POLITECNICO DI TORINO

Master's degree in ICT for Smart Societies

Master's degree Thesis

Features selection for SARS-CoV-2 spread in Italy: observation at regional and provincial level

Supervisors:
Monica Visintin
Guido Pagana

Candidate:
Giulia Ciaramella

A.Y. 2019-2020

SUMMARY

At the end of 2019 a novel virus -later called SARS-CoV-19- began circulating in the Chinese area of Wuhan, causing a few months later the COVID-19 pandemic.
The new disease, which causes severe pneumonia, stirred up problems in the health system of many countries, with hospitals overcrowded with people, many of them not receiving the necessary care due to the limited number of respiratory machines. Having spread across the entire globe, the disease captured the attention of scientists of any type. In the domain of machine learning, researchers applied their knowledge in image diagnosis for classifying a Covid-19 case from other pneumonia cases, or in forecasting methods, to know in advance the number of future diseased people or also in research with therapeutical aim, studying possible usable drugs among some already available on the market, and much more.
Today, 10 months after the discovery of the virus, it continues to be pretty unknown. It spread severely in some countries and less harshly in others.
Moreover, in European territories such as Spain, France, or Italy, the spread of the virus is not homogeneous within the regions and provinces. More precisely, considering the Italian situation, the first period of the pandemic (26th February- 17th April) showed an extremely different situation among the Northern, Center and Southern areas. In particular, Northern Italian territories counted a number of infected which was nine times greater than the infections in the South. Also, the mortality rate changed areas by areas having been 14% in the North, 8.6% in the Center and 8.2% in the South. The large discrepancy of Covid-19 diffusion within the Italian territories has been the inspiration for this work.
The aim of this study hence was to take Italy as a case study, considering first provinces and then regions, and finding the features, among some selected ones that mostly affected the virus spread in the territory.
The selected features vary from health fields such as the number of doctors and nurses, distribution of physical condition of the population through the number of overweight, underweight, normal weight and obese people, the number of nursing homes (hosting the elderly), to mobility information as airport flow, car traffic and usage of public transport, to the average income, to meteorological data, to the number of Severe Accident Risk (S.A.R.) industries present and the number of animals farmed, to the demographic data as the population mean age, number of inhabitants and population density.
Among these features, the goal was to find the most relevant ones that explain the virus diffusion among provinces and regions, in terms of *total cases*, *daily variation of cases* and *deaths*.

To cope with the problem, several regression methods were used: linear models (Linear Least Squares, Lass and ridge), ensemble trees (Random forest and Extremely random forest) and Gaussian process regressions. All three supervised machine learning methods, once chosen the target variable to regress, output the regressed value and the feature relevance: in this way feature selection was possible. The period of study was from February 26th to April 17th for regions and February 16th, March 11st, 21st and 31st and April 11th for provinces, due to the unavailability of some data.
For each regression method, some performance indicators are considered, specifically: r2 (r-squared), RMSE (Root Mean Squared Error) and error variance. From the analysis, it results that linear regression methods are not accurate enough and that the best methods, in terms of accuracy and error, are ensemble trees for regions and ensemble trees and Gaussian process regression for provinces.
The analysis of the most important features, aim of the study, reveals that at provincial level the most relevant are: number of inhabitants and number of S.A.R. for the regression of *new positive cases*; while, features as S.A.R., number of farmed animals are the most important for regressing *total cases*.
Concerning regions instead, the number of *deaths* seems to be explained especially by the number of nursing homes and mobility. These two features are also the most relevant for the *cumulative number of diseased*; finally, the regression of *new positive cases* revealed as the most important features the nursing homes and S.A.R. .
To sum up, for regions the mobility resulted as a feature promoting virus diffusion both in terms of deaths and cumulative cases of Covid-19; hence the government measures to monitor and limit mobility resulted efficient. Moreover, nursing homes feature was also expected to be relevant for the virus diffusion since many diseased and dead were elderly from nursing homes. However, its relevance is high only at regional level.
On the other hand, the meteorological data (temperature, wind, rain,..) available only for provinces, did not show a good relevance with the virus diffusion within this analysis, as many tried to claim especially at the beginning of the epidemic. The relevance that S.A.R. and number of farmed animals have on the virus diffusion over the provinces is interesting and it might be worth analyzing more in-depth.

Although the obtained results, it has to be pointed out that the COVID-19 pandemic situation can rapidly change and that the results obtained with this thesis are strictly bounded to the observation period of the case study previously specified.
However, the results of this analysis can be seen as a starting point for further investigations for those features that appear to be the most related to virus diffusion in Italy.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANAS Azienda Nazionale Autonoma delle Strade

Business enterprise owning thousands of kilometers of streets and highways; it joined 'Gruppo Ferrovie dello Stato' which is a national enterprise that operates in the public and private transport field.

ARD Automatic Relevance Determination

CART Classification And Regression Trees

Three-based method that can be used for both classification and regression problems.

ENAC Ente Nazionale per l'Aviazione Civile

Italian authority for technical regulamentation in the field of civil aviation.

GPR Gaussian Process Regression

Nonparametric Bayesian approach to regression

LLS Linear Least Squares

Linear regression technique

r2 coefficient of determination also known as r-squared

Metric explaining the goodness of a regression. Also used in telecommunication to evaluate the reconstruction of the received signal

WHO World Health Organization

1 INTRODUCTION

On 31st December 2019 Hubei, a province situated in China, informs World Health Organization (WHO) of a series of not identified pneumonia cases in its area. Only some time later, on 9th January 2020 China declares that the agent of the respiratory problems is a new virus identified as SARS-CoV-2 which causes the disease named COVID-19.
The name SARS-CoV-2 stands for Severe Acute Respiratory Syndrome - CoronaVirus 2 and indicates the membership of the new virus to the Coronaviridae family. Other viruses belonging to this family are those causing the normal cold, or the Middel East Respiratory Syndrome (MERS). It is named CoV-2 because it is the second virus after the SARS-CoV (SARS Corona Virus) discovered in the 2002, which caused an epidemic in the East area of the Globe, roping in several countries (37) especially in Chinese territory and causing the death of 744 people.
The symptoms of COVID-19 disease are similar to the normal flu: cold, fever, fatigue. Other symptoms can include shortness of breath or difficulty in breathing, muscle aches, chills, sore throat, runny nose, headache or chest pain. As the diffusion of the infection has evolved and reports on clinical case histories have accumulated, a new symptom began to emerge peculiar of Sars-CoV-2: the partial or total loss of the sense of smell and taste
Since the symptoms appear after an incubation period of 14 days, the person does not know to be ill, and can unintentionally infect other people that get in contact with him. This is a key element making even harder the disease containment.
COVID-19 can cause a wide range of manifestation of varying gravity in people. Death risk generally increases with the person age but it also depends on the patients pre-existent health conditions and diseases such as serious heart diseases (heart failure, coronary artery disease or cardiomyopathy), cancer, chronic obstructive pulmonary disease (COPD), type 2 diabetes, severe obesity, chronic kidney disease, sickle cell disease or weakened immune system from solid organ transplants.
The virus is highly transmissible and since it is a virus and not a bacteria, it cannot be fought with antibiotics. The transmission has been confirmed to come only from man to man and not trough animals. Its origin is still uncertain but it is believed to be bounded to the fish market in Wuhan, which calls thousands of visitors, and this would explain the fast transmission to a huge amount of people [33].
On 23rd January 2020 in fact, the central government of China imposed a lockdown to the city of Wuhan in order to limit the virus spread. Also other Chi-

nese cities employed the lockdown as control measure and at the end, the lockdown situation affected 57 million people [17]. Meanwhile, also many countries cancelled all the flights for and from China, in order to avoid importing the virus that could spread in other zones.
This is the case of Italy. However, somehow the disease fastly spread out in the Italian territory. At the beginning the situation appeared to be limited in some cities, but it precipitously collapsed forcing the government to firstly close the most affected provinces, and then the entire country (9th March), following the example of China. Some time later, on 20th March 2020 the WHO stated the global emergency and the COVID-19 disease pandemic was officially declared. Until 17th April 2020 the pandemic affected 219 areas among countries and territories all over the world: Africa, Americas, Eastern Mediterranean, Europe, South East Asia and Western pacific. The outbreak situation up to that date is explained by WHO with the following image (fig. 1):

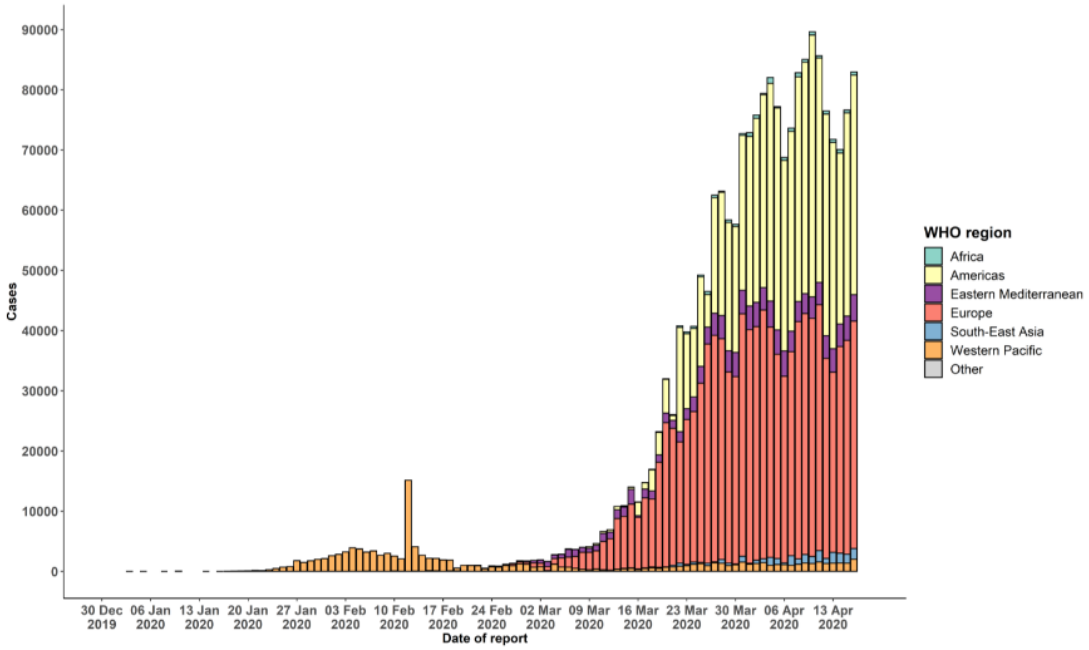


Figure 1: Epidemic curve of confirmed COVID-19 cases by date of report and WHO region through 17th April 2020. Each bar represents the number of daily cases [14].

As it can be noticed, the outbreak in the Western Pacific areas begun to be reported to the WHO in the week between 13th and 20th January 2020; until 13th April, the peak of the virus diffusion occurred on February 12 2020, counting roughly more than 15000 diseased only on that day. When the number of infected started decreasing for Western pacific zone, the outbreak started for all the other regions and in far greater numbers: America and Africa result to be the most affected territories counting about 90000 cases in a day (10th April), followed by Eastern Mediterranean and Europe.

The general answer to the ongoing outbreak has been the limitation of mobility, in order to avoid crowded space where the virus could be spread. On 11st April 2020, which has been the day with the highest daily contagious (according to fig. 1), 167 countries, territories and areas have implemented additional health measures mostly interfering with international traffic. An updated summary is outlined in the 'Subject in Focus' in the WHO report [15].
For what concerns the Italian situation, fig. 2 shows the number of daily reported cases and cumulative deaths from the beginning of the pandemic (24th February) until 17th April 2020:

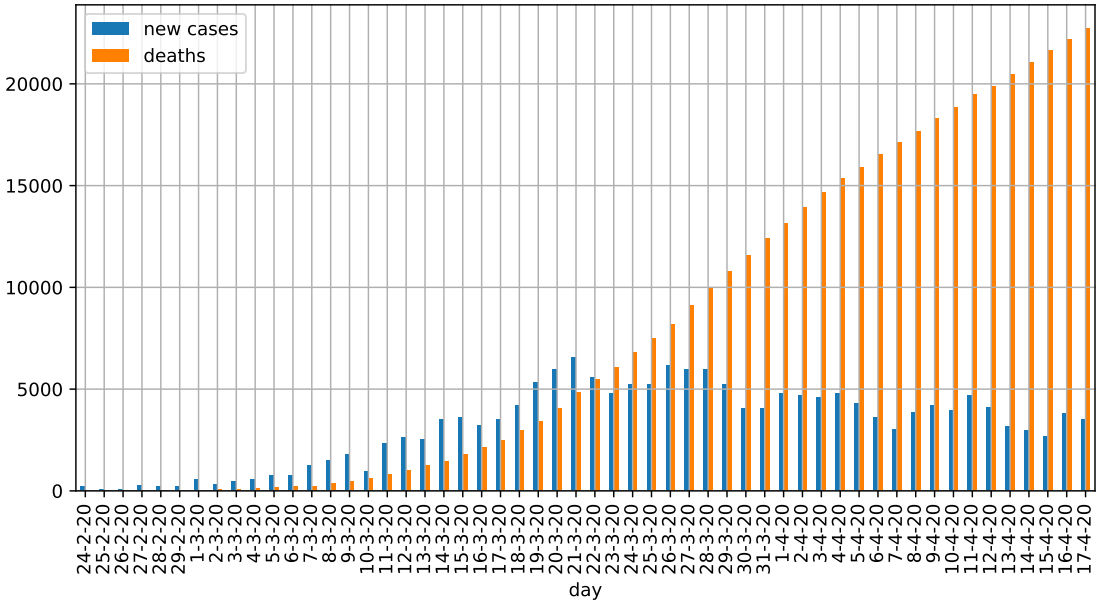


Figure 2: Epidemic curve of confirmed dead (orange) by COVID-19 and daily new cases (blue) in Italy through 17th April 2020

The peak of the daily new cases curve has been reached on 21th March 2020, counting 6557 diseased. Until the 17th April there have been 22745 deaths.
The Italian case is interesting because the virus spread differently over the country. The most affected Italian regions have been the Northern ones and the less affected have been the Southern's; the Central regions instead have been moderately affected. Focusing on the three main areas of the country, the Italian situation until the 17th April is described by fig. 3 (Northern Italy), fig. 4 (Central Italy) and fig. 5 (Southern Italy):

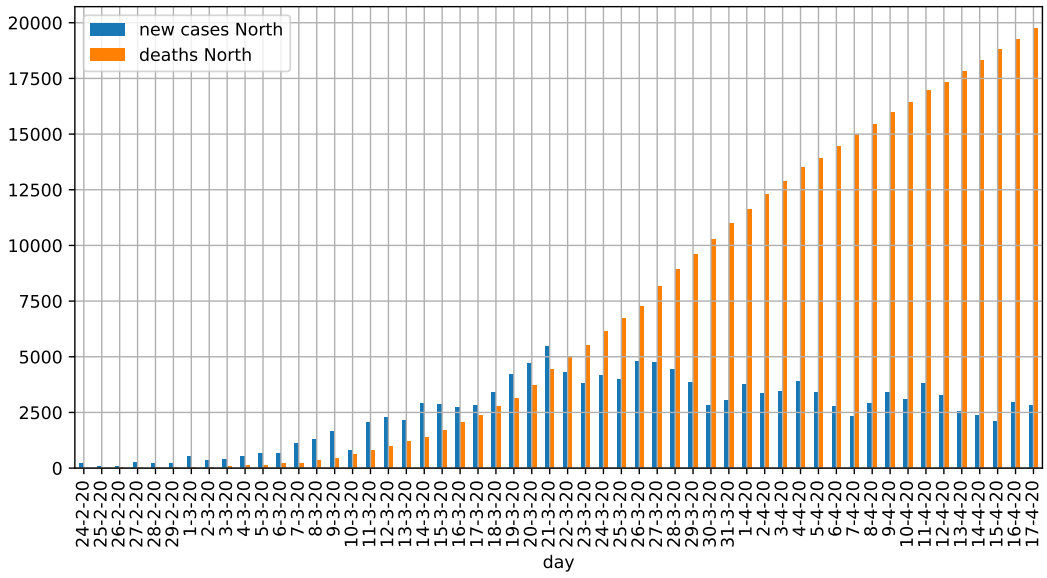


Figure 3: Epidemic curves for Northern Italian regions (Valle d'Aosta, Liguria, Piemonte, Lombardia, Friuli Venezia Giulia, Veneto, Emilia-Romagna, P.A. Bolzano and P.A. Trento)

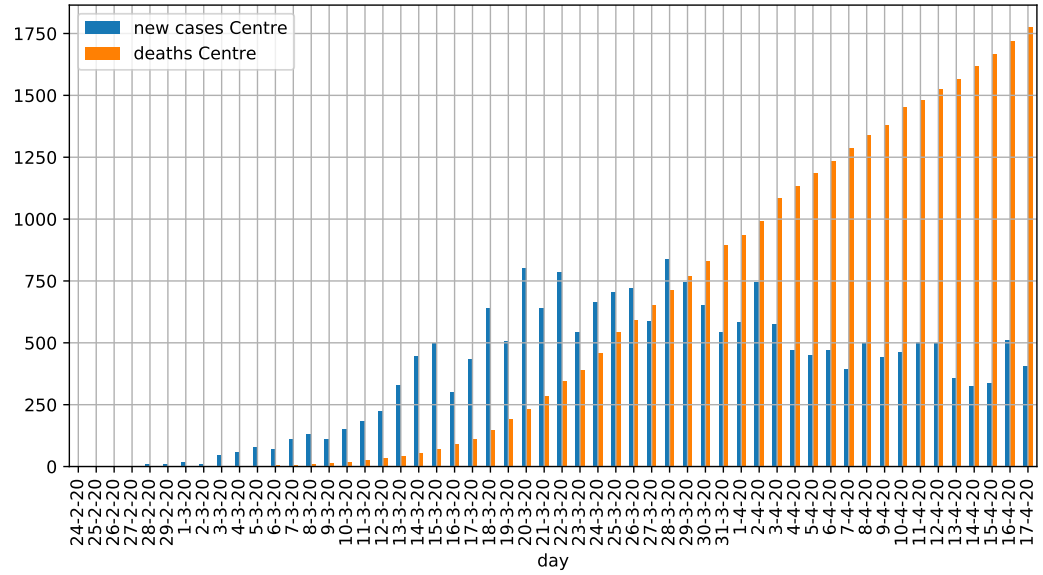


Figure 4: Epidemic curves for Central Italian regions (Lazio, Marche, Umbria and Toscana)

