UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SARS-CoV-2(Covid-19)pandemic data analysis with the use of machine learning algorithms

Diploma Thesis

Giorgos Peramatzis

Supervisor: Dimitrios Katsaros

Volos 2020

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SARS-CoV-2(Covid-19)pandemic data analysis with the use of machine learning algorithms

Diploma Thesis

Giorgos Peramatzis

Supervisor: Dimitrios Katsaros

Volos 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάλυση δεδομένων της πανδημίας του SARS-CoV-2(Covid-19) με τη χρήση αλγορίθμων μηχανικής εκμάθησης.

Διπλωματική Εργασία

Γεώργιος Περαματζής

Επιβλέπων/πουσα: Δημήτριος Κατσαρός

Βόλος 2020

Approved by the Examination Committee:

Supervisor **Dimitrios Katsaros**

Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Michael Vasilakopoulos**

Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Christos Antonopoulos**

Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Date of approval: 20-9-2020

Acknowledgements

I would like to warmly thank my professor Mr. Dimitrios Katsaros for our excellent cooperation, the pleasant atmosphere which is extremely important but mainly for the trust he showed me from the first moment, as well as for the help he gave me with his advice , his knowledge and his guidance. He was always there for what I needed and that is another reason why I should sincerely thank him again. For all these reasons, I will always be grateful for our cooperation .

DISCLAIMER ON ACADEMIC ETHICS
AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Giorgos Peramatzis
15-9-2020

Abstract

The new SARS-CoV-2 coronavirus and the disease that causes (COVID-19) have swept across the globe and tested every country's health infrastructure. Countries have taken extreme measures to limit the spread of the disease despite the great socio-economic problems caused by them. Monitoring the new coronavirus disease remains vital and that is why the entire scientific community has focused all its efforts on tackling it.In this thesis, we tried to understand the behavior of the new coronavirus but also to predict future cases using various machine learning algorithms as well as neural networks in order to correctly and timely predict the epidemiological burden in Greece.

Περίληψη

Ο νέος κορωνοϊός SARS-CoV-2 και η ασθένεια που προκαλεί (COVID-19) έχει σαρώσει στο πέρασμα του όλο τον πλανήτη και δοκιμάζει τις υποδομές υγείας κάθε χώρας.Οι χώρες κατέφυγαν σε ακραία μέτρα για τον περιορισμό της εξάπλωσης της νόσου παρά τα μεγάλα οικονομικοκοινωνικά πρόβληματα που προκαλούνται απο αυτά . Η παρακολούθηση της νέας νόσου του κορωνοϊού παραμένει ζωτικής σημασίας και γι'αυτό τόν λόγο όλοι η επιστημονική κοινότητα έχει στρέψει όλες τίς δυνάμεις της στην αντιμετωπίση του.Σε αυτήν την διπλωματική , προσπαθήσαμε να κατανοήσουμε την συμπεριφορά του νέου κορωνοϊού αλλα και να προβλέψουμε μελλοντικά κρούσματα χρησιμοποιώντας διάφορους αλγόριθμος μηχανικής μάθησης καθώς και νευρωνικά δίκτυα με σκοπό την σωστή και έγκαιρη πρόβλεψη του επιδιομιολογικού φορτίου στην Ελλάδα.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The new coronavirus has entered our lives for good and has terrified the whole world, so it is imperative that the new vaccines that are just around the corner work, but it is also imperative its understanding as well as its contagious rate[1, 2]. SARS-CoV2 first appeared in late 2019 in a Chinese city on Wuhan[3, 4, 5], although it is not yet clear where it first started and how it was transmitted to humans[6].SARS-CoV-2 is a single-stranded RNA virus and is genetically close to other coronaviruses such as SARS-CoV and MERS-CoV [6, 7]. Bats are considered to be the predominant source of origin for SARS-CoV2 as well as for the other two [6]. Human infection most likely occurred through vectors, such as pangolins in the case of SARS-CoV-2 [5].Furthermore, In just a few months, the World Health Organization described this new threat as a pandemic on 11 March 2020 [8] and by 13 December 2020 more than 70 million people have fallen ill and more than 1.6 million have died[9, 10]. Recent studies have shown that antibodies produced in asymptomatic individuals tend to weaken after two to three months [11, 12]. Antibodies are a critical component of immunity and a good vaccine will try to reproduce this type of natural protection while extending its action.As it is easily understood, it is very important to be able to predict its future behavior based on some data, however this is not at all an easy and simple task because of its contagion depends on many different factors.To the best of our knowledge, the main method of spreading COVID-19 is through human-to-human respiratory droplets[13].According to [14] the correlation between different socio-economic ,meteorological variables and the weekly number of COVID-19 cases observed shows that 8 independent variables below the total population, household income, occupation and transport categories are highly correlated with the cases of COVID-19. It is also worth noting that all countries are taking draconian measures, such as large-scale quarantines and lockdowns, which are responsible for the estimated economic damage of 3-10 trillion dollars in the global economy [15], in an effort to reduce the spread of covid19 in order to save the required time until a large part of the population is vaccinated.In addition, most countries have taken these harsh health measures more than once, Greece is no exception. Although the first time went quite well and the reduction of cases was fast and successful, in the second it shows that the reduction seems to be done very slowly [16]. This shows us how dependent this virus is on the various circumstances and why it is so dangerous. For all these reasons, anticipation of upcoming cases is considered necessary so that new measures can be planned in a timely manner or when existing ones need to be relaxed.

