):306–307.

[152] United States Environmental Protection Agency (2020). Outdoor air quality data. https://www.epa.gov/outdoor-air-quality-data/download-daily-data. Accessed: 2020-11-24.

[153] Verganti, R., Vendraminelli, L., and Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management*, 37(3):212–227.

[154] Villalonga, B. (2004). Intangible resources, tobin'sq, and sustainability of performance differences. *Journal of Economic Behavior & Organization*, 54(2):205–230.

[155] Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic management journal*, 5(2):171–180.

[156] White, A., Wyse, J., and Murphy, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed gibbs sampler. *Statistics and Computing*, 26(1-2):511–527.

[157] Wilson, H. J. and Daugherty, P. R. (2018). Collaborative intelligence: humans and ai are joining forces. *Harvard Business Review*, 96(4):114–123.

[158] Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

[159] Wolff, D., Nee, S., Hickey, N. S., and Marschollek, M. (2021). Risk factors for covid-19 severity and fatality: a structured literature review. *Infection*, 49(1):15–28.

[160] Wu, X., Nethery, R. C., Sabath, B. M., Braun, D., and Dominici, F. (2020). Exposure to air pollution and covid-19 mortality in the united states. *MedRxiv*.

[161] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M., Dahly, D. L., Damen, J. A., Debray, T. P., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369.

[162] Yang, L. (2008). Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):438–450.

[163] Yang, W., Shaff, J., and Shaman, J. (2021). Effectiveness of non-pharmaceutical interventions to contain covid-19: a case study of the 2020 spring pandemic wave in new york city. *Journal of the Royal Society Interface*, 18(175):20200822.

[164] You, S., Tong, Y. W., Neoh, K. G., Dai, Y., and Wang, C.-H. (2016). On the association between outdoor pm2. 5 concentration and the seasonality of tuberculosis for beijing and hong kong. *Environmental Pollution*, 218:1170–1179.

[165] Yu, L., Sun, X., Tian, S., Shi, X., and Yan, Y. (2018). Drug and nondrug classification based on deep learning with various feature selection strategies. *Current Bioinformatics*, 13(3):253–259.

[166] Zhang, Z. and Jordan, M. I. (2009). Latent variable models for dimensionality reduction. In *Artificial intelligence and statistics*, pages 655–662. PMLR.

[167] Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., and Xie, G.-S. (2017). Discriminative elastic-net regularized linear regression. *IEEE Transactions on Image Processing*, 26(3):1466–1481.

[168] Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273.

[169] Zurcher, A. (2020). Coronavirus spreading in new york like 'a bullet train'. https://www.bbc.com/news/world-us-canada-52012048. Accessed: 2020-11-24.