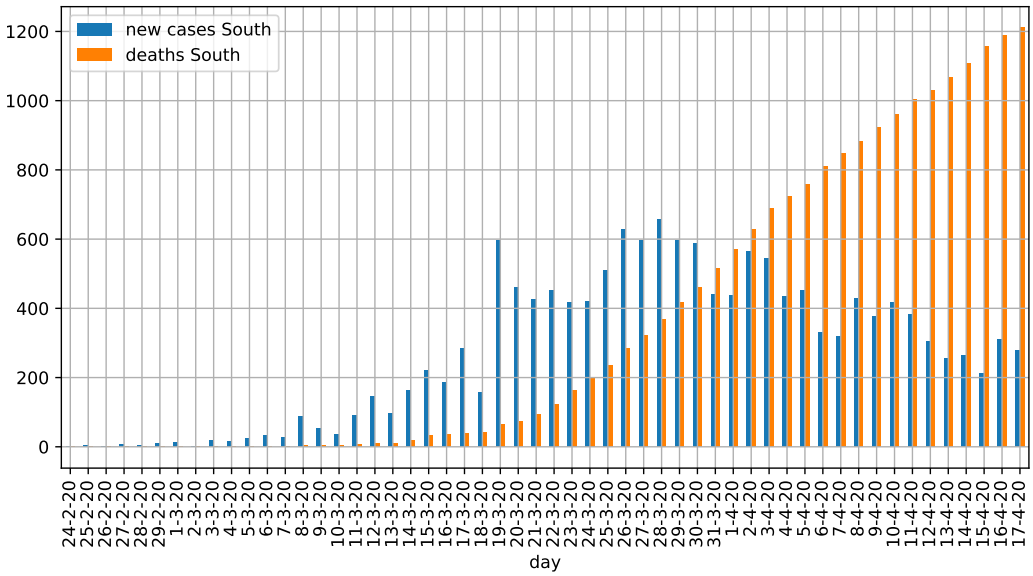


Figure 5: Epidemic curves for Southern Italian regions (Campania, Abruzzo, Calabria, Sicilia, Sardegna, Puglia, Basilicata and Molise)

It is clear that the virus affected the Northern Italian population more harshly than the two other Italian areas. In fact, both the numbers of deaths and the total cases are significantly greater.
This can be better appreciated in fig. 6 that reports the cumulative cases (both deaths and diseased) for the three areas:

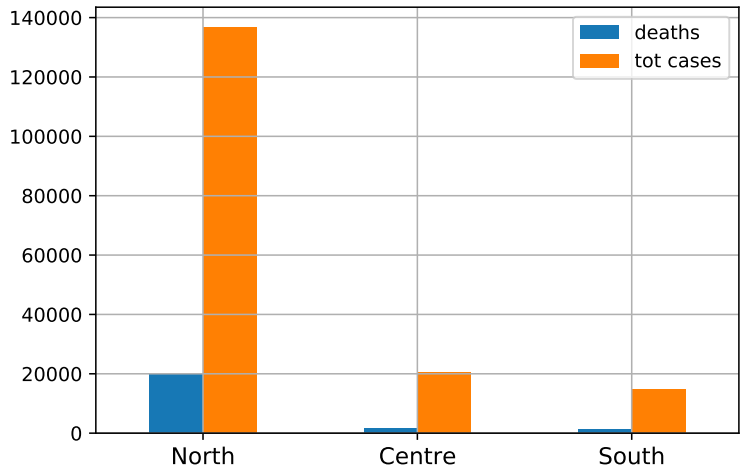


Figure 6: Cumulative number of cases and deaths for COVID-19 through 17 April 2020 in Northern, Central and Southern Italy

It is interesting to notice that the number of cases in Northern of Italy is almost nine time greater than the number of diseased in the Southern Italy. Moreover, since the population is 27774970 in the Northern regions (data referred to 2019), 11986958 in the Central and 20482711 in the Southern territory, the percentage of the people that got ill has been approximately: 5% of the

population in the North, 0.17% in the Centre and 0.07% in the South. Finally, among the total cases, the 14% died in the North, the 8.6% in the Centre and the 8.2% in the South.

The clear discrepancy about the virus diffusion among the three main Italian areas raised up several questions about why such a difference, what made the virus more diffused in the Northern Italy.

Behind this study there is the idea to try answering this questions by researching, among some selected features, those that could have caused a more severe virus diffusion in the North and a more thigh one in the Centre and South of Italy.

2 STATE OF THE ART

Since the beginning of the COVID-19 pandemic, the disease captured the attention of all sort of scientists. After ten months from the discovery and although the high effort of biologists, virologists and other specialists, the virus is still fairly unknown and it seems to behave randomly.
Machine Learning is one of the sciences exploited in the fight against the virus spread. In fact, the government scientist crew has used it since the first weeks of the pandemic to predict the infection curve, which explains the number of diseased people over time. This curve is used to predict the spread of the virus over a region and the numbers are used to understand the medical effort needed to treat ill people. The infection curve has a bell shape: starting from zero, it increases exponentially as people get infected; after a certain time, the curve begins decreasing because the hypothesis is that a person that gets ill can die or recover becoming immune to the illness, and in both cases cannot infect others - hence with the time passing the 'herd immunity' phenomena occurs. Because the COVID-19 forces people to hospital treatments in most cases, it is important that the pick of the curve, which is the maximum number of simultaneous infected people, stays below the maximum capacity of the hospitals, otherwise, health crisis occurs, having many people that cannot be helped in the proper way.
The infection curve is also obtained considering some boundary conditions. For instance, the government simulated the infection curve by adding the reduction of mobility to study the effect on the number of infectious [16]. The curve is obtained with the application of the *SEIR* model which is not a novelty in modeling and forecasting epidemic phenomena. It has been widely used to model the spread of infective viruses such as in the case of Ebola ([5]), Zika ([37]) and also for other coronaviruses such as SARS ([2]) and MERS ([6]). SEIR models have several versions depending on the disease itself and also on the hypothesis. The scientists try to adapt the model to the current epidemic to obtain more realistic curves. An example is a study carried out by A. Godio et al ([47]) that improves the SEIR model for the case of Italian regions of Piemonte and Veneto by adding a stochastic approach (the Particle Swarm Optimization solver) to improve the reliability of predictions in a period of 30 days.
SEIR models are used not only to predict the infection curve over time but also to estimate the R_0 coefficient, known as *Reproduction Number*. This coefficient tells about the number of people that an ill person can infect on average. The R_0 is revaluated over time because its value continuously changes: if it is under 1 and keeps decreasing it means that the epidemic is vanishing.
During the Sars-Cov-2 pandemic, machine learning has been largely exploited

not only by employing mathematical models but also in the field of deep learning for imaging classification. Several studies ([18],[25]), have used neural networks to distinguish a Sars-Cov-2 infected person from other known pneumonia cases, using chest X-ray imaging. Another study in particular obtained the 99% accuracy in the classification by utilizing the DenseNet121 neural network feature extraction with Bagging Tree classifier [28]. These researches are notably interesting because the automatic classification could be an effective alternative to other testes (serological and swabs), helping to recognize a COVID-19 case and immediately address him to the proper treatment without risking other infections.
Similarly to the previous classification problems, another research used neural networks based classifier to diagnose COVID-19 people by analyzing their respiratory pattern [3].

An additional machine learning application is the therapeutical one. The proper cure to this disease has not been found yet and there is a certain hurry to find the right antiviral treatment because of the high impact the virus has on the worldwide population and the number of deaths it is causing. On the other hand, traditional drug research requires time, especially due to the complexity of drug design and clinical trial protocols. Machine learning can also help in this regard. In [42] Sovesh Mahapatra et al. exploit a model based on the Naive Bayes algorithm that has an accuracy of the 73% to predict the drugs - among the ones already available on the market- that can be used for the SARS-CoV-2 treatment.
Similarly, Yiyue Ge et al. in [1] used the past SARS-CoV and MERS-CoV data to demonstrate that their model - which is based on both machine learning and statistical analysis- successfully predicts effective drug candidates against a specific coronavirus. Being already used in other diseases such as epilepsy [24], or for the prediction of the cancer immunotherapy response [39], this kind of machine learning application is not a novelty.
The primary objective for most of the studies is hence to develop predictive, diagnosis and therapeutical models, whose primary goal is to address the hospital resources for the best or accelerate the drug research in the case of the therapeutical application.
Besides those kinds of studies, researchers are attempting to find the causes behind the virus spread among different countries. Many regions have been only scarcely interested by the virus diffusion, many others instead, were drastically affected by it.
Many scientists studied the case of Italy, being, in the first period of the worldwide virus spread, the second most affected country after China. Moreover, the case of Italy raised interest because the diffusion of the virus has not been homogeneous over the country itself, distinguishing the gravity of the infection diffusion in North, South and Center of the country. Many experts indicated, among the theories, the pollution as the first cause. The areas counting more cases of COVID-19 in fact are also the most polluted ones. The reason should

lie in the fact that a long-lasting exposition to the pollution is the first cause of several health issues, that make people more susceptible to respiratory infections [49].
A study made in Harvard University focused on the United State territory, pointed out that an increase of $1\mu g/m^3$ in $PM_{2.5}$ is associated with an increase of 8% of mortality of COVID-19 [21].
Other theories take into account the role of the sunlight UV-rays, that annihilate the virus infectious power, in the same way that the UV-rays are used to activate or deactivate some molecules by changing the genetical material (DNA or RNA). The study proposed by Shanna Ratnesar-Shumate et al. in *Simulated Sunlight Rapidly Inactivates SARS-CoV-2 on Surfaces* [19] pointed out that the virus in simulated saliva exposed to the simulated sunlight is inactivated for the ninety percent every 6.8 minutes, while the virus present in culture media is deactivated each 14.3 minutes.
Many other experts instead, especially in the field of health, try to find the cause of the virus diffusion in the pre-existing health condition of the afflicted population. Since the beginning of the epidemic, the highest ranks of public health, such as *Epicentro* at the Italian level, constantly update the demographic information regarding the people affected by the virus [9]. That news tells about the mean age of the ill people, the percentage of people with a certain number of pre-existing health problems such as obesity or chronic disease.
All these information and hypothetical causes such as pollution, UV-sunlight and its effect, pre-existing disease, mean age, and also the lifestyle that people conduct, are typical of a limited territory. For instance, also the incidence of a particular disease is often circumscribed, differing from region to region.
Although the studies made focused on a particular hypothetical cause of the virus spread, there are not researches focusing in understand the overall features that could be related to virus diffusion. The aim of this study is precisely this one: exploiting machine learning algorithms in the attempt to understand the features that mostly contributed to the virus spread considering limited areas.
The features selected for this study are chosen both by listening news and reading articles -e.g. the percentage of uncovered sky, taking inspiration by the UV theory- and both by intuition -e.g. the rain, which is thought to contribute in limiting the virus spread. Many data that were meant to be included -for instance the map of some pre-existing chronic disease, data about the eating habits, other data about gathering centers and flow of people and the pollution itself- but there was not the possibility to collect them. And this is also the limitation of this thesis.

3 REGRESSION ALGORITHMS REVIEW

One of the approach for feature selection is exploiting regression techniques that output feature relevance. Regression methods in fact, are the ones used to cope with the goal of this thesis.
The algorithms of concern in this specific context are Linear regression (LLS, Lasso and Ridge), ensemble trees (Random forest and Extra trees) and Gaussian process regression. The following section will cover the theoretical part of the used algorithms.

3.1 LINEAR REGRESSION

Linear regression is the widest used machine learning technique that exploits the linear relation among the target variable -the one to be estimated which is also called **regressand** and is the dependent variable- and the observed features - which are called **regressors** and are the independent variables.
In this context the regressand is the number of infected people or the deaths, while the regressors are province or region characteristics, such as the population density, the population mean age, etc. Given N observations (a set of cities with the corresponding target variable) and F features, the linear regression problem finds how much a feature 'weighs' on the estimation of the regressand.
Once found the weights for each feature, they can be interpreted as feature importance for the algorithm.
Since the dependent variable can be explained by the independent variables multiplied by a certain weight (to be found), plus a random measurement error, it can be written:

$$y = Xw + \epsilon$$

where y indicates the observations and it is a column vector ϵ $\mathbb{R}^N$ where N is the number of observations, the matrix X represents the input data, so the value of the features for each observation and X $\epsilon \mathbb{R}^{NxF}$ where F is the number of features.
To find the weights, the linear regression algorithm aims to minimize the error of the estimation of y starting from the observations X; this can be written as:

$$f(w) = (y - \mathbf{X}w) \tag{1}$$

Different regression techniques find different values of weights because the problem they solve, hence, the cost function to minimize (eq. (1)) is different. In the Linear Least Squares (LLS) problem, the cost function is represented by:

$$\min_{w} || y - Xw ||^2 \tag{2}$$

This minimization problem is simply solved by finding the gradient of the function $f(w)$, setting it equal to 0 and finding w. The final solution is then given by:

$$f(w) = y^T y - yTX(X^T X)^{-1} X^T y \tag{3}$$

Sometimes the LSS can be affected by overfitting. This happens when some features assume very large values so that **w** takes large values too and overfitting occurs because the model is trying to model also the noise. Some regularization that increases the LSS stability exists and these are ridge and Lasso regression. Both add a simple term to the LLS cost function such that it is enough to reduce model complexity and prevent from overfitting [8] eq. (4):

$$\min_{w} || y - Xw ||^2 + \lambda || w ||^q \tag{4}$$

That is equal to write:

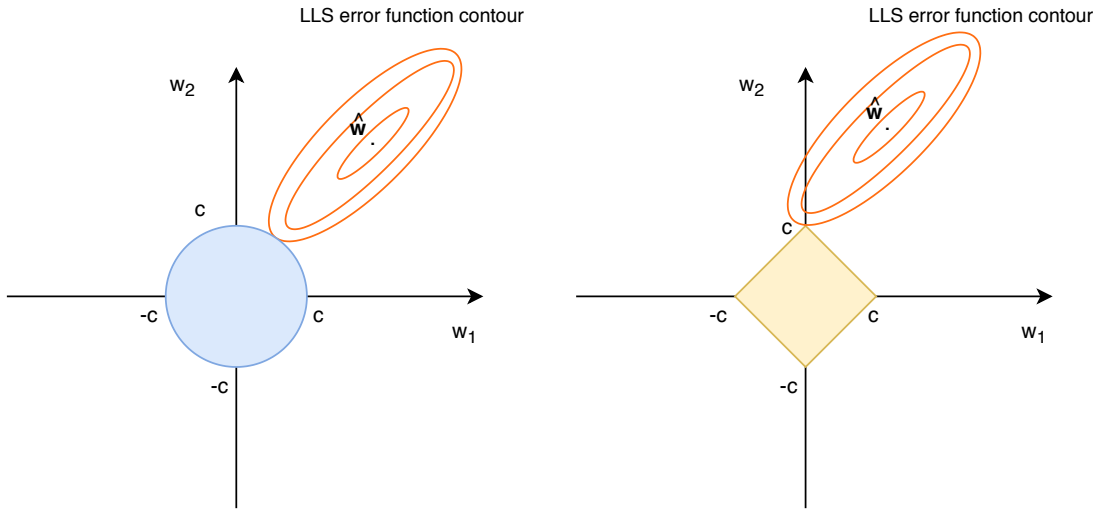
$$\min_{w} || y - Xw ||^2 \text{ s.t. } \sum_{i=0}^{F} | w_i^q | < c, \text{ for } c > 0 \tag{5}$$

According to the value of q the above equation eq. (4) becomes Lasso or ridge regression. When $q = 1$ eq. (4) is also called **L1** regularization and it is Lasso regression; instead when $q = 2$ eq. (4) is also called **L2** regularization and it is ridge regression. So in the case of Lasso regression, the LLS function is optimized by adding a penalty term that considers the magnitude of the weights; with ridge instead, the square magnitude of the weight is considered. The term λ is a positive scalar and its value is to be found with trial and error procedure. In general, it is important to notice that when λ is equal to 0, one is again in the case of LLS regression for both L1 and L2 regularization.
Last thing to say is that Lasso regression can be used for feature selection [8], [50]. Focusing on the penalty term in eq. (5) and imaging to be in a two-dimensional problem, hence when $F = 2$, the constraints for Ridge and Lasso become, respectively eq. (6):

$$\begin{aligned} &\text{(Ridge)} \quad | w_1 |^2 + | w_2 |^2 < c, \\ &\text{(Lasso)} \quad | w_1 | + | w_2 | < c \end{aligned} \tag{6}$$

where the first is the expression of a circle and the second is the expression of a diamond. Considering the contour of the LLS error function (a *contour* are the ensemble of points where the function has the same value), where the center corresponds to the LLS solution (w that minimizes the error). The estimate of

Ridge and Lasso is simply given by the points in which the contour lines hit the diamond (Lasso) and circle (Ridge) regions. The fact is that the diamond has corners while the circle has not. If the solution occurs in the corner, one weight has a non zero value, while the other is not. In a feature space greater than 2, the occurrence of the corner with the contour line of the LLS function can happen more easily. This means that there is a greater possibility that more parameters go to 0. From here, the feature selection. Figure 7 shows the example with 2D feature space.



(a) Ridge constriant region **(b)** Lasso constraint region vector

Figure 7: Lasso and Ridge penalty function regions. The curves in red are the contours of the LLS error function. Image from "*The Elements of Statistical Learning*" [50].

The advantage to use liner regression is given by its simplicity. However, the supposition about the linear relation between the regressand and regressors is sometimes very harsh. Nevertheless, these algorithms are usually the first to be applied in problems where the relation between the features and observation is to be guessed.

3.2 REGRESSION TREES

The Trees methods belong to the family of non-linear and non-parametric learning algorithms. One of the advantages to use Trees is the interpretability of their results. They natively include the 'feature importance' attribute that allows defining which characteristic about the input data is relevant for the model.

There are many tree-based methods, differing for some decision rules. The most famous ones are Classification And Regression Trees (CART) and C4.5. In this context the focus is given to the CART method.