1.1 Objective of the thesis

In this research I will try to forecast the the daily Covid-19 cases in Greece. To do so I will use various algorithms from various field. I will also use an epidemiological model to predict the outcome in different scenarios.

1.1.1 Contribution

The contribution of this thesis can be summarized as:

1. Creation of statistical, machine learning and deep learning models.
2. Creation of of epidemiological SIR-D model.
3. The models tested against various parameters
4. The CNN model resulted as the model with the best accuracy.

1.2 Organization of the thesis

In the Chapter 2 I will explain all the models that will be used in this thesis. In the Chapter 3 I will create an epidemiological model to forecast the cases. In the Chapter 4 I will conduct experiments using the models and show the results. In the Chapter 5 I will analyze the results of the previous chapter.

Chapter 2

Theoretical Background

The order of the models that will be used is important and is structured in increasing complexity and from classical to modern methods. The statistic approaches are simple and give good results fast. Machine learning approaches used in the research are slower and more complex, than the first approaches, but also have a higher bar to clear to be skillful.

2.1 Statistic Models

2.1.1 SimpleExpSmoothing

Simple exponential smoothing is a forecasting method for univariate data without trend or seasonality[17]. It is a simple model for predicting a time series, with the basic idea that the future will be about the same as the recent past.

2.1.2 ExponentialSmoothing

Exponential smoothing is a time series forecasting method for univariate data.Exponential smoothing methods for forecasts use weighted averages of previous observations, the weights decreasing exponentially with the age of the observations[18]. Thus, the relative weight is greater when the observation is more recent.

2.1.3 ARMA

Autoregression (AR) model is a time series model that uses observations from previous time steps as input to a regression equation to predict the value in the next step. The Moving

Average (MA) model attempt to record the shock effects observed in white noise conditions. These shock effects could be considered as unexpected events that affect the observation process. This means that the MA(q) model does not use previous predictions to predict future values like the AR(p) model, but uses errors from previous predictions. The ARMA model is simply a merger of the above models[19].

2.1.4 HOLT

The Holt two-parameter model, also known as linear exponential smoothing, is a smooth- ing model for predicting trend data[20]. Holt's model has three distinct equations ,a forecast equation and two smoothing equations (one for the level and one for the trend), that work together to create a final prediction [20].

2.1.5 ARIMA

AutoRegressive Integrated Moving Average (ARIMA) model is a generalization of AutoRegressive Moving Average (ARMA) that adds the concept of integration. By using the raw observation differentiation we succeed in making the time series stationary[21].

2.1.6 SARIMAX

The Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) model is the ARIMA model with the addition of exogenous variables, which can be any variables we are interested in [22]. Now, Seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX) model is the ARIMAX with the seasonal aspect. Seasonality is very important to consider when certain patterns are inconsistent and occur periodically[23, 24].

2.2 Machine Learning

The deep learning neural networks are able to automatically learn arbitrary complex mappings from inputs to outputs and support multiple inputs and outputs. These are powerful features that offer a lot of promise for time series forecasting, particularly on problems with complex-nonlinear dependencies, multivalent inputs, and multi-step forecasting.

2.2.1 KNeighbors Classifier

The nearest neighbor k algorithm (KNN) is a supervised machine learning algorithm that can be used to solve classification and regression problems. To classify an unknown query, the k-NN classifier calculates the distances between the query and all attributes in the training data set[25]. The KNN algorithm assumes that similar things are close.