Given a dataset of N observations and F inputs for each observation, the regression trees work in a manner that the feature space is iteratively divided into multiple regions and for each region, a constant value is assigned. At the end of the building phase, the constant value for each region corresponds to the prediction for an observation given its inputs X. More in detail, imagining to have some observations and a feature space of 2 inputs x_1 and x_2 such that $x_i \epsilon \mathbb{I}$, where $\mathbb{I}$ indicates a finite fixed input dimension. First, the feature space is split into two regions R_1 and R_2 and constant value is assigned to each region to model the response y. The splitting rule is based on finding first the split feature j and then the splitting point s to achieve the best fit. The splitting process is repeated for each region and subregions until some stop condition is met. This method is called *recursive binary splitting*. In the end, there will be a number M of regions and for each of them, a constant value c_m is assigned for the response variable y such that:

$$f(x) = \sum_{m=1}^{M} c_m I(x \ \epsilon \ R_m) \tag{7}$$

Where $f(x)$ in eq. (7) represents the modelled continuous response variable *, as a constant c_m for each region R_m. The best c_m is found as the average of y_i in region R_m eq. (8):

$$\hat{c}_m = avg(y_i \mid x_i \epsilon R_m) \tag{8}$$

For each split i a region R_i is split in R_{i1} and R_{i2} given a splitting feature f and a splitting point t; the half-planes R_{i1} and R_{i2} are defined as:

$$R_{i1}(f,t) = \{X \mid X_f < t\} \text{ and } R_{i2}(f,t) = \{X \mid X_f > t\} \tag{9}$$

Where X represents the entire features space and X_f represents the feature f of the feature space X.
That said, the feature f for the splitting and its threshold point t are found solving the greedy optimization eq. (10)

$$\min_{f,t} [\min_{c_{i1}} \sum_{x_i \epsilon R_{i1}(f,t)} (y_i - c_{i1})^2 + \min_{c_{i2}} \sum_{x_i \epsilon R_{i2}(f,t)} (y_i - c_{i2})^2] \tag{10}$$

Where the inner minimizations are solved by using eq. (8) for both regions:

$$\hat{c_{i1}} = avg(y_i \mid x_i \ \epsilon \ R_{i1}) \text{ and } \hat{c_{i2}} = avg(y_i \mid x_i \ \epsilon \ R_{i2}) \tag{11}$$

So intuitively, the CART algorithm the splits are chosen to minimize the residual sum of squares between the observation and the mean in each region. By repeating this procedure for all the subregions, the final tree is built.

* Recall that for regression problems the response variable is continuous and thus can be modelled as a continuous function, in fact, differently that classification problems, regressions' can take whichever value.

At the end of the procedure, the algorithm outputs also the *feature importance* which is a weighted sum over the features that takes in consideration how many time a feature was selected for the best split, being the one that brings more information for the regression.

The greatest problem with regression trees is their instability due to the high variance. In fact, even a small change in the dataset can be reflected in a completely different tree at the end of the learning process. This is mainly explicable with the fact that an error at the top has an impact on the entire tree construction. In order to reduce variance and not sacrificing the use of trees, the *ensemble* learning can be used [50].

Ensemble method refers to a procedure in which instead of relying on one single model, a bunch of models is trained, so that, if the first reached a weak performance, the predictive performance of the ensemble is much more strengthened. Ensemble learning is defined *homogeneous* if the base model is the same, or *heterogeneous* if the models composing the ensemble are different. Moreover, two types of ensemble learning exist: sequential ensemble and parallel ensemble. When using the former, one model in the ensemble is trained at time and becomes the starting point of the next model, this exploits the dependence of the base models; when using the latter instead, many models are trained at the same time, this exploits the independence of the base models [23].

Bagging is an example of ensemble learning. It is used for decreasing the variance of a model. **Bagging** (**bootstrap aggregation**) is based on **bootstrapping**, which refers to a statistical technique for the resampling of the data to train with. With bagging, starting from a dataset, many subsets are created and a model for each of them is trained. These subsets are also called bootstrap samples. Once the bootstrap models are trained, the outputs of the base learners' models are averaged. This technique is mostly used for unstable models, such as decision trees. Examples of bagging model are **Random Forest** and **Extra trees**.

Random Forest works as follow: many bootstrap samples are retrieved from the original dataset and for each of them a decision tree is trained. The bootstrap samples are built by randomly choosing a subset of features. The final prediction is drawn from the average of the outputs of all trees. It is called Random forest because a set of random trees is built.

Extra Trees (or Extreme Random Forest) operates in a similar way as the random forest, but it adds more randomness: the splitting threshold for all the features is picked randomly instead of being searched to minimize the node (region) impurity †. This is why it is called **extreme** random forest (extremely randomized forest).

The randomness introduced with this algorithm reduces variance. Extra Trees that bring even more randomness with respect to Random Forest, in fact, presents usually less variance, and both are much better than a single decision tree.

† The *impurity* measures the heterogeneity of the labels

3.3 GAUSSIAN PROCESS REGRESSION

‡ The Gaussian Process Regression (GPR) is a non-parametric Bayesian approach algorithm, hence, the regression output is not a numerical value but a probability distribution of all possible values.
By applying the Bayes approach to the linear regression problem, the parameter w is not estimated as a numeric value but as a probability distribution of the possible values. This is done through the Baye's rule eq. (12):

$$p(w \mid y, X) = \frac{p(y \mid X, w)p(w)}{p(y \mid X)}, \tag{12}$$

Where $p(w \mid y, X)$ is the posterior probability of parameter w given the observation y and inputs X, $p(w)$ is the prior distribution of w, that express the belief of data before observing them, $p(y \mid X, w)$ is the likelihood probability deducted from the dataset and $p(y \mid X)$ is the marginal likelihood.
In the case of GPR the prior distribution over the parameters w is 0 mean and covariance matrix Σ_p eq. (13):

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \Sigma_p\right), \tag{13}$$

The posterior distribution is then used to predict the distribution of new unknown given new input x^* eq. (14):

$$p\left(f_* \mid \mathbf{x}_*, X, \mathbf{y}\right) = \int p\left(f_* \mid \mathbf{x}_*, \mathbf{w}\right) p(\mathbf{w} \mid X, \mathbf{y}) d\mathbf{w} \tag{14}$$

Where (f^* represents the predicted distribution given the new input x^*, the old input (training data) X and the target vector y.
The predictive distribution is again Gaussian with a mean given by the posterior mean of weight multiplied by test input, while the variance is a quadratic form of the test input with the posterior covariance matrix; this leads to infer that the predictive uncertainties grow with the magnitude of the test input.
This is the idea behind the Bayesian approach. In the case of GPR, the only thing to do is specify the prior distribution of the data, which is assumed to be Gaussian. Since a Gaussian process $f(x)$ is completely described by its mean $m(x)$ and covariance $k(x, x')$ described, it can be written as the following:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{15}$$

where, the mean and the covariance are defined as:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k\left(\mathbf{x}, \mathbf{x}'\right) &= \mathbb{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f\left(\mathbf{x}'\right) - m\left(\mathbf{x}'\right)\right)\right] \end{aligned} \tag{16}$$

‡ Most of the equations are from [51] and [34]

The covariance function is also called *kernel*. It specifies the covariance between pairs of random variables:

$$\operatorname{cov}\left(f\left(\mathbf{x}_p\right), f\left(\mathbf{x}_q\right)\right) = k\left(\mathbf{x}_p, \mathbf{x}_q\right) \tag{17}$$

The covariance function of the outputs is evaluated between the inputs data. In the case of regression, the covariance function of the training target data is evaluated as a function of training features data.
Many covariance functions exist and one has to select the most appropriate one for the data that is handling.
A general covariance function can be expressed as the following (eq. (18)):

$$K\left(\mathrm{x}, \mathrm{x}'; \theta\right) = \theta_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{I}\frac{\left(x_i - x_i'\right)^2}{l_i^2}\right] + \theta_2 \tag{18}$$

the term θ_1 indicates the elongation of the function along the vertical axis; the term θ_2 indicates the offset of the function from the 0; the term l_i is the *lengthscale* and indicates how much the function y, which is the expected value of the target variable, is fluctuating along with the space of feature x_i. It indicates the relative sensitivity of the model to change when the feature value changes. Low fluctuations (large values of l_i) means that y does not change much for high variation of x_i, hence, the feature x_i is irrelevant for the model. On the contrary, frequent fluctuation (low values of l_i) means that the y is particularly affected by the changing value of x_i [34].

The kernel functions can be distinguished in stationary kernel and non-stationary kernel. The former assumes that the function depends only on the distance of data and not on their absolute value. It can furthermore be divided between isotropic, where data are also invariant to rotation in the input space, and anisotropic kernel. When the kernel is isotropic, all the features are assumed to be as well relevant in the model fitting, and the lengthscale in the covariance function is a scalar. Instead, when the kernel is anisotropic, the term lengthscale in the covariance function is a vector whose length is equal to the number of features. This allows the model to look at the lengthscale of each feature with respect to the expected outcome of y. By imposing an anisotropic kernel, the Automatic Relevance Determination (Automatic Relevance Determination (ARD)) is applied, and this allows to analyze, after training, the feature relevance.
In the context of this thesis, the covariance function exploited is the RBF (Radial Basis Function, also known as 'squared-exponential' or 'Gaussian' kernel), and the Matérn kernel because they resulted to be the best fitting ones. The RBF is defined as (eq. (19) from [46]):

$$K\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}'\right\|^2}{2l^2}\right) \tag{19}$$

Note that the numerator of the exponential is the Euclidean distance of the two variables x' and x. This kernel is infinitely differentiable, which implies

that Gaussian processes with this kernel as a covariance function are very smooth.

The Matérn kernel is defined as (eq. (20) from [45]):

$$k\left(x_i, x_j\right) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right) \tag{20}$$

The Matérn kernel is a generalization of the RBF. The parameter ν regulates the smoothness of the covariance function: the higher the value, the smoother the function. For a very large value of ν ($\nu \to \infty$), the Matérn kernel is equivalent to the RBF kernel. $\nu = 1.5$ and $\nu = 2.5$ are two important values that indicate that the function is (respectively) once or twice differentiable.
The function $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is a modified Bessel function.
The GPR is an extremely powerful machine learning tool which only requires the specification of the covariance function. Another advantage is that different than other algorithms, it provides itsown measurement uncertainty on the predictions.

4 RESULTS

This chapter reports the results obtained with the supervised learning algorithms introduced in the previous chapter (chapter 3). The regressions techniques are applied twice: for provinces and regions. The datasets used in the two cases are different due to the unavailability of some data; hence most of the features are not the same in the two geographical levels.
Details about the features composing the datasets are given in the proper sections (section 4.1 reporting regions results and section 4.2 wiring provinces results), where also the data sources are given.
Sci-kit learn library [44] available for python, is the tool exploited for the implementation of the regression methods. Before applying the regression, the data have been pre-processed. First thing, the dataset was cleaned by dropping all the non-valid entries (such as missing data). The entire dataset (hence both the data used for training and testing) then was standardized so that all the features values had a 0 mean and standard deviation of 1. This technique prevents errors in the prediction of features belonging to different lengthscale. For instance, by standardizing, the features that take large value -e.g. the number of people living in a city- are not taken as more important than features that instead have lower values - e.g. the mean age of a population. Standardization of features allows also to compare the results of different algorithms, in terms of feature importance. In the end, the dataset was shuffled.
After this procedure, the dataset was divided into training and test set. The training set was in any case chosen to be 70% of the total. Since the data was shuffled, the training and test sets contain both recent and less recent observations.
The results concerning the regions are presented in section 4.1 while those related to province level are presented in section 4.2.
Results include a true vs predicted scatter plot, in which, other than the scatter points, two lines are present: one is the perfect prediction line (bisector of the Cartesian plane) and the other is the best fit line. The latter is a regression line drawn starting from the scatter plot. It is obtained by exploiting the *numpy* library; in a few words it is obtained by evaluating the *slope* and the *intercept* of a line that minimizes the OLS (Ordinary Least Squares) error of all the points. Once the slope m and the intercept b are found, the best fit line is simply derived as: $y = m * x + b$. The closer the best fit line is to the perfect prediction line, the better the regression. Flanked to this regression figure, there is another one concerning the feature importance (ensemble trees) or the weighted vector (linear regression) or feature relevance (GPR), depending on the algorithm.

Furthermore, in order to compare different algorithms RMSE (Root Mean Squared error) , coefficient of determination also known as r-squared (r2) and error variance are chosen as evaluation metrics.

- **RMSE** Root Mean Squared Error is defined as:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{N}}$$

  where $\hat{y}$ is the predicted value, y is the true value and N is the number of samples. Since the RMSE indicates the error in the prediction, the less is its value the better the model.

- **Variance**: defined as the average of the squared deviations from the mean,

$$var = \sigma^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \mu)^2}{N}$$

  where μ is the mean of the predicted values. The variance indicates how much the model predictions are spread far from the mean value. The lower the value, the more precise the estimation. A high value of variance is associated with overfitting problems, especially if the models show also low biases.

- **r2** is defined as:

$$r^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

  where SS_{tot} indicates the total sum of squares, which is proportional to the variance of the data, and SS_{res} is the square of residual (error in prediction). Moreover, y represents the true value of the target, $\hat{y}$ is the predicted value, $\bar{y}$ is the mean value among the true values. The metric r2 can be thought of as a metric that tells about how good a model fits the data. The highest value is 1 and is obtained when there is no residual, hence the model perfectly fits the data. Instead, values of r2 outside the range $[0, 1]$ can occur when the model fits the data worse than a horizontal hyperplane [10].

Finally, at the end of the regression of a specific target, there is an analysis of the most important features among the different algorithms. For this purpose, an image (for each target) is used: this illustrates, as a percentage, how many times a feature appears to be among the most important ones (among the first four) output by the regression methods. Additional information is given in specific sections.

4.1 RESULTS – REGIONAL LEVEL

4.1.1 Data acquisition

The data related to regions can be categorized in three groups: **Civil protection** data, **region characteristics** and **health-sanity** data.
Civil protection is a department of *Presidenza del Consiglio dei Ministri*, that operates with regions and autonomous provinces to coordinate national resources in order to guarantee assistance in case of severe emergencies [40]. This department in collaboration with regions and autonomous provinces, collects data about the Italian situation of COVID-19 cases and provides it through a public Github repository ([27]), updated every day. The data of interest include the following information (table 1):

Feature name	Name meaning
deaths	cumulative number of deaths
tot cases	cumulative number of positive cases
new positive	tot cases current day - tot cases previous day

Table 1: Civil protection data for regions

Data related to region characteristics are wide and include those listed in table 2:

Feature name	Name meaning	Data source
mobility	car traffic	[20], [11]
n ani farm	total number of animals farmed	[29]
n inhab	number of inhabitants	[41]
pop density	population density (inhabitants divided per squared kilometer)	[38]
income	average per-capita income	[12]
nurs home	total number of available beds in residential house for old people	[32]
S.A.R.	number of Serious Accident Risk (industries)	[35]
mean age	mean age	[13]
old index	index of oldness	[13]

Table 2: Region characteristics data.

Where the oldness index is defined as

$$\frac{\text{population over 65}}{\text{population under 14}} * 100$$

Going more in-depth about mobility, it has to be clear that the data are retrieved in this way: during the lockdown phase, Google companies made available daily data on a certain period, [11]) related to the mobility trend. In particular, the big company collected data about daily traffic aggregated per region by exploiting its users. What is available on the site is not the correct traffic number measure, but the difference in percentage with respect to a baseline, which is assumed to be the amount of traffic held the January 13th 2020. The data are in a tabular form whose most relevant headers are: 'regions', 'date' and 'percentage change from baseline'.
The date, at the moment of the download, went from February 15th to April 17th. On the other hand, data from the GitHub repository starts from February 26th and is updated every day. For this reason, the data that make up the dataset is from February 26th to April 17th.
Since Google did not provide the baseline of the traffic quantity, another solution was adopted. Azienda Nazionale Autonoma delle Strade (ANAS) website ([20]) makes available data on mobility utilizing the mean number of cars that cross a road in a day; the period of observation is the year 2018. ANAS provides the mobility information for the streets and highway that it holds, city by city. At the region level, the values of the mobility of the cities belonging to the specific region have been aggregated by sum. In this way, ANAS mobility data are considered the baseline to which the Google mobility trend refers to, and so, Google data are handled so that they are not a percentage variation but a number; in other words:

$$\text{mobility}_{\text{day i}} = \textit{google mobility} * \textit{baseline}(\%), \text{ baseline} = \text{ANAS mobility} \quad (21)$$

Data about health and sanitation are included in table 3:

Feature name	Name meaning	Data source
obese, underweight normal, overweight	percentage of people with a certain body constitution	[30]
health workers	number of health workers (medical staff)	[48]
nurses	number of nurses	[48]
doctors	number of doctors	[48]

Table 3: Health and sanitation industry in regions.

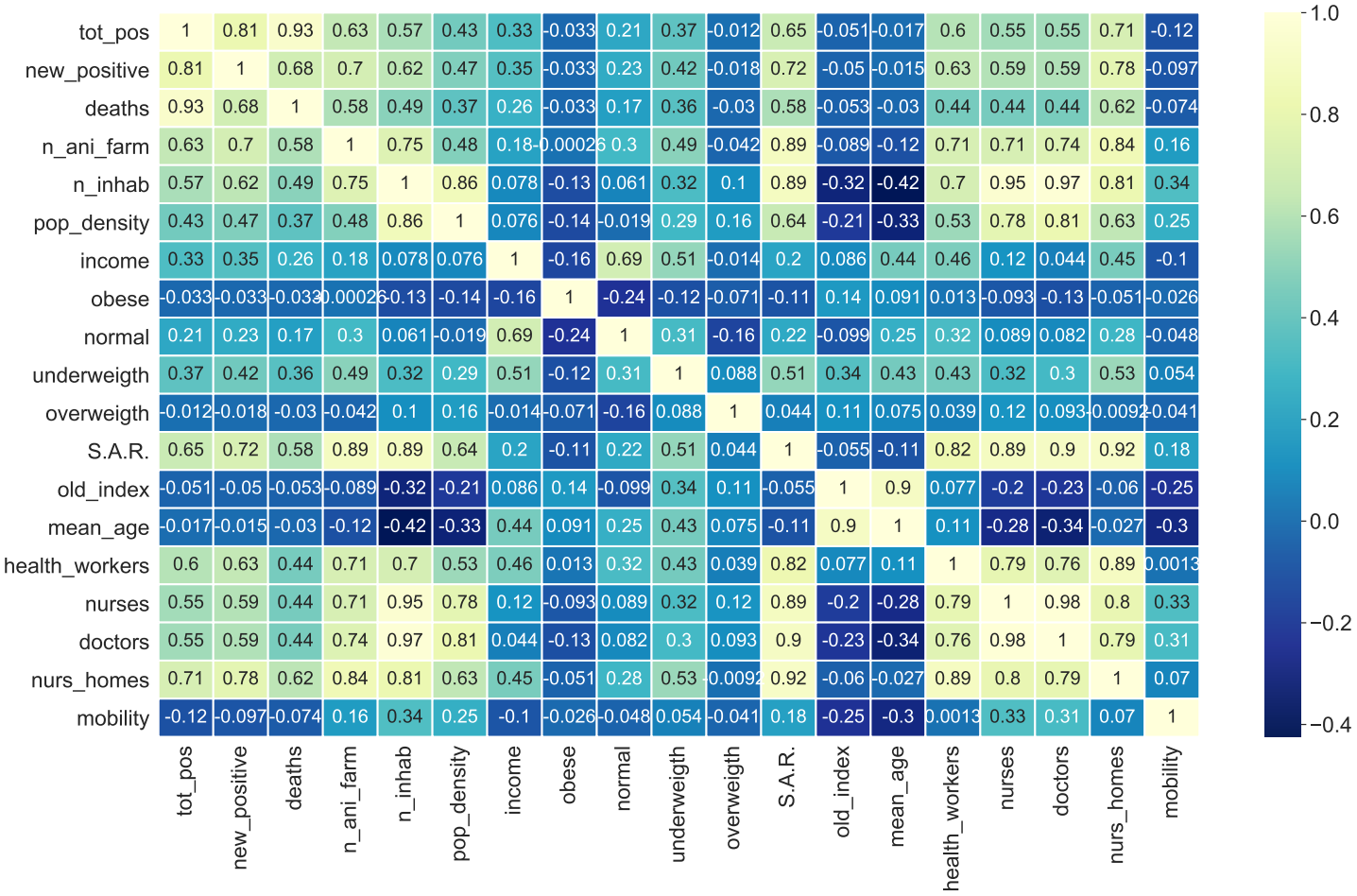


Figure 8: Correlation heatmap among data from table 1, table 2 and table 3.