2.2.2 Decision Tree Classifier

Like KNN, the decision tree is a supervised algorithm that can be used to either regression or classification problems. The decision tree is used to create a training model that can be used to predict the order or value of the query by learning simple decision rules. In decision trees,we always start from the root of the tree to predict a label. We compare the values of the root with the attribute of the record[26]. Based on the comparison, we follow the branch that corresponds to this value and go to the next node.

2.2.3 Support Vector Machine

Support Vector Machine (SVM) is also an algorithm used to solve sorting and regression problems. Although it is a linear model, it can nevertheless solve linear and non-linear problems. This can be done by viewing the data in a higher dimension in which is linearly separated. The algorithm creates a line or hyperplane that divides the data into categories[27].

2.2.4 GaussianNB

Bayes Theorem assumes that each input variable is dependent on all other variables[28]. However, this greatly complicates the calculations, which makes it difficult to manage. However, if we consider that each input variable is independent, then the model goes from a conditional probability dependent model to a conditional independent probability model. This simplification is often found under the name Naive Bayes.Now , if we make the assumption that the continuous values associated with each class are distributed according to a Gaussian distribution,then we have the Gaussian Naive Bayes algorithm[29].

2.2.5 Bagging Classifier

In the bagging classifier, a sample is selected from a set, using the replacement method to ensure that the selection is random. The learning algorithm is then executed on the selected samples and then the predictions are aggregated and combined so that the final prediction examines all possible outcomes[30]. It is a general procedure that can be used to reduce variance and avoid overfitting.

2.2.6 Random Forest Classifier

Random forest is a supervised learning algorithm.It is an ensemble of decision trees, usually trained with the "bagging" method which means that reduces variances and overfitting [31].The random forest algorithm creates decision trees in data samples and then takes the prediction from each of them and finally selects the best solution by voting[31, 32] .

2.2.7 Extra Trees Classifier

Extra trees classifier is an ensemble method and it is very similar to Random Forest.The differences are that the Extra tree selects a split point at random, unlike Random Forest which uses a greedy algorithm to select an optimal split point[32]. In addition, the Extra Trees algorithm fits each decision tree in the entire training data set, as opposed to the random forest that uses a bootstrap sample to create each decision tree[32].

2.2.8 Gradient Boosting Classifier

Gradient boosting is a machine learning technique for regression and classification problems that uses the form of a set of weak prediction models to generate a prediction [33]. It builds the model in a gradual manner as other boosting methods do and generalizes it allowing for the optimization of an arbitrary differentiable loss function.Gradient reinforcement basically combines weak students into a single strong student in a repetitive way [34].

2.3 Deep learning Neural Networks

2.3.1 MLP

Multilayer Perceptrons (MLP) are widely used in the field of time series forecasting. MLP consists of at least one input level, one output level and one hidden level. A multi perceptron network is also a feed-forward network and uses a supervised learning technique called backpropagation for training [35, 36].

2.3.2 Convolutional Neural Networks

Convolutional Neural Network is a feed-forward neural network and is an improvement over MLP. The difference between CNN and MLP is that CNN uses the concept of weight distribution. The advantage of CNN is that its weight number does not have to be as large as for a fully connected construction [37].As a result, training is relatively easier and im- portant features are extracted more efficiently. CNN consists of four levels an input layer, a converging layer, a concentration layer, and a fully connected layer.

2.3.3 Long Short-Term Memory

Recurrent neural networks like the Long Short-Term Memory network add the explicit handling of order between observations when learning a mapping function from inputs to outputs[38].In addition to individual data points, entire data sequences can be easily processed.Moreover, each network in the loop receives input and information from the previous network, performing the specified function, and generates an output by feeding the information to the next network[39].

2.3.4 CNN-LSTM

A CNN is a type of neural network developed primarily for the solution of two-dimensional data. However, CNN can also be very effective in learning features from one-dimensional sequence data, such as univariate time series data. Its combination with an LSTM backend, where CNN is used to interpret the input sequences provided together as a sequence in an LSTM model for interpretation gives us the CNN-LSTM hybrid model[40].