Figure 8 illustrates correlation among features belonging to table 1 , table 2 and table 3 . As it can be seen, there are features highly correlates such as 'number of inhabitants' with 'number of doctors' and 'nurses', and generally among 'nurses', 'doctors' and 'health workers'.
Among all features composing the dataset, there are time-variant and invariant ones. The time-invariant ones are related to health and to region characteristics except for mobility, which, together with civil protection data, constitute the time-variant part of the dataset. Finally, the dataset is made up of 21 regions * 54 days = 1134 entries

4.1.2 How the results are presented

In the following, there are several sections. Each of them describes the results of the regressions for different target variables. Those are deaths, total cases and new positive. When analysing the features explaining the deaths, the total cases and the variation of new positive are dropped from the dataset. On the other hand, when regressed the total cases, the deaths and the new positive cases are dropped; the same is done when regressing the new positive, where

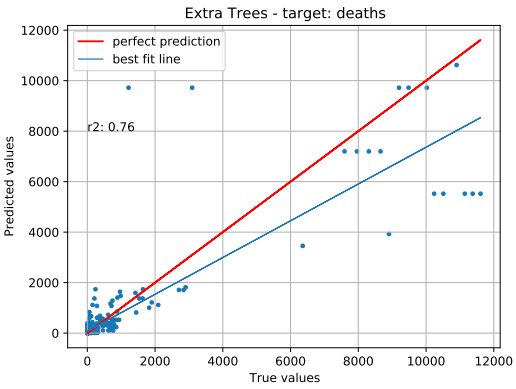
this time, the total cases and the deaths are dropped from the dataset. The reason is clear: the three data are highly correlated so the performance of te algorithms would be overrated, and the most important features would be the number of cases rather than the territorial ones.
Results regarding the features explaining the deaths are shown in section 4.1.3. Instead, the results for positive cases are presented in section 4.1.4.1 for *tot cases* and section 4.1.4.3 for *new positive cases*.

In each section, there are 3 paragraphs: one dedicated to the random forest (RF) and Extra trees (ET), one for linear regressions and one for Gaussian process regression (GPR).

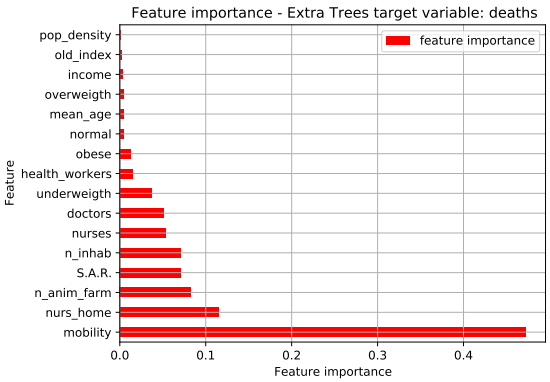
4.1.3 Regressand: number of deaths

In this section the aim is to analyze the features that explain the deaths in the regions.

ENSEMBLE TREES Results are shown in fig. 9.

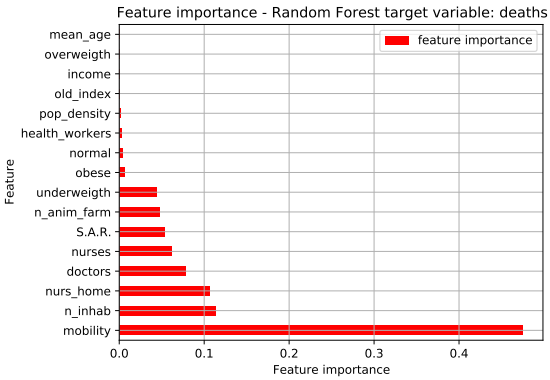


(a) Extra trees regression r2: 0.76



(b) Extra Trees feature importance

(c) Random Forest regression: r2: 0.78

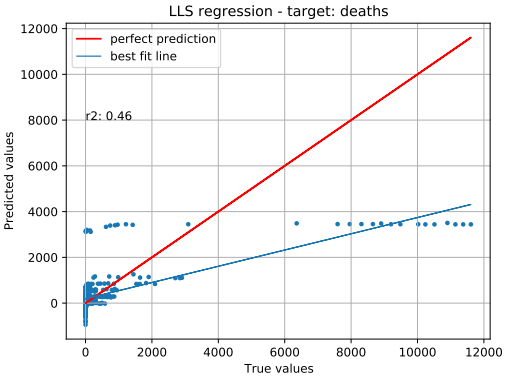


(d) Random Forest feature importance

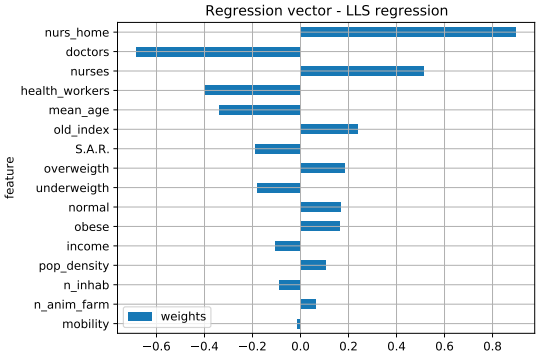
Figure 9: Ensemble methods regression of *deaths*

The most important feature is the mobility followed by number of inhabitants and nursing homes for ET and RF, number of animals in the farms. The regression figures (fig. 9a and fig. 9c) are very similar and in fact, the r2 of the regressions are close too, reaching 0.76 (and ET) and 0.78 (RF).

LINEAR REGRESSION Results are shown in fig. 10.

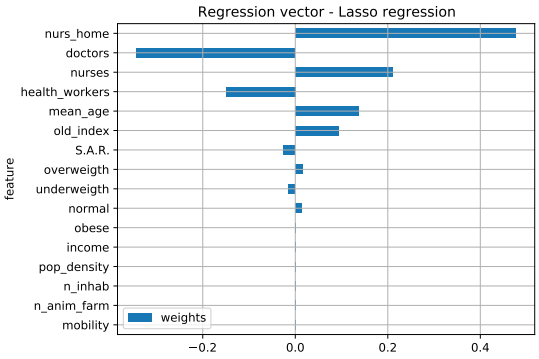


(a) LLS regression r2:0.47



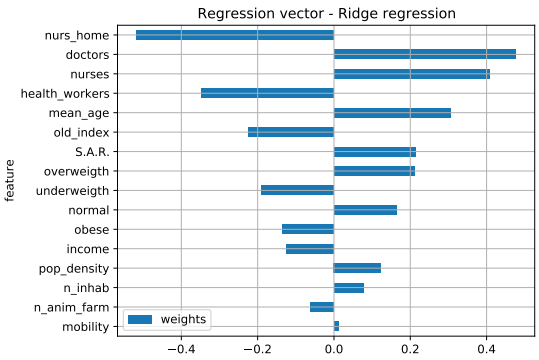
(b) LLS regression weigth vector

(c) Lasso regression r2: 0.44



(d) Lasso regression weigth vector

(e) Ridge regression r2: 0.46



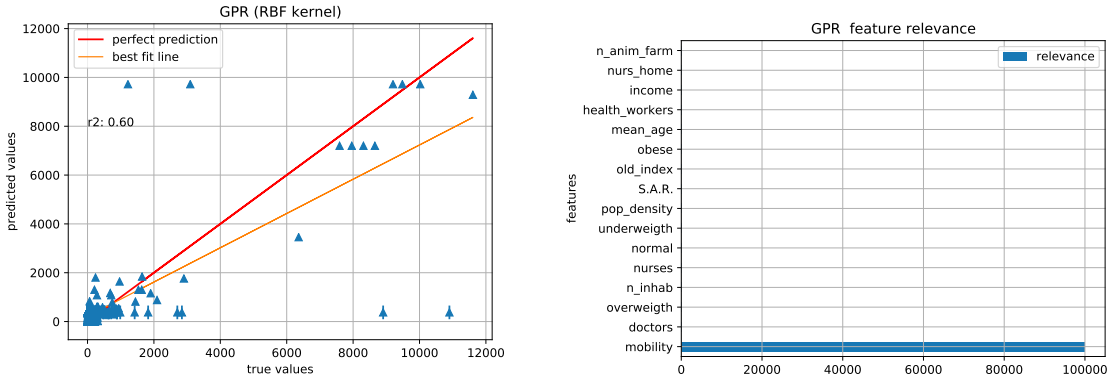
(f) Ridge regression weigth vector

Figure 10: Linear regression of *deaths* in regions

R2 associated with linear regressions with dataset decreases a lot, reaching about 0.47, and the regressions are not so good in the end. It is worth noticing that the most weighted features are the same among the three algorithms: nursing homes, doctors and nurses, even if the signs assigned to the weights

are different. This could be explained with the high correlation among the features.

GPR The kernel used is the RBF, while the constant kernel is 10



(a) GRP r2: 0.60 **(b)** GPR feature relevance

Figure 11: Gaussian process regression of *deaths* in regions

Parameter r2 is equal to 0.60 and hence is not high. From the feature relevance, it appears that the only feature counting for the model is mobility. Note that many points are always predicted close to 0. The predicted value is about 300 and corresponds to the mean value of the deaths in the dataset. The error bar instead represents the standard deviation of the deaths column . So in the end, for those points, the method predicts only the mean value of the dataset.

4.1.3.1 Conclusions on deaths regression

To give a measure of the most important features this procedure was followed: the feature was counted in each algorithm only if it was among the first four most relevant. The number of times that a feature was counted was then divided by the number of algorithms and multiplied for 100, to have a percentage measure. Since linear regression algorithms gave the same results in terms of feature relevance, only one of them is considered in the calculation (more specifically, Lasso is chosen). Hence, the total number of algorithms is four (GPR, Random Forest, Extra trees and Lasso).
Findings of this analysis are shown in fig. 12, which illustrates that more than 70% of times there are 'nursing homes' and 'mobility' among the most important ones, followed by 'doctors' and 'number of inhabitants' (50% of times), and in the end (25% of times) there are 'health workers', 'S.A.R.' and 'nurses'. The best performing algorithms are the ensemble methods, in order Random Forest (r2=0.78) and Extra Trees (r2=0.76). The two methods in fact show the lowest error in terms of RMSE and variance (fig. 13).

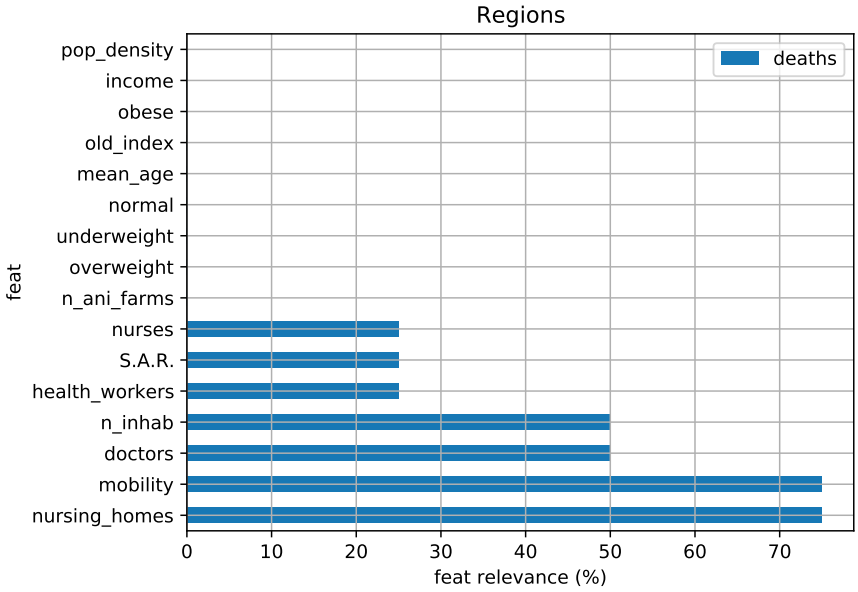


Figure 12: Feature relevance among 4 different algorithms for the regression of *deaths* in regions.

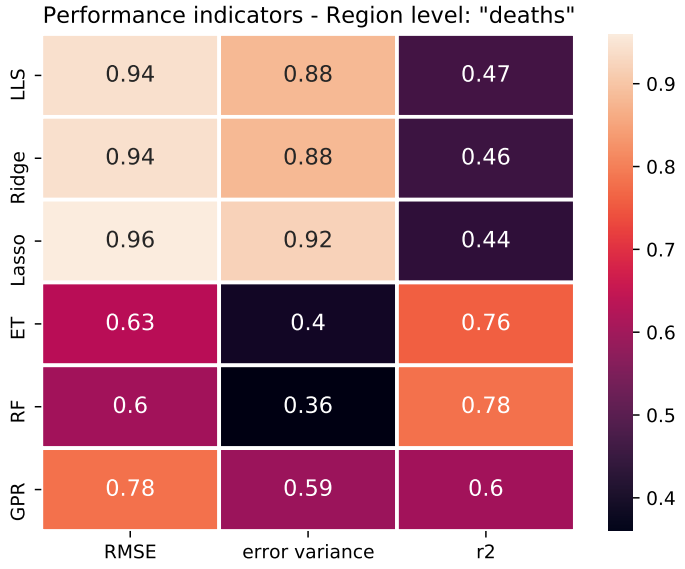


Figure 13: Performance indicators for each algorithm when regressing *deaths* in regions. The values of RMSE and error variance are obtained with standardized data.

Note that the values about error variance and RMSE are obtained with normalized data.

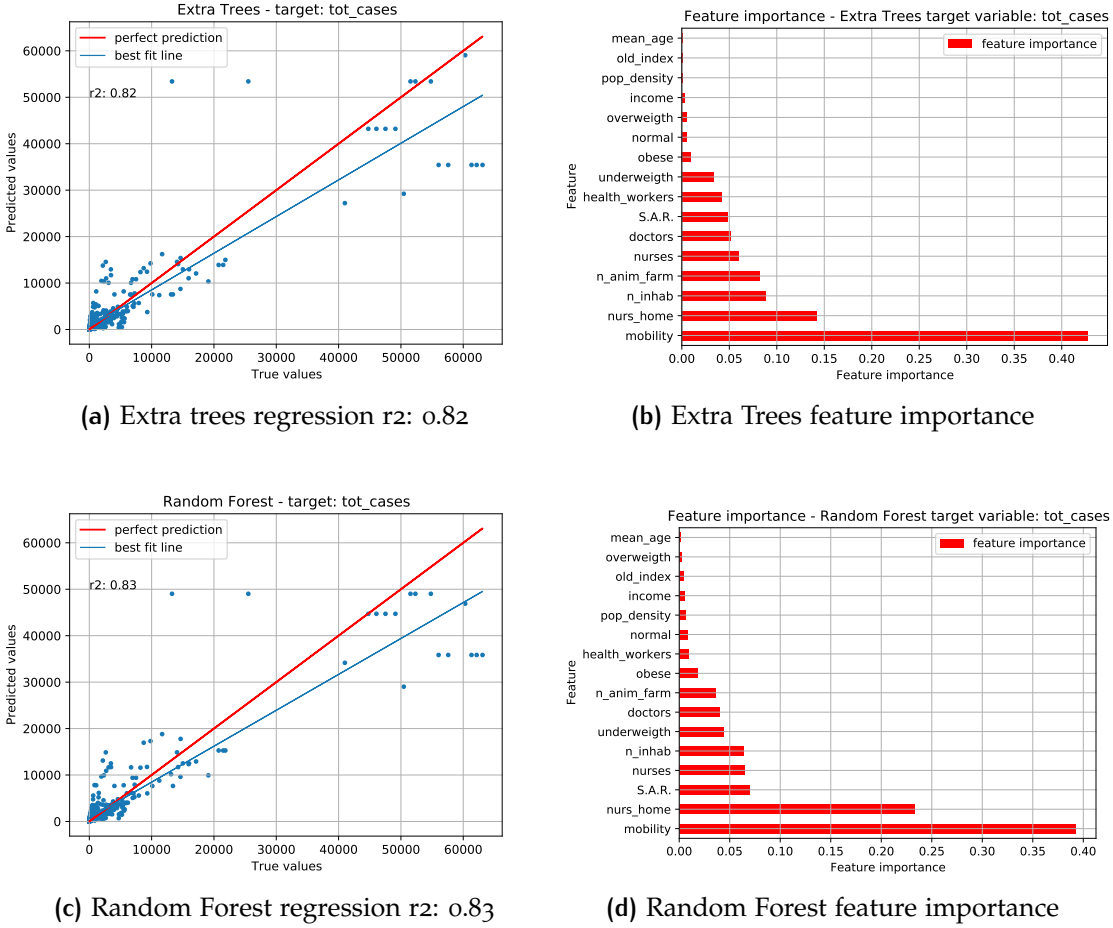
4.1.4 Regressand:number of positive cases

The analysis is repeated with the positive cases of COVID-19 as the target variable. Two sections are present, one concerning the study of the cumulative number of positive cases (section 4.1.4.1), and one section concerning the regression of the daily variation of positive cases (section 4.1.4.3). Note that within these analysis, the used datasets are the same used to regress the deaths in section 4.1.3, they just differs from the target variable.

4.1.4.1 Regressand: total cases

In this section, the regressor variable is the *total cases*. Recall that this variable indicates the cumulative number of infected people, including the ones that healed from the illness, the deaths and the current positive.

ENSEMBLE TREES Results are reported in fig. 14.



(a) Extra trees regression r2: 0.82

(b) Extra Trees feature importance

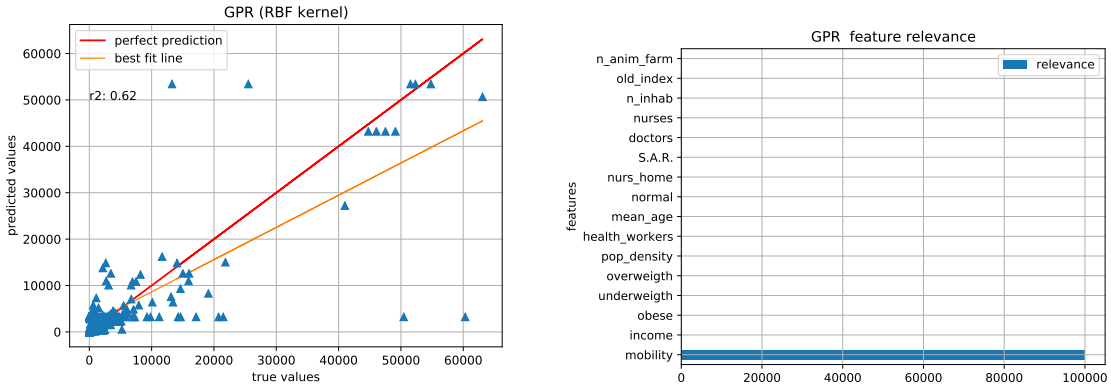
(c) Random Forest regression r2: 0.83

(d) Random Forest feature importance

Figure 14: Ensemble methods regression of *'total cases*.