2.4 Evaluation

The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

$$\frac{1}{n}\sum_{t=1}\left|\frac{A_t - F_t}{A_t}\right|$$

Time series forecasting models can be evaluated on a test set using walk-forward validation. Walk-forward validation is an approach where the model makes a forecast for each observation in the test dataset one at a time. After each forecast is made for a time step in the test dataset, the true observation for the forecast is added to the test dataset and made available to the model. Simpler models can be refit with the observation prior to making the subsequent prediction. More complex models, such as neural networks, are not refit given the much greater computational cost. Nevertheless, the true observation for the time step can then be used as part of the input for making the prediction on the next time step.

Chapter 3

Epidemic Models

By definition, compartmental models simplify the mathematical modelling of infectious diseases. The population is assigned to compartments with labels – for example, S, I, or R, (Susceptible, Infectious, or Recovered). People may progress between compartments. The order of the labels usually shows the flow patterns between the compartments. for example SEIS means susceptible, exposed, infectious, then susceptible again.

With these models we will try to predict things such as how a disease spreads, or the total number infected, and to estimate various epidemiological parameters such as the reproductive number. Those models can also be used to show how different public health interventions may affect the outcome of a epidemic, but those scenarios will not be covered in this thesis.

3.1 SIR

The simplest model that can be created for the epidemics is the Susceptible Infected Recovered (SIR) model. This is an epidemiological model that computes the theoretical number of people infected with a contagious illness in a closed population over time. The model consists of three compartments:

* S: Susceptible (= Population - Confirmed)

* I: Infected (= Confirmed - Recovered - Fatal)

* R: Recovered or Fatal (= Recovered + Fatal)

Model:

$$S \xrightarrow{\beta I} I \xrightarrow{\gamma} R$$

β: Effective contact rate γ: Recovery(+Mortality)

3.2 SIR-D

In some cases the people that are infected with a virus are not recovering, but they are dying. The model that also takes this into consideration is the Susceptible Infectious Recovered Deceased (SIR-D) model. The model consists of four compartments:

* S: Susceptible (= Population - Confirmed)
* I: Infected (= Confirmed - Recovered - Fatal)
* R: Recovered
* D: Fatal

Model:

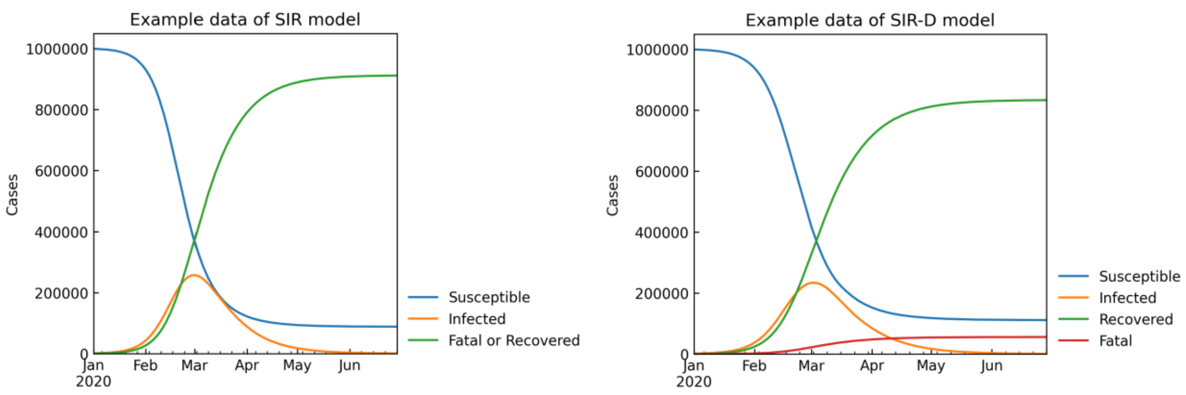
$$S \xrightarrow{\beta I} I \xrightarrow{\gamma} R$$

$$I \xrightarrow{\alpha} D$$

α: Mortality rate β: Effective contact rate γ: Recovery rate

3.3 Reproduction Rate

In the field of epidemiology, the basic reproduction number, denoted as R_0, of an infection can be thought of as the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. In an epidemic, when $R_0 > 1$ means that the infection is starting to spread in a population, but not if $R_0 < 1$. As a result, the larger the value of R_0, the faster the epidemic is spreading.



(α) SIR model

(β) SIR-D model

Figure 3.1: Examples of the models

3.4 Different Scenarios

For this section we will use a data set that consists of Date, Confirmed, Infected, Fatal, Recovered values and the library [41] to create an epidemic model to analyze and to also test in different scenarios.

In the period of the pandemic in Greece various measure have been taken. In the table 3.1 we can see the most drastic ones that were taken to stop the spread.