The regression is not bad, reaching an r2 of 0.82 (ET) and 0.83 (RF). The most important features are the mobility, the nursing homes, the number of animals in farms and inhabitants.

GAUSSIAN PROCESS REGRESSION Result of the regression is shown in the following fig. 15



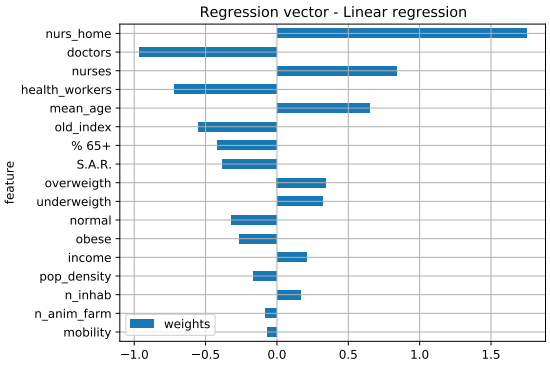
(a) GPR r2: 0.62 **(b)** GPR feature relevance

Figure 15: GPR regression of 'total cases'

The resulting regression shows an r2 of 0.62 and the only relevant feature is the mobility.

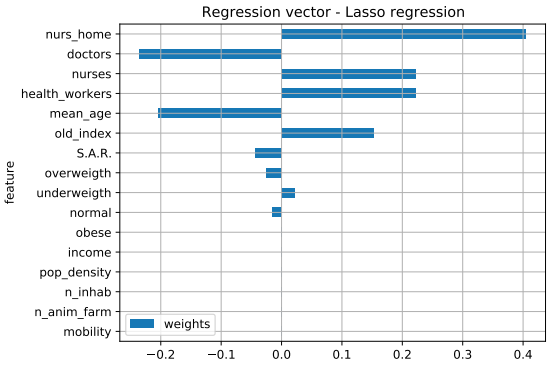
LINEAR REGRESSION Results are shown in fig. 16

(a) LLS regression r2: 0.54

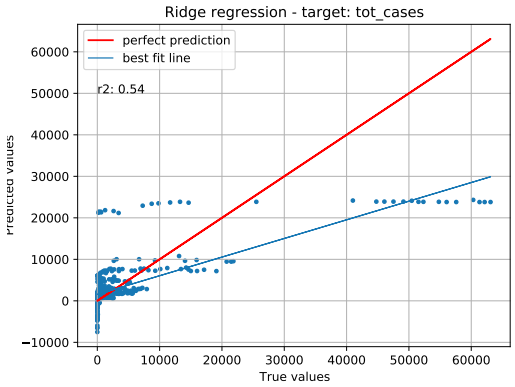


(b) LLS regression weight vector

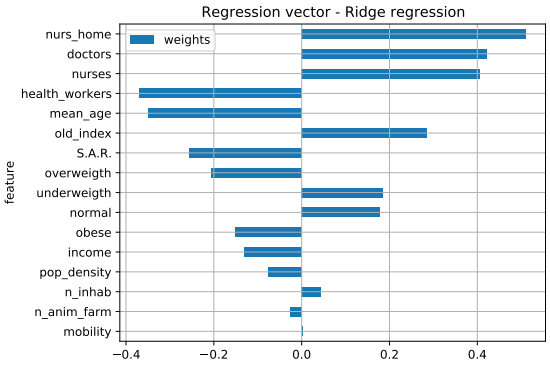
(c) Lasso regression r2: 0.52



(d) Lasso regression weight vector



(e) Ridge regression r2: 0.54



(f) Ridge regression weight vector

Figure 16: Linear regressions of *total cases* in regions.

The linear regression methods do not work excitingly, presenting an r2 metric of 0.54 (LLS and Ridge) and 0.52 (Lasso). Again, the largest weighted features are nursing homes, doctors and nurses.

4.1.4.2 *Conclusions on total cases regression*

Also in this case the best performing algorithms are the ensemble methods. It is worth mentioning that for both ensemble methods and linear regression, the most relevant features (and the highest weighted in linear regression) are the same obtained with the regression of deaths (section 4.1.3): mobility, nursing homes, number of animals per farm for ensemble methods and nursing homes, doctors, nurses and health workers for linear regression.

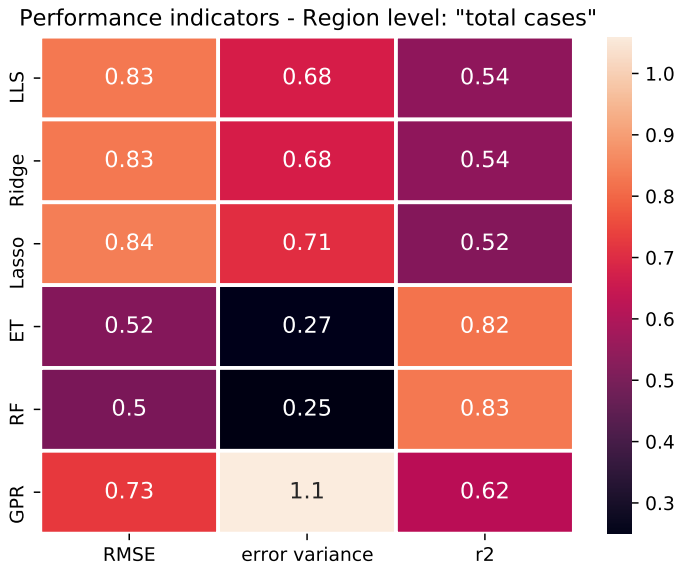


Figure 17: Performance indicators for each algorithm for the regression of *total cases* in regions. The values of RMSE and error variance are obtained with standardized data.

Figure 17 that reports the performance indicators of all the algorithms, illustrates that ensemble methods do have the highest r2 value and the lowest RMSE and error variance; in second place in terms of r2 and RMSE is GPR which however reports the highest error variance.

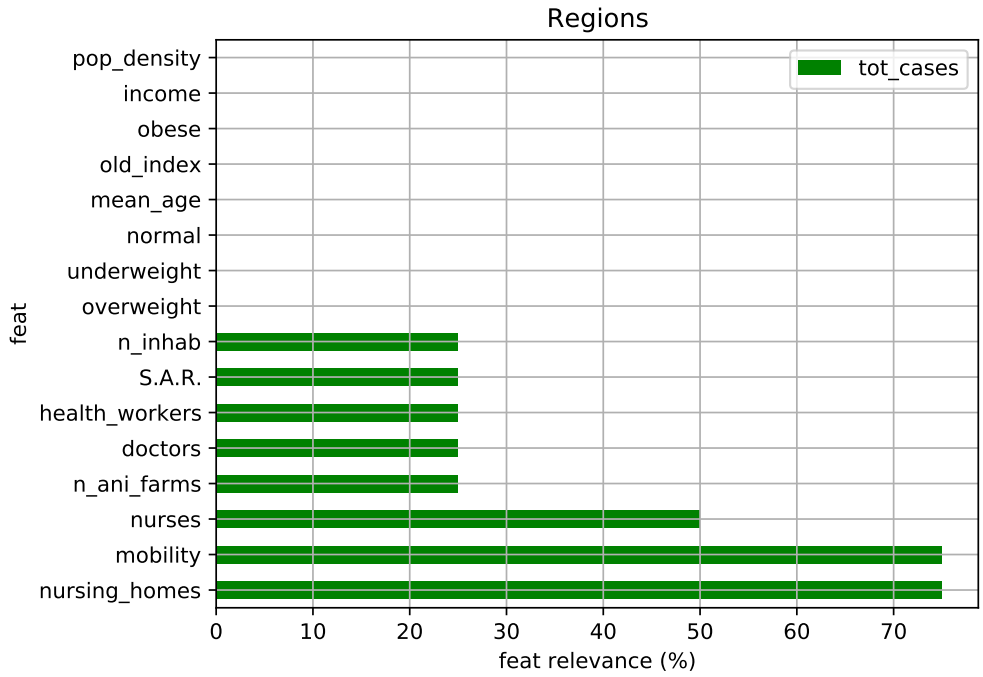


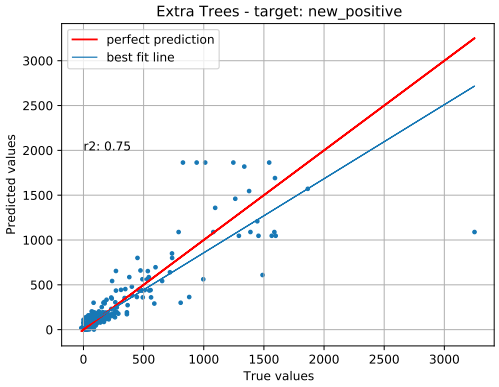
Figure 18: Feature relevance among 4 different algorithms for the regression of *total cases* in regions.

Again, linear regression algorithms output the same ordered features relevance, so the results of only one of the three methods (LLS, Lasso and ridge) is considered in the calculation. Figure 18 shows that 'nursing homes' and 'mobility' have been the most important more than the 70% of the cases (three times out of four), at the second place there is 'nurses' (50% of times), followed by 'doctors', 'health workers', number of animals farm, 'S.A.R.' and 'number of inhabitants'. Note that the results are similar to fig. 12.

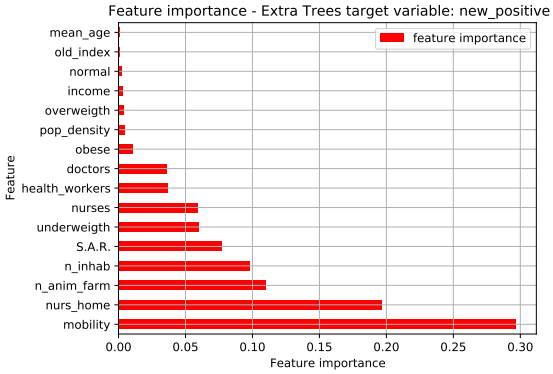
4.1.4.3 *Regressand: new positive cases*

This section covers the regressions results when regressing *new positive cases* target. Also in this case the dataset used is table 2. The *new positive cases* target is actually the daily variation of the *total cases* previously regressed. The aim is to understand if the relevant features change in the two cases.

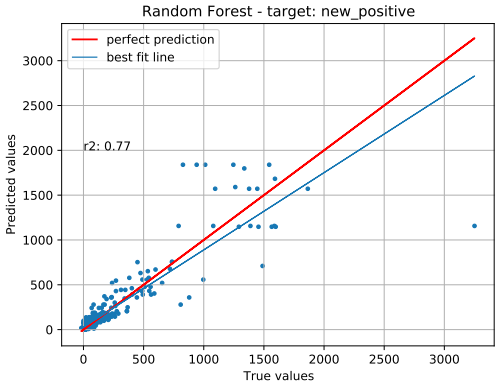
ENSEMBLE TREES In fig. 19 the results about ensemble methods are presented.



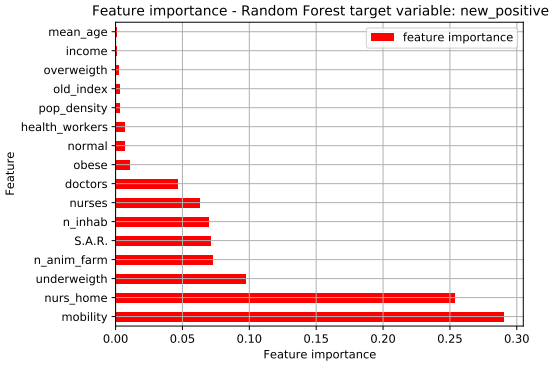
(a) Extra trees regression r2: 0.75



(b) Extra Trees feature importance



(c) Random Forest regression r2: 0.77

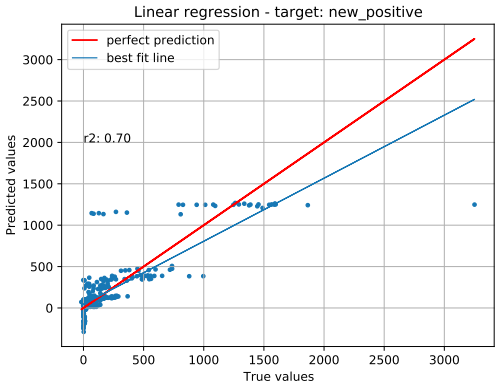


(d) Random Forest feature importance

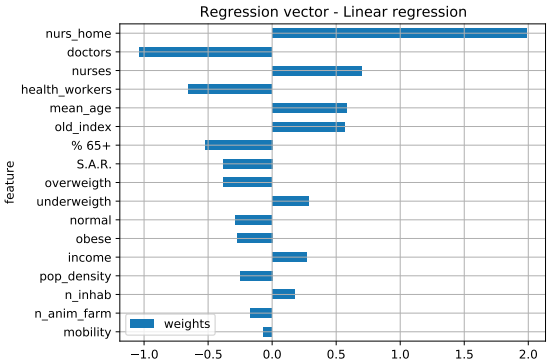
Figure 19: Ensemble methods regressions of *new positive cases* in regions

Regressing the 'new positive cases' gives a worse performance concerning the 'total cases' in terms of r2 metric reaching 0.75 (ET) and 0.77 (RF). The most important features are again the mobility, the number of inhabitants, the S.A.R., the number of animals in farms and nursing homes.

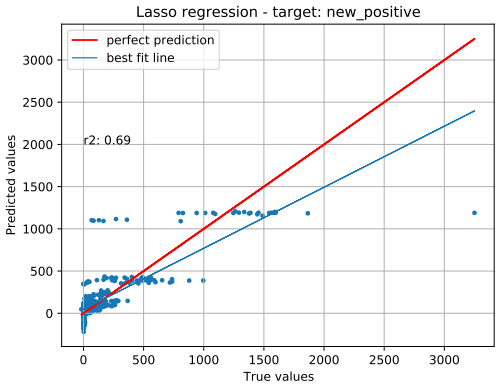
LINEAR REGRESSION Results of linear regressions are reported in fig. 20



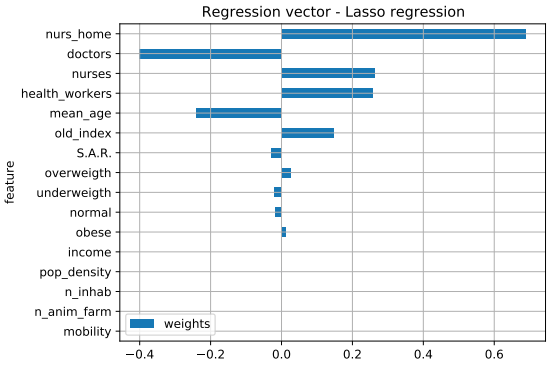
(a) LLS regression r2: 0.70



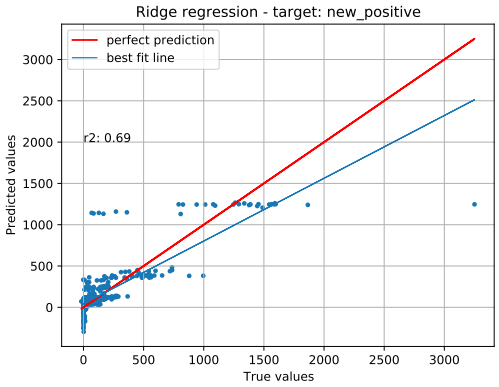
(b) LLS regression weight vector



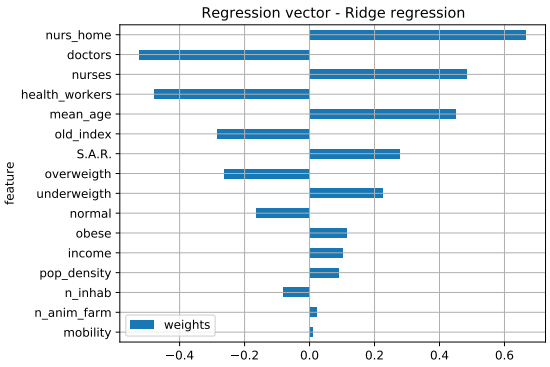
(c) Lasso regression r2: 0.69



(d) Lasso regression weight vector



(e) Ridge regression r2: 0.69

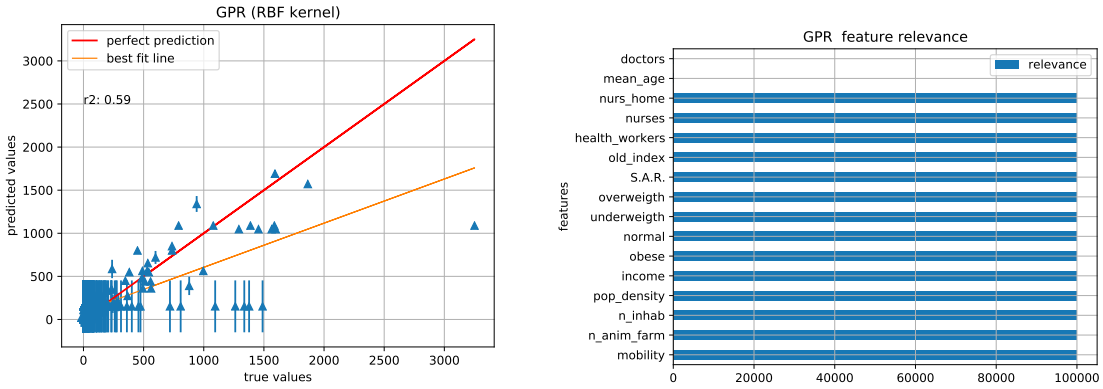


(f) Ridge regression weight vector

Figure 20: Linear regressions of *new positive cases* in regions

For linear regression, the results are better than the linear regression applied to the 'total cases' (section 4.1.4.1), increasing the accuracy r2 of about 0.2 points. The most weighted features are the usual ones: nursing homes, doctors, nurses.

GAUSSIAN PROCESS REGRESSION Gaussian process regression is shown in fig. 21



(a) GPR r2: 0.59 **(b)** GPR feature relevance

Figure 21: Gaussian process regression of *new positive cases* in regions

The regression shows that r2 is 0.59. Again, the algorithm could not regress many points, and these are the ones horizontally aligned close to the 0.