Date	Measures
18/3	All shops are closed
23/3	1st catholic lockdown
4/5	Shops are opening, movement is allowed
25/5	Cafeterias are opening
1/07	Borders opened for tourists
7/11	The 2nd catholic lockdown

Table 3.1: Some of the measures issued by the Government

In the scenarios that we going to examine, the model that we are going to use of the ones that we mentioned above is the SIR-D, because some patients that are being infected by the virus are not recovering and they are dying.

Using these data and the SIR-D model as the prεffered model we tried to calculate the parameters of the model. The parameters that are going to be calculated are α, β and γ as mentioned above. As a first step the model separated the time elapsed in phases depending on the R_0 and we can see the periods in the Figure 3.2.

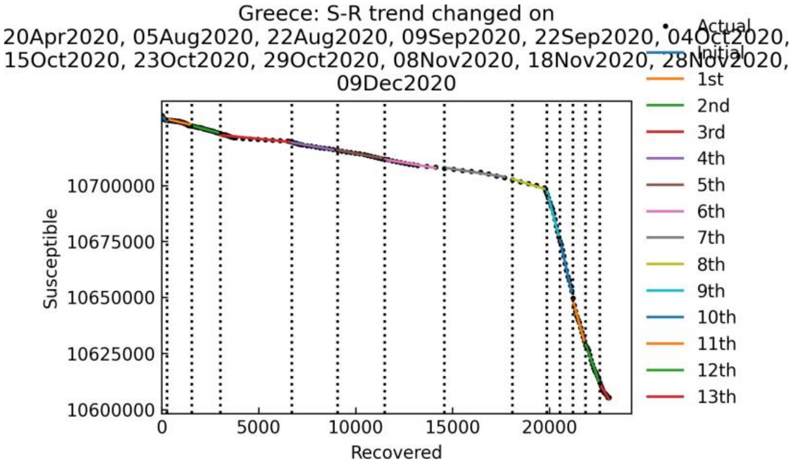


Figure 3.2: Phases of the epidemic

Trying to interpret the figure we can notice that the phase 8 is the first phase that the virus has a large R_0 and the virus is starting to spread faster.

Using the data that are gathered and the parameters that are calculated from the model we will create different scenarios to check what would happen in each case.

The first scenario that we are going to examine is that the 2nd catholic lockdown didn't occur and the virus continues to spread.

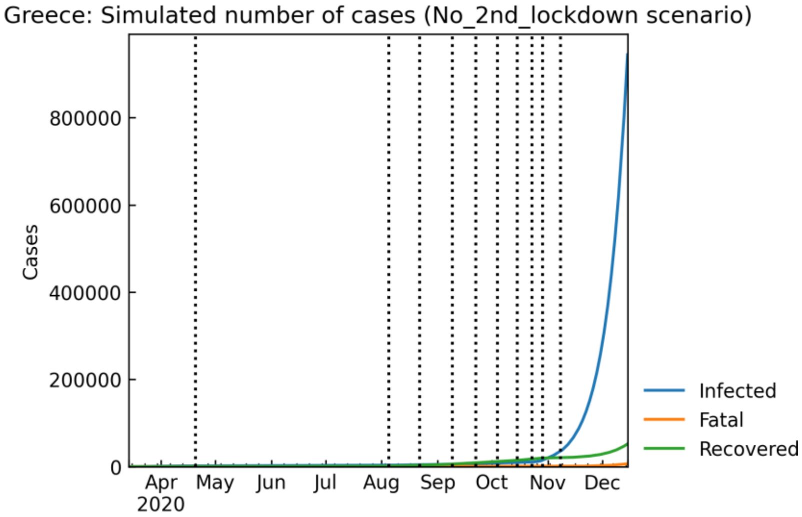


Figure 3.3: Curve of the model in scenario no lockdown

At the Figure 3.3 we are seeing how the virus would be if the second lockdown have not been issued. As we can see the number of infected has sky-rocked summing up above 80 %

of the total population of Greece.

In the second scenario that we are going to examine we are supposing that the second lockdown will be over in the first month.