4.1.4.4 *Conclusions on new positive cases regression*

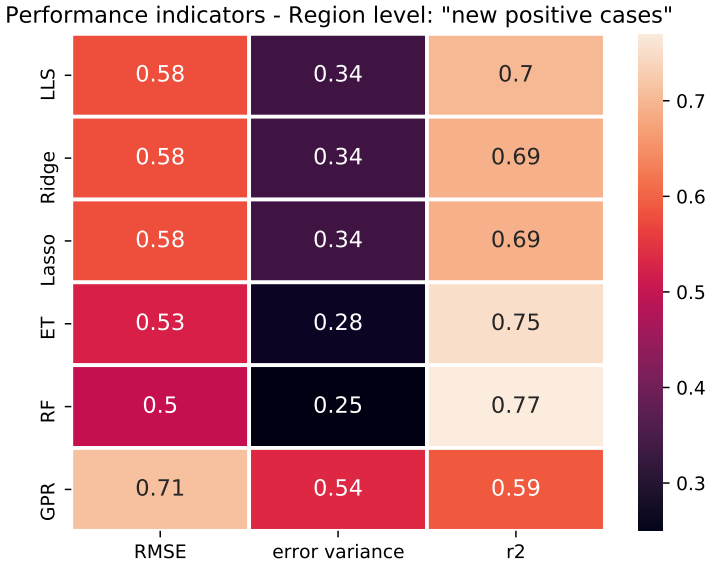


Figure 22: Performance indicators for each algorithm for the regression of *new positive cases*. The values of RMSE and error variance are obtained with standardized data.

Figure 22 illustrate that the best algorithms are the ensemble methods in terms of all the three performance indicators (RMSE, error variance and r2). Very

similar outcomes are given by the linear regressions algorithms. GPR instead is the one with the highest RMSE and error variance and lowest r2, resulting to be the worst method.

The results are similar to fig. 12 (illustrating the features relevance for *deaths* regression) and fig. 18 (illustrating the features relevance for *total cases* regression)

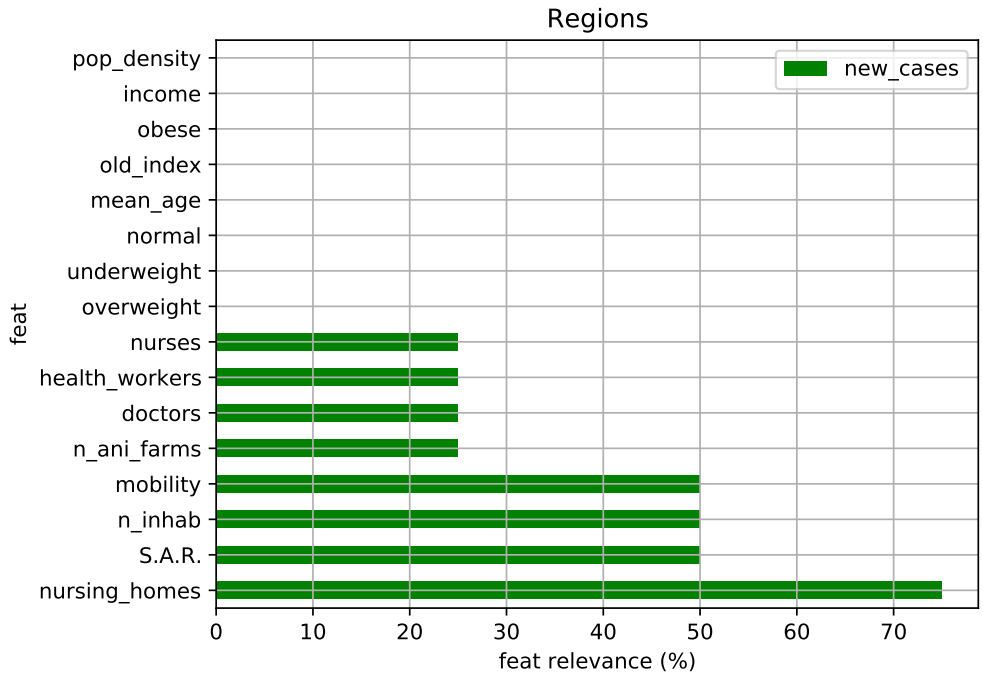


Figure 23: Feature relevance among 4 different algorithms for the regression of *new positive cases* in regions.

Concerning features importance, again, only one linear regression method results are considered, since the three algorithms gave the same outcome. Figure 23 illustrates that three times out of four (more than 70%) nursing homes has been among the most important features, followed by S.A.R., mobility inhabitants the 50% of times. Features like 'doctors', 'health workers' and 'nurses' have been among the most important only one time out of four, corresponding to the linear regression that, as already seen, always elect them as the most weighted features.

4.2 RESULTS – PROVINCIAL LEVEL

This section reports the results of the relevance of the virus diffusion in the provinces, in terms of *total cases* and *new positive cases*, corresponding in the daily variation of Covid-19 cases.

4.2.1 Data acquisition

The dataset includes different features with respect to the region level dataset. This is manly due to the difficulty in retrieving and finding some of them. The features can be divided in demographic information (table 4), weather information (table 5)and province characteristic (table 6):

Feature name	Name meaning
mean age	mean age
n inhab	number of inhabitants
pop density	population density
old index	oldness index

Table 4: Demographic features for provinces from [4]

All the demographic information are from ADMINSTAT Italia ([4]).

Feature name	Name meaning
min T	min temperature
max T	max temperature
rain	rain in millilitres
cloud	percentage of NOT covered sky
wind	wind in $\frac{km}{h}$

Table 5: Weather features for provinces from the site [36]

The weather information was retrieved from the site [36]. The research for each city was done manually because the site did not provide a suitable way to search for all the Italian cities at once. It must be said that the meteorological information provided by the site are not observed data: they come from the estimation given 10 years of historical observations.

Feature name	Name meaning	Data source
n ani farm	total number of animals in all the farms	ISTAT [29]
income	average per-capita income	[12]
pub transp	per capita usage of public transportation	ISTAT [31]
airport flow	number of people in the airports	[7]
S.A.R.	Serious Accident Risk industries	[35]
com industry	commerce indusry (wholesale, retail trade)	ISTAT [43]
accom rest	number of accommodations and restaurants	ISTAT [43]
nurs home	number of sanity and social assistance services	ISTAT [43]

Table 6: Province characteristics

Concerning the airport flow, the site of ASSAEROPORTI was consulted. The data are related to the flights of January 2019. However, since the lockdown limited also the airway traffic, the data of [7] are assumed to be the baseline of common airway traffic and the dataset, in the different dates, is handled in such a way to reflect the Italian situation.
For this purpose, the Ente Nazionale per l'Aviazione Civile (ENAC) website was consulted too [22]. In this website, the organization notifies the changes and limitations of airflow required by the d.p.c.m. (Decreto del Presidente del Consiglio dei Ministri, i.e. an ordinance by the prime minister). On 11 March ENAC informs of the closure of all Italian airports from 14th March onwards, hence from that date the entire civil airport flow was stopped. Among the flights still feasible there are those for emergencies such as governmental flights or organ transfer. For this reason, from 14th March onwards the airport flow is assumed to be close to 0 until April 11th, which is the latest date of the dataset.

Feature name	Name meaning	Data source
positive	cumulative number of infected people	[26]

Table 7: Civil protection data for provinces

For the provinces, the civil protection, through the same Github repository exploited for region data, only provides data about the number of total positive people (table 7).
In this context, both the 'total positive' and the 'new positive cases' are regressed. The latter is not directly provided from the civil protection source but can be obtained by subtracting the total positive of the previous day to the total positive of the current day.

The dataset is hence formed by a set of features that are time-invariant and some that are time-variant. The latter is related to the number of infected people, the usage of public transportation, the airport flow and meteorological information. All the others are time-invariant.

Differently than the case of the regional level, the observation time in which the time-variant data are observed is limited to four single dates, and they are: 16/02, 11/03, 21/03, 31/03 and 11/04. The reason is that to observe an event, time-variant features are needed, and for the provincial level these are scarce. In fact, data about airport flow are given not daily based, but aggregated by months. Instead, the meteorological information, which is the only feature that could be taken on different days, could not be retrieved once for all the provinces, but the process of collecting data was done manually. To have a longer dataset, more days of observation were needed; however, because of the unfeasibility to manually grow the dataset, only four dates were chosen.
The final dataset is composed of all the Italian provinces (107) analyzed in 6 different days; however, for 4 provinces some data is missing, hence the total records of the dataset is 103*6= 618.
The dataset is analyzed with regression algorithms. The objective is twofold: estimate the new positive cases and the estimation of cumulative positive cases. Since, as already specified, the latter is calculated by subtracting the total positive of an old date to the total positive of the newest date, the time passing between one day and the next almost corresponds to the incubation period of the virus, that results to be about 10-14 days.

Finally, fig. 24 reports correlation coefficients among features composing the province dataset. As can be seen, there is a particularly high correlation among 'nursing homes', 'accommodation restaurants' and 'commerce industries'.

4.2.2 How results are presented

The results are shown similarly to region results (section 4.1.2): there are two sections, one reporting results for the new positive cases (section 4.2.3) and one for the total positive (section 4.2.4); each section reports results with the 3 main groups of algorithms (Linear regression, ensemble trees, GPR). At the end of each section there is a small part of conclusion in which the results of the three groups of algorithms are compared.

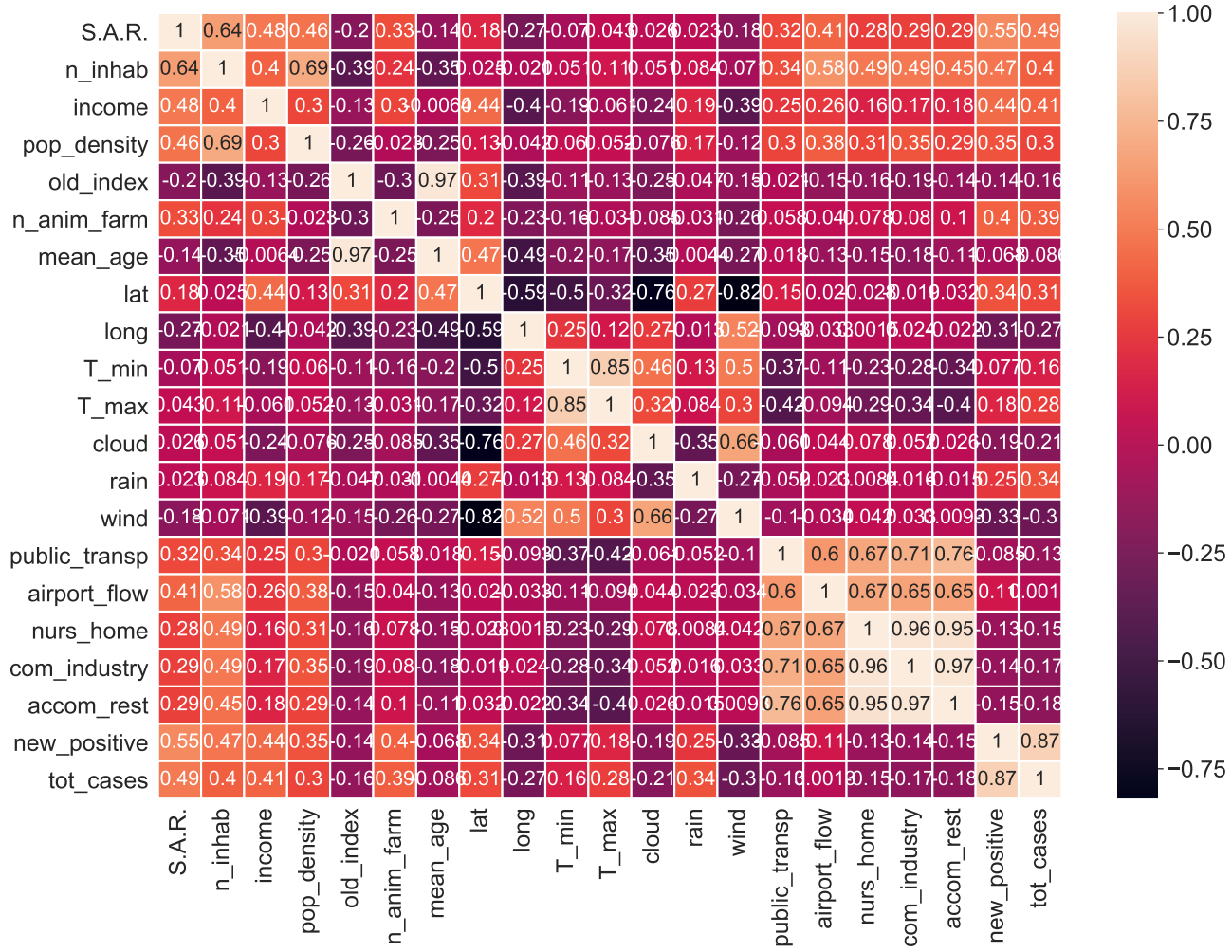
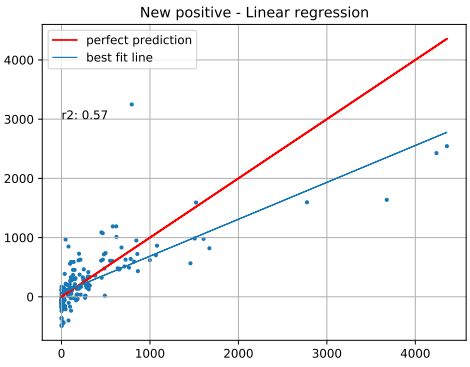


Figure 24: Correlation heatmap for province dataset including all the features (table 4, table 5, table 6 and table 7)

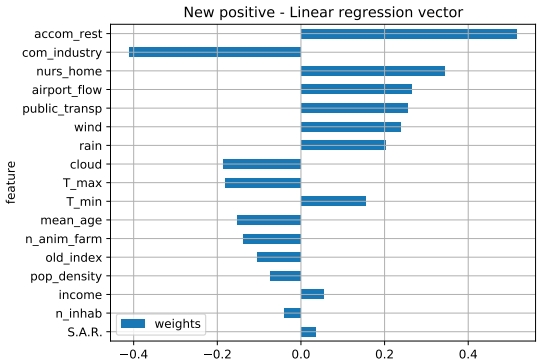
4.2.3 Regressand: new positive cases

This section will cover the regressions results obtained when regressing the new positive cases.

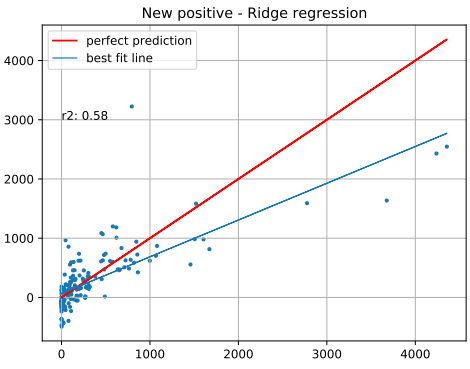
LINEAR REGRESSION Linear regression results are reported in section 4.2.3



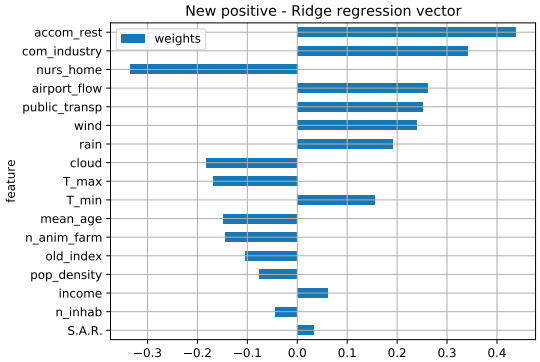
(a) LLS regression r2: 0.57



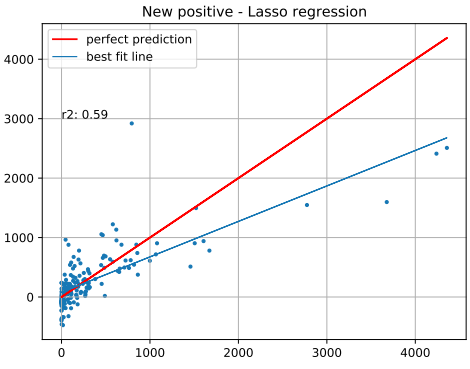
(b) LLS regression weight vector



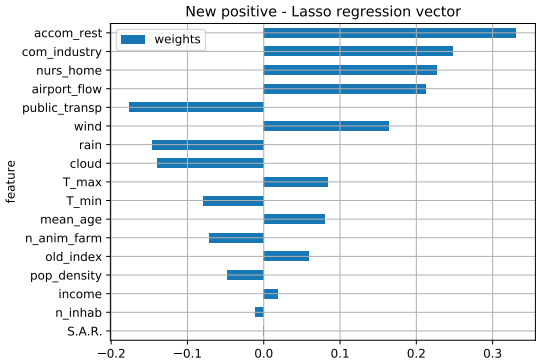
(c) Ridge regression r2: 0.62



(d) Ridge regression weight vector



(e) Lasso regression r2:0.57

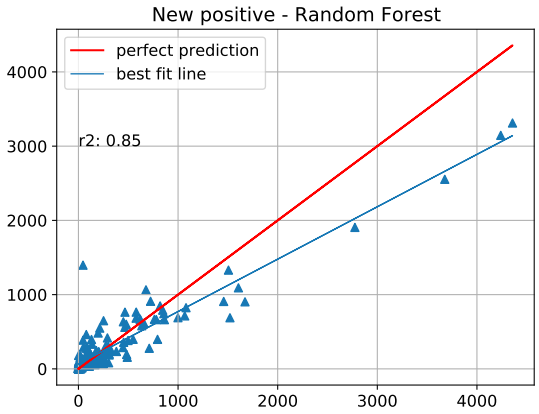


(f) Lasso regression weight vector

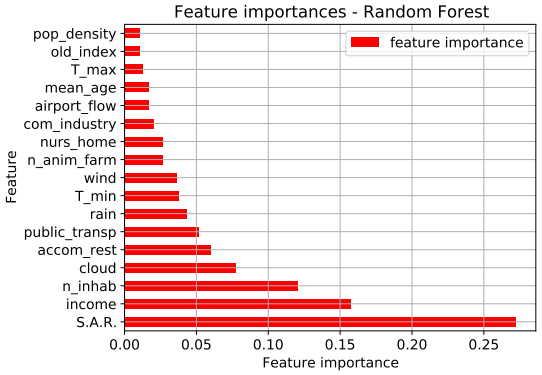
Figure 25: Linear regression of *new positive cases* in provinces

The weight vectors of the three linear methods are similar, in fact, the order (decreasing module of weights) in which the features are, is the same. The largest weighted features are accommodation and restaurants, commerce industry, nursing homes and airport flow. The r2 of the regressions are very close: 0.57 for LLS, 0.58 for Ridge and 0.59 for Lasso.

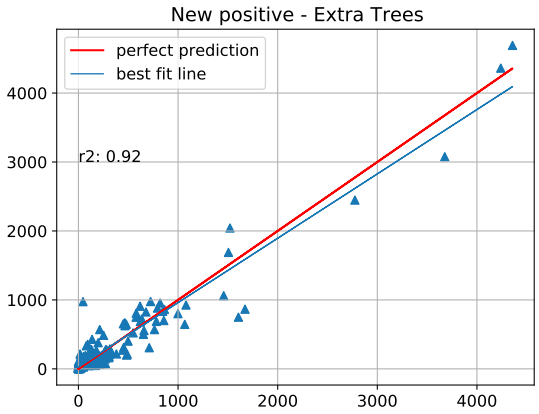
ENSEMBLE TREES Results are in section 4.2.3.



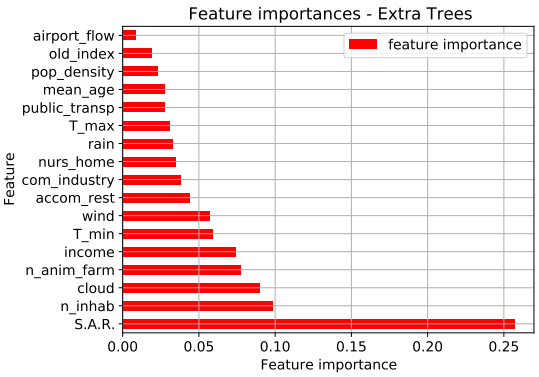
(a) Random Forest regression r2: 0.86



(b) Random Forest feature importance



(c) Extra tree regression r2: 0.92

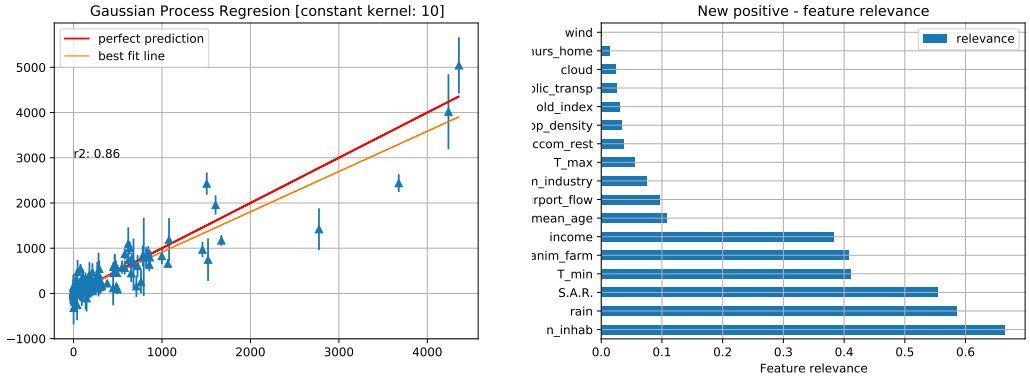


(d) Extra Trees feature importance

Figure 26: Regression of *new positive cases* with Trees in provinces.

The regressions obtained with the ensemble methods are very good and the r2 scores are very high. The best one is however the Extra Trees with r2=0.92. The most important features are S.A.R. for ET and RF followed by inhabitants and income.

GAUSSIAN PROCESS REGRESSION The kernel used with this dataset is the Matérn kernel with $\nu = 2.5$. Other types of kernel did not give acceptable regression results. The GPR regression is reported in fig. 27:



(a) GP regression r2: 0.86 **(b)** Feature relevance

Figure 27: GPR with Matérn kernel of *new positive cases*

The regression is very good, reaching an r2 score of 0.86. The most relevant features are: number of inhabitants, rain, S.A.R., min temperature, number of animals, income.

4.2.3.1 *Conclusions on new positive cases regression*

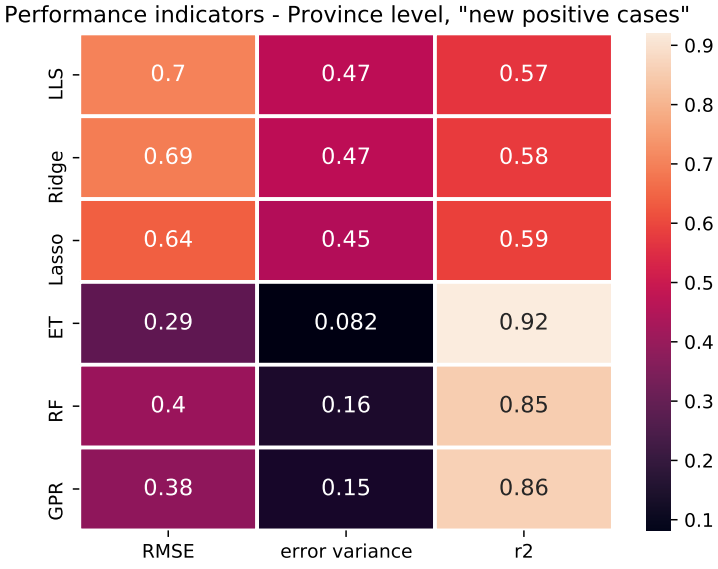


Figure 28: Performance indicators for each algorithm for the regression of *new positive cases* in provinces. The values of RMSE and error variance are obtained with standardized data.

Considering fig. 28, it shows that the lowest RMSE and error variance is given by Extra Trees which also shows the highest value of r2; afterwards there is GPR followed by Random Forest.

Regarding the most relevant features, the algorithms considered in this analysis are again four (GPR, Random Forest, Extra trees and one for linear regression) since all three linear regressions gave the same results. Figure 29 illustrates that the number of inhabitants and the number of S.A.R. have been the most relevant ones 70% of the times (with three methods out of 4); soon after comes the feature 'cloud'.

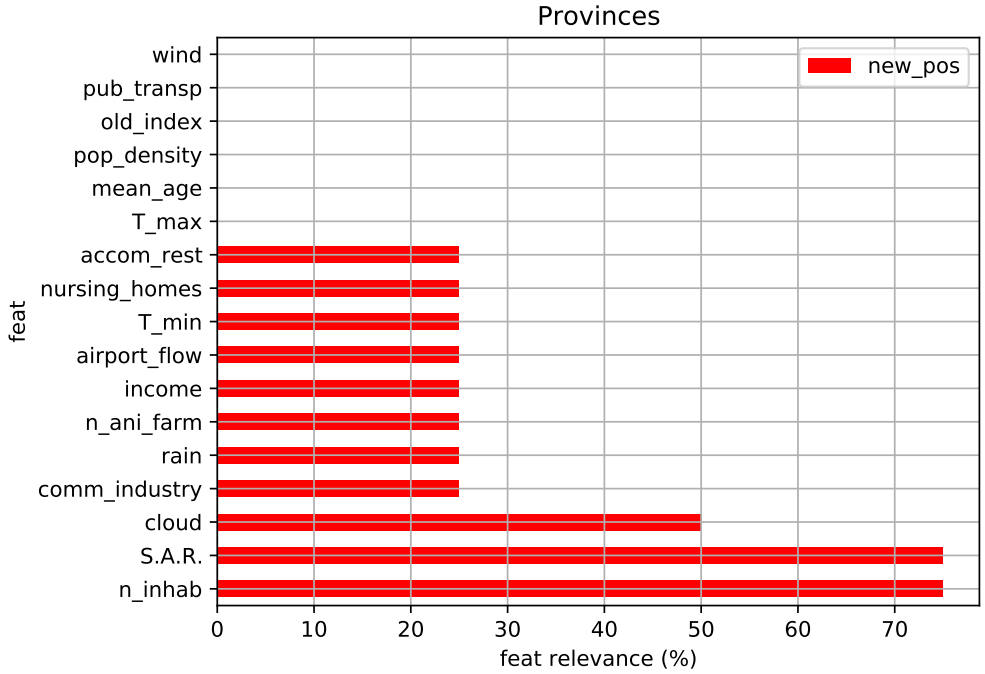
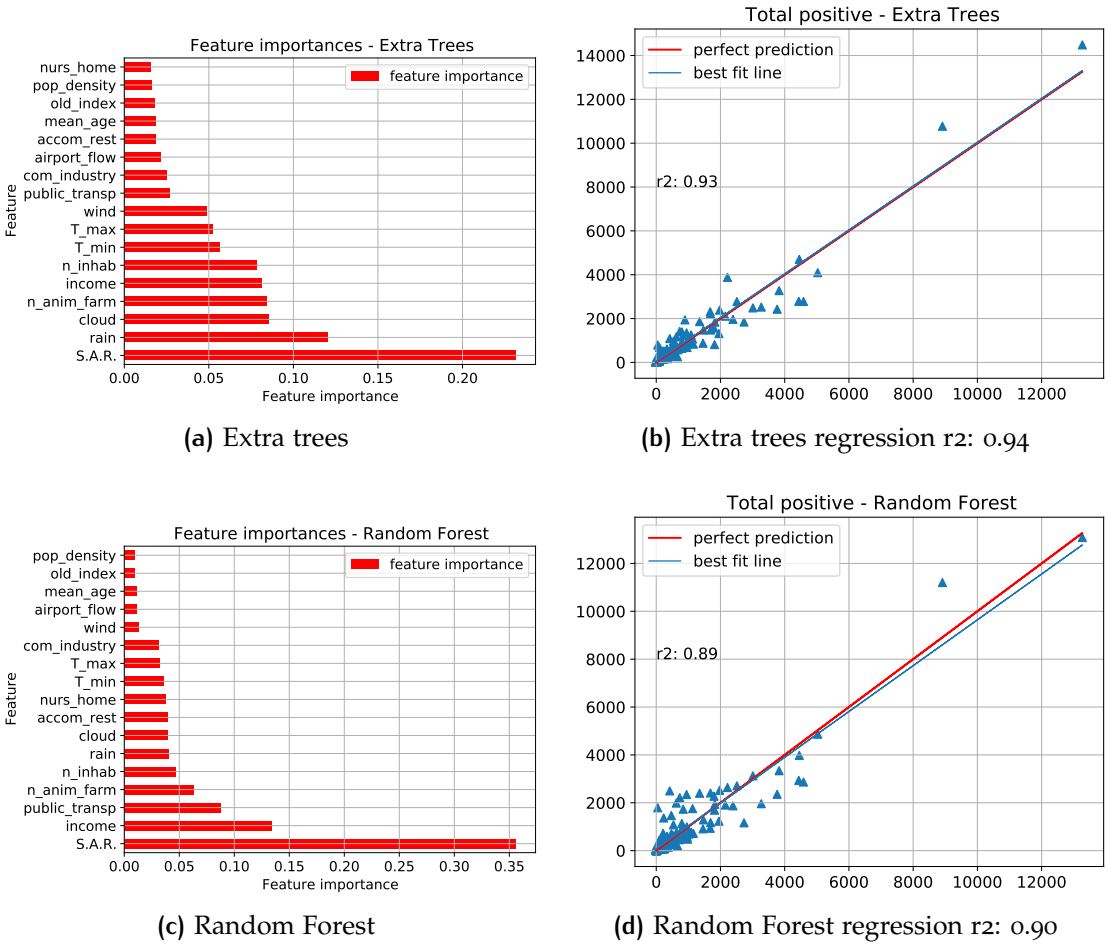


Figure 29: Feature relevance among 4 different algorithms for the regression of *new positive cases* in provinces.

4.2.4 Regressand: total cases

This section reports the results about the regression of *total cases* of Covid-19.

TREES Results with ensemble trees are reported in fig. 30



(a) Extra trees

(b) Extra trees regression r2: 0.94

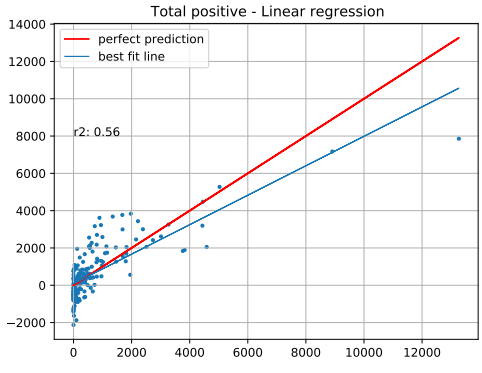
(c) Random Forest

(d) Random Forest regression r2: 0.90

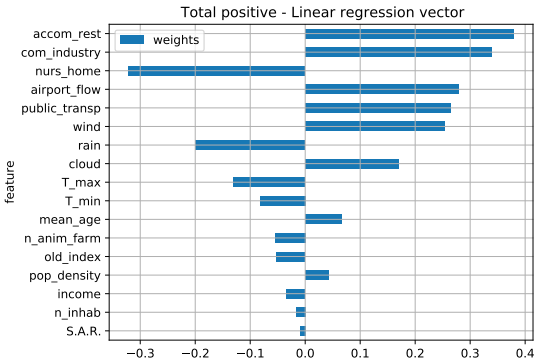
Figure 30: Ensemble methods regressions of *total cases* in provinces.

The regressions are even better than before when regressing the 'new positive cases'; the best one is still the Extra trees. The most important features are S.A.R. for all of them, income for RF and rain for ET.

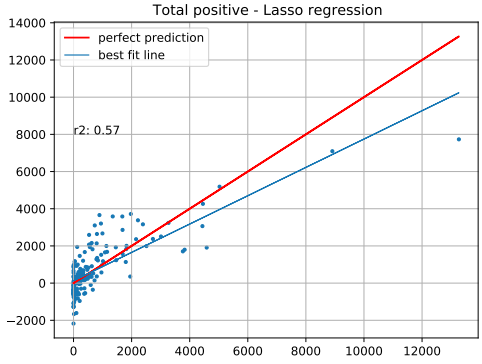
LINEAR REGRESSION Results of linear regression are reported in fig. 31. Such as the case of regions, also here the linear regressions show the same largest weighted features, identical among them and identical to the ones that came out when regressing the 'new positive cases': accommodation restaurant, commerce industry, nursing homes and airport flow. Also in this case, they differ from the signs. The r2 scores are not high, being 0.56 for LLS and Ridge and 0.57 for Lasso.



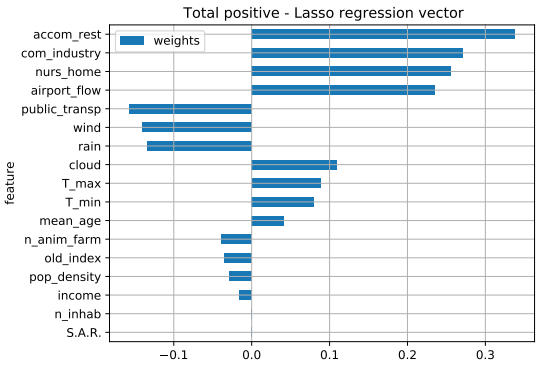
(a) LLS regression r2: 0.52



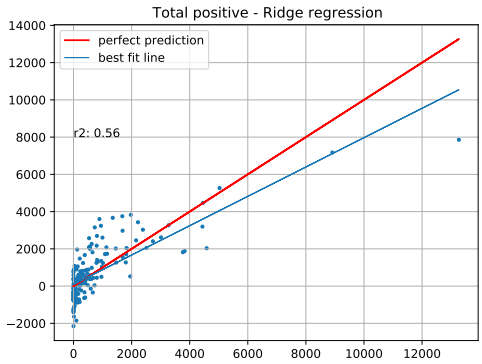
(b) LLS regression weight vector



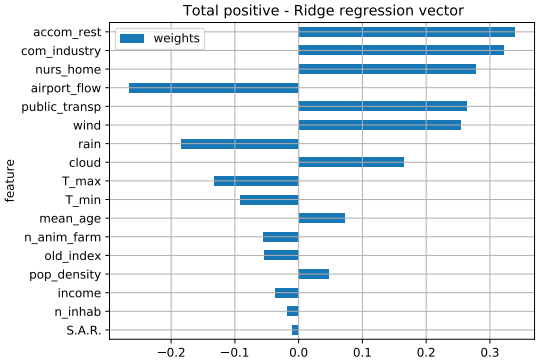
(c) Lasso regression, r2: 0.54



(d) Lasso regression weigth vector



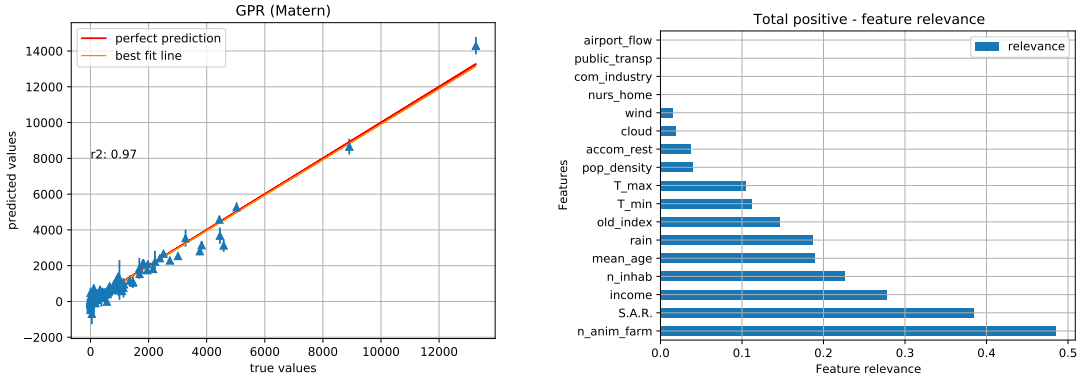
(e) Ridge regression, r2: 0.58



(f) Ridge regression weigth vector

Figure 31: Linear regression of *total cases* in provinces.

GAUSSIAN PROCESS REGRESSION Result obtained with GPR is reported in fig. 32. The kernel used is again Matérn with $\nu = 2.5$.



(a) GP regression r2: 0.97 **(b)** GPR feature relevance

Figure 32: Gaussian process regression of *total cases* in provinces.

The regression with GP is very good reaching a high r2 score (0.97). The most relevant features are similar to the ones obtained with *new positive cases*: nursing homes, income, rain, animals, S.A.R., population density

4.2.4.1 *Conclusions on total cases regression*

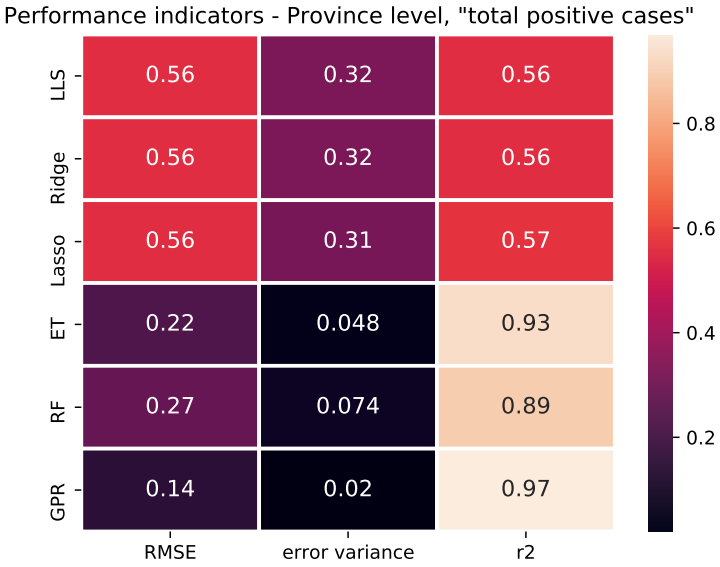


Figure 33: Performance indicators for each algorithm for the regression of *total cases* in provinces. The values of RMSE and error variance are obtained with standardized data.

Figure 33 shows that the best algorithms are GPR and ensemble methods, that report the lowest RMSE and error variance and the highest r2. The best efficient algorithm among all is the GPR.

Also in this case, linear regression algorithms gave the same results in terms of feature relevance, so again, only one of them is considered in the calculation. Figure 34 illustrates the most frequently important features:

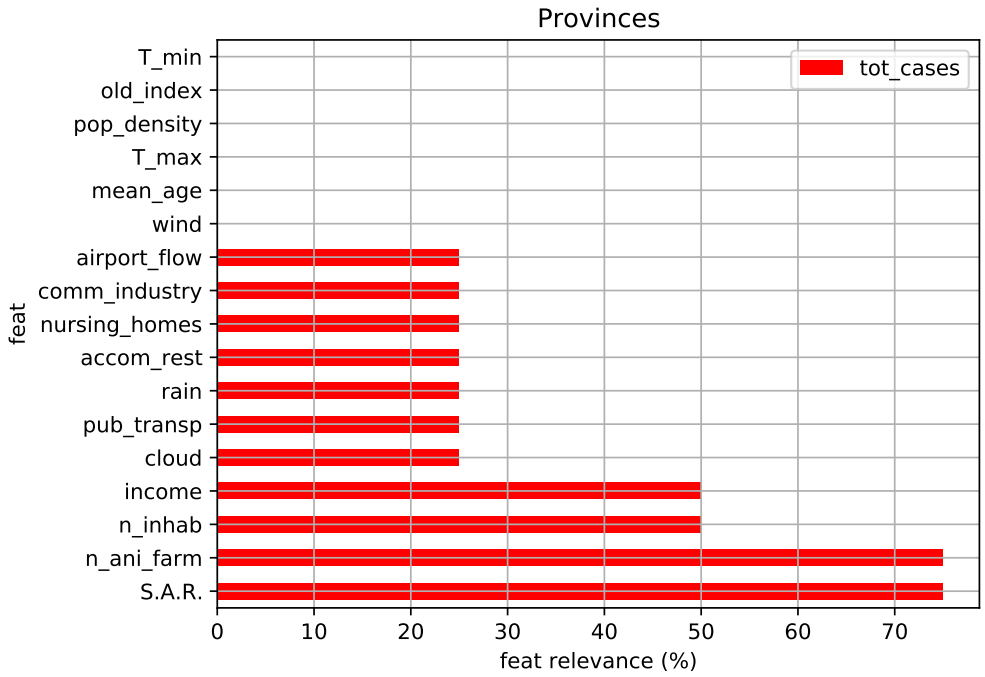


Figure 34: Feature relevance among 4 different algorithms when regressing the *total cases* in provinces.

The number of S.A.R., together with the number of animals (three times out of four), immediately followed by the number of inhabitants and income, seem to be the most important features.

It is interesting to notice that for both *new positive cases* and *tot cases* some features such as 'wind', 'mean age', 'oldness index', 'population density' have never been among the most four important features, despite the relevance that people try to assign to them for the virus spread (as mentioned in chapter 2).

5 CONCLUSIONS

With the discovery of the novel COVID-19 disease which counted the first cases of severe pneumonia during December in China, and with the fast spread all over the globe, many scientists have worked hard trying to give a contribution to the research in any field. Machine learning researchers, for instance, worked in application fields from image diagnosis, to therapeutical research for drug, to forecasting methods and so on.
Interesting studies include the investigation of the virus spread. In particular, the fact that diffusion of the virus has been different worldwide led to think that probably there are some characteristics bounded to the territory that could influence the virus diffusion among the population.
In this regard, many scientists are considering climatic (presence of sun or wind) and environmental conditions (such as pollution) as principal factors for the diffusion of the virus.
Italian situation regarding COVID-19 disease represents an interesting case, because the virus spread differently also within the Italian soil, distinguishing itself in the three main areas of North, Center and South. Such disparity includes both the number of contagious and the death rate.
This thesis has aimed to look deeper, at the regional level and the provincial level, on the possible territorial characteristics, among a list of chosen ones, that could have made the difference for the virus diffusion over the Italian territory in terms of both number of infection and number of deaths.
For this scope, supervised machine learning algorithms were exploited. In particular, regression techniques were used.
To cope with this study, two datasets were built: one for provinces and one for regions.
The features composing the two datasets are not the same. For instance, the province dataset contains meteorological features (rain, percentage of not covered sky, temperature) but the regions dataset does not because it was not possible to retrieve this information; moreover, the regions dataset contains data regarding the mobility that is not available for province level.
Features that appear in both datasets are data related to nursing homes, S.A.R. (number of Serious Accident Risk industries), number of animals farmed and demographic information such as mean age, oldness index, number of inhabitants, population density and income. On the other hand, some features that were meant to be included such as the pollution, were not available through open data.
News, researchers and personal hypothesis have been the sources for features identification.

For what concerns the timescale of the data, there is a difference between provinces and regions too. At the regional level in fact, the dataset records refer to 54 days (from February 26th to April 17th) due to the availability of Google mobility patterns and the Github repository for COVID-19 cases data. Provinces dataset instead only includes the meteorological data as time-variant features and, being these manually collected for each of the 107 provinces, for timing reasons, in the end, the collected data are related to 6 days spaced by roughly 10 days.
The objective of the study was threefold in the case of regions and twofold in the case of the provinces. For the former in fact, regression algorithms have been applied to *new positive cases*, *total cases* and *deaths*; for the latter instead, the regression algorithms have been applied for *new positive cases* and *total cases*.
Both linear and not linear algorithms were used for regression. The reason is that the linear models are always an initial guess with unknown datasets; also, they are simple and cheap from the computational point of view. However, linear models do not always capture the data patterns when data are too spread, such as in this study. The non-linear models used are Bagging Trees (Random Forest and Extra Trees) and Gaussian Process Regression, instead, the linear ones include LLS (Linear Least Squares) and its variations Lasso and ridge regression.
Once applied the regressions, the models have been compared with specific performance indicators (RMSE, r2 and error variance).
All three methods (linear, Gaussian and bagging trees) output not only the regression value but also the relevance that each feature has in the regression itself. In this way in fact, it was possible to cross-check the most important features among the different models, to reach conclusions about the virus spread, which was the primary goal of the thesis.

- As regards the provinces, the results showed that linear models did not prove to be sufficiently accurate, in fact for *new positive cases* the best performant algorithms are the Extra trees and Random Forest, while when regressing *total cases* with the same dataset, the best algorithm instead is the GPR (even if the results among GPR and ensemble trees are very close). Generally, GPR and the Bagging threes have higher accuracy when regressing *total cases* instead of *new positive cases*, while for linear regression it is the contrary, losing some points of r2 score when regressing *total cases*.

  Concerning feature selection, the most important features are 'number of inhabitants', number of 'S.A.R.' and 'cloud' for the regression of *new positive cases*; while, features as 'S.A.R.', 'number of farmed animals', 'number inhabitants' and 'income' are the most important for regressing *total cases*.

- As regards the regions the results showed that the regression of *deaths*, *new cases* and *total cases*, the results showed that the most performant algorithms are the bagging trees that report always the lowest error (in

terms of RMSE and error variance) and higher r2 value.
Among the three targets regressed, the one regressed with the lowest errors (RMSE and error variance) and higher r2 is *new positive cases*; the highest error is obtained when the *deaths* are regressed.

Concerning feature selection, the most important features when regressing the *deaths* are: 'nursing homes' and 'mobility'; also *total cases* are mostly explained by 'nursing homes' and 'mobility'; when regressing *new positive cases* the most important features are again 'nursing homes', 'S.A.R.', 'mobility' and 'number of inhabitants'.

In the end, one can claim that the most recurrent features are 'S.A.R', 'number of inhabitants' and 'number of animals farmed' for provinces, and 'nursing homes', 'mobility' and 'S.A.R.' for regions.
As already mentioned, at the province level the climatic conditions did not prove to be correlated with the regression of positive cases. The fact that the number of inhabitants is a good regressor for either the *positive cases* and the *deaths* is quite obvious: the more the people, the more possible the contagious. Moreover, the fact that 'mobility' appears to be a good regressor too, goes along with the Government directives followed during the lockdown, demonstrating that the less the people move, the less the epidemic can grow.
The presence of 'S.A.R.' and 'number of animals farmed' among the most relevant features for provinces instead is something that nobody ever mentioned. The idea behind the choice of this feature, at the beginning of data collection, was that these industries are the ones more correlated with air pollution and since pollution data could not be collected, S.A.R. was included.
On the other hand, the presence of S.A.R. in the territory can increase the local traffic, being a meeting place for many people. Hence, the correlation between 'S.A.R.' and COVID-19 could be either the pollution and the aggregation factor.
In conclusion, it is important to highlight that the obtained results are relative to the so called 'first wave' of infection, regarding the disease cases held in the period February-April (in Italy). Since the pandemic is still ongoing, and as time goes by, the datasets built with new data (same features but considered in different periods such as Summer or Autumn) could give other results both in terms of regression accuracy and feature relevance.
Considering that, to have more appropriate results, it is advisable to wait until the end of the pandemic, in view of having data related to different periods and a different pandemic situation associated with it.

A future work that could be done is collecting data related to other countries and applying the same methods to those data. Spain or France are countries as well interesting as Italy: they observed the same virus diffusion diversification among their territories. In this way, the arising results could be valuable to enhance or undermine those obtained with this thesis.

WEB REFERENCES

[4] *ADMINSTAT Italia: Mappe tematiche, curiosità, confronti e classifiche per i comuni, le province e le regioni sulla base di 20 indicatori socio-demografici*. URL: https://ugeo.urbistat.com/AdminStat/it/it/classifiche/dati-sintesi/province/italia/380/1.

[7] *ASSAEROPORTI (Associazione italiana gestione aerporti) - Statistics*. URL: https://assaeroporti.com/statistiche/.

[8] Saptashwa Bhattacharyya. *Ridge and Lasso Regression: L1 and L2 Regularization*. URL: https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

[9] *Caratteristiche dei pazienti deceduti positivi all'infezione da SARS-CoV-2 in Italia*. URL: https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia.

[10] *Coefficient of determination*. URL: https://en.wikipedia.org/wiki/Coefficient_of_determination.

[11] *Community Mobility Reports. These reports will be available for a limited time, so long as public health officials find them useful in their work to stop the spread of COVID-19*. URL: https://www.google.com/covid19/mobility/.

[12] *Comuni italiani - la classifica dei redditi nei comuni italinai del 2017*. URL: http://twig.pro/la-classifica-dei-redditi-dei-comuni-italiani-del-2017/.

[13] *Comuni italiani - Mean age, oldness index, percentage of over 65. Data related to 2017*. URL: http://www.comuni-italiani.it/statistiche/eta.html.

[14] *Coronavirus disease 2019 (COVID-19) Situation Report –88*. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200417-sitrep-88-covid-191b6cccd94f8b4f219377bff55719a6ed.pdf?sfvrsn=ebe78315_6.

[15] *Coronavirus disease 2019 (COVID-19)Situation Report –88*. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200417-sitrep-88-covid-191b6cccd94f8b4f219377bff55719a6ed.pdf?sfvrsn=ebe78315_6.

[17] *COVID-19 pandemic lockdown in Hubei*. URL: https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdown_in_Hubei.

[20] *dati 2019 mobilità da ANAS*. URL: https://www.stradeanas.it/sites/default/files/pdf/Anas%20Dati%20TGMA%202019.pdf.

[22] *ENAC: Ente Nazionale per l'Aviazione Civile Italian Civil Aviation Authority*. URL: https://www.enac.gov.it/aeroporti/infrastrutture-aeroportuali/aeroporti-in-italia.

[23] *Ensemble Learning to Improve Machine Learning Results*.

[26] *Github repository for COVID-19 data: province data*. URL: https://github.com/pcm-dpc/COVID-19/tree/master/dati-province.

[27] *Github repository for COVID-19 data: regions data*. URL: https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni.

[30] *ISTAT: Aspetti della vita quotidiana - Persone : Indice di massa corporea - regioni e tipo di comune. Data related to 2018*. URL: http://dati.istat.it/.

[31] *ISTAT: Banche dati/ambiente ed energua / ambiente nelle città. Data related to 2012*. URL: http://dati.istat.it/.

[32] *ISTAT (in Salute e sanità/Servizi sanitari e loro ricorso/Strutture sanitarie residenziali e semiresidenziali): Assistenza sanitatia di base e strutture sanitarie residenziali e semiresidenziali. Type of information: total beds, related to 2017*. URL: http://dati.istat.it/.

[33] *La Cina dice che il nuovo coronavirus non sarebbe "nato" al mercato del pesce di Wuhan*. URL: https://www.wired.it/scienza/medicina/2020/05/29/coronavirus-salto-specie-wuhan/?refresh_ce=.

[35] *Mappa italiana industrie pericolose: il nuovo Rapporto 2013 di ISPRA*. URL: https://figliodellafantasia.wordpress.com/2013/07/09/mappa-italiana-industrie-pericolose-il-nuovo-rapporto-2013-di-ispra/.

[36] *Metereological data*. URL: https://it.weatherspark.com/.

[38] *Population density in the Italian regions*. URL: https://www.tuttitalia.it/regioni/densita/.

[40] *Presidenza del Consiglio dei Ministri Dipartimento della protezione civile - website*. URL: http://www.protezionecivile.gov.it/dipartimento.

[41] *Regioni Italia per Popolazione*. URL: http://www.comuni-italiani.it/regionip.html.

[43] *Sanity insudtry, trade, restauration and accomodation services - ISTAT data, fro 2017*. URL: http://dati.istat.it/Index.aspx?DataSetCode=DICA_ASIAUE1P#.

[44] *scikit-learn Machine Learning in Python*. URL: https://scikit-learn.org/stable/.

[45] *Scikitlearn library - Matern kernel*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.Matern.html#rc15b4675c755-1.

[46] *Scikitlearn library - RBF kernel*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html#examples-using-sklearn-gaussian-process-kernels-rbf.

[48] *Sempre meno infermieri e operatori sanitari: tutti i dati regione per regione. 2018 data.* URL: https://www.infodata.ilsole24ore.com/2018/12/14/sempre-meno-infermieri-e-operatori-sanitari-tutti-i-dati-regione-per-regione/.

[49] G. Settimo M.E. Soggiu M. Masocco A. Spinelli. *Inquinamento atmosferico e diffusione del virus SARS-CoV-2*. URL: https://www.epicentro.iss.it/coronavirus/sars-cov-2-inquinamento-atmosferico.

BIBLIOGRAPHY

[1] *A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19*. DOI: https://doi.org/10.1101/2020.03.11.986836. URL: https://www.biorxiv.org/content/10.1101/2020.03.11.986836v1.abstract.

[2] *A double epidemic model for the SARS propagation*. 2003. DOI: 10.1186/1471-2334-3-19. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC222908/.

[3] *Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner*. URL: https://arxiv.org/abs/2002.05534.

[5] Christian L. Althaus. *Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa*. 2014. DOI: 10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288.

[6] *Applying discrete SEIR model to characterizing MERS spread in Korea*. URL: https://doi.org/10.1142/S1793962316430030.

[16] *COVID-19 Outbreak Prediction with Machine Learning*. URL: https://poseidon01.ssrn.com/delivery.php?ID=123003017068028117076004098015084002035005000074066087090122001004123086030024024119100028043014103061021010001017066001070080000053057080086025026070107111076000091063028037025090084095125070075111113124072093003111126071126071125119122112006002099124&EXT=pdf.

[18] *COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images*. DOI: 10.1101/2020.05.09.20096560. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273278/.

[19] S. Ratnesar-Shumate G. Williams B. Green M. Krause B. Holland S. Wood J. Bohannon J. Boydston D. Freeburger I. Hooper K. Beck J. Yeager L.A Altamura J. Biryukov J. Yolitz M. Schuit V. Wahl M. Hevey P. Dabisch. *Simulated Sunlight Rapidly Inactivates SARS-CoV-2 on Surfaces*. Vol. 222, 214–222. DOI: https://doi.org/10.1093/infdis/jiaa274. URL: https://academic.oup.com/jid/article/222/2/214/5841129.

[21] X. Wu R.C. Nethery M.B. Sabath MA D. Braun F. Dominici. *COVID-19 PM2.5 A national study on long-term exposure to air pollution and COVID-19 mortality in the United States*. URL: https://projects.iq.harvard.edu/covid-pm.

[24] *Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data*. DOI: https://doi.org/10.1016/j.neuroimage.2015.06.008. URL: https://www.sciencedirect.com/science/article/abs/pii/S105381191500484X.

[25] *Extractingpossibly representativeCOVID-19Biomarkers from X-Ray images with Deep Learning approachand image data related to Pulmonary Diseases*. URL: https://arxiv.org/ftp/arxiv/papers/2004/2004.00338.pdf.

[28] Sara H. Kassani et al. *Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach*. URL: https://www.researchgate.net/publication/340859631_Automatic_Detection_of_Coronavirus_Disease_COVID-19_in_X-ray_and_CT_Images_A_Machine_Learning-Based_Approach.

[29] ISTAT. "Agricoltura". In: *Annuario statistico italiano 2018* 13 (2018). URL: https://www.istat.it/it/archivio/225274.

[34] David J.C. MacKay. *InformationTheory, Inference,and LearningAlgorithms*. 2003. URL: http://www.inference.phy.cam.ac.uk/mackay/itila/.

[37] *Modeling Zika Virus Transmission Dynamics: Parameter Estimates, Disease Characteristics, and Prevention, authors=M. Rahman1 K. Bekele-Maxwell, L.L. Cates3, H.T. Banks N.K. Vaidya, year=2019, url=https://doi.org/10.1038/s41598-019-46218-4,*

[39] *Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers*. DOI: 10.1093/annonc/mdz108. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594459/.

[42] *Repurposing Therapeutics for COVID-19: Rapid Prediction of Commercially available drugs through Machine Learning and Docking*. DOI: https://doi.org/10.1101/2020.04.05.20054254. URL: https://www.medrxiv.org/content/10.1101/2020.04.05.20054254v2.

[47] *SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7277829/.

[50] Jerome Friedman Trevor Hastie Robert Tibshirani. *The Elements of Statistical Learning*. Ed. by Springer.

[51] C. E. Rasmussen C. K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006. ISBN: 026218253X.