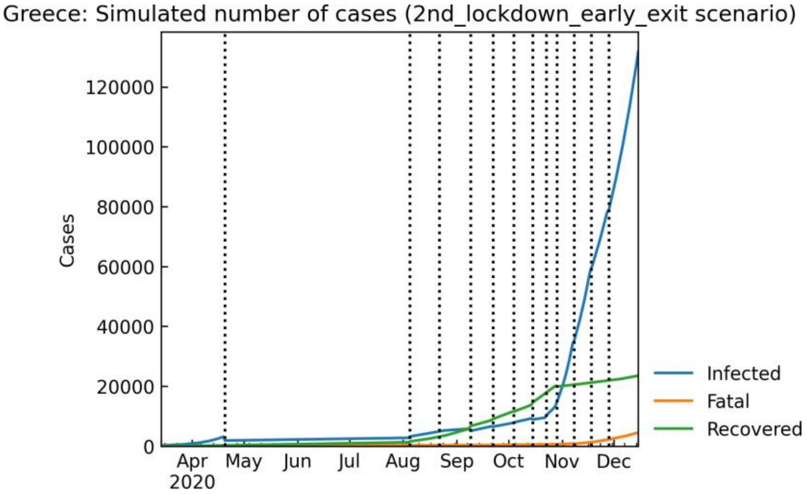


Figure 3.4: Curve of the model in scenario of early termination of the lockdown

At the Figure 3.4 we are seeing that the number of infected people is significantly lower than the scenario above and only a month of restricted social interactions can have a huge impacted on the spread of the virus. By that we conclude that social distancing is working.

At the Figure 3.5 we can see the real values of the Infected Recovered and Fatal to have a better understanding of how the R_0 of each scenario is changed.

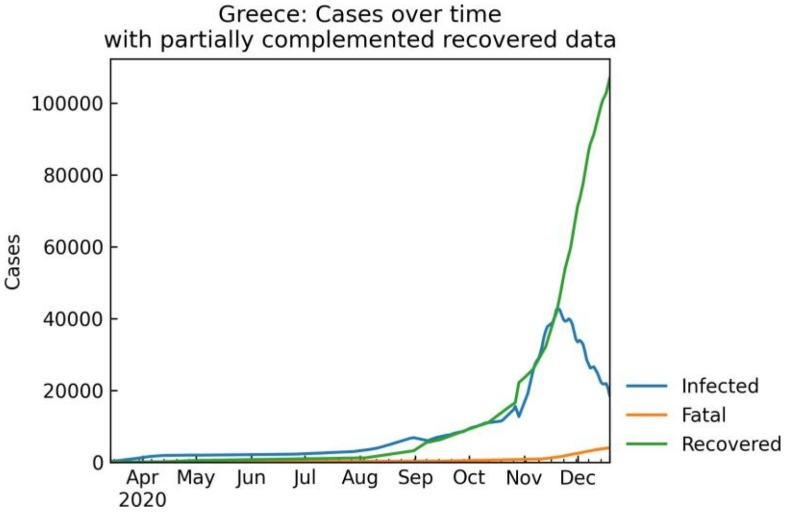


Figure 3.5: Curve of the model in scenario of early termination of the lockdown

Chapter 4

Results

4.1 Setup

The machine learning models and neural networking models have a big variety of parameters and a grid search has been conducted trying to find the best parameters for each model for our data. The two important factors that are being searched in this research are the: days of forecasting, and the steps for historical reference

A way of performing hyper-parameter optimization is grid search which is simply an exhaustive searching through a manually specified subset of the hyper-parameter space of a learning algorithm. As an evaluation method for measuring the performance of the model in each set of parameters we are using Mean absolute percentage error (MAPE) as described in the above sections.

We will not implement the grid search algorithm at the statistic models because most of them have a small window for optimization except of the ARIMA and SARIMA model, but we will not investigate them further in this research.

Let’s explain further the usage of those two factors and importance of them :

- Days of forecasting, is referring to the number of days that we are trying to forecast in the future. The incubation period of the virus is 14 days according to scientists, so the results of a social gathering will be shown is a spawn of 14 days. So trying to forecast further than 14 days of the last update of the model may be a drawback of the model. If a model takes time to be created we don’t want to be irrelevant in a short period of time. We will investigate the difference between 14 and 20 days of forecasting in the future.

- Steps for historical reference, is the sequential input of data into the model that will analyze and try to find a pattern on. If this period is too long we are feeding outdated data and if too short we will loose the pattern. We will investigate this parameter in the range of 5 to 12.

For the CNN-LSTM model further search will be conducted and we will also add more variables in the grid search of the best hyper-parameters like the filter and the kernel.

We will have all the models predicting 14 and 20 days into the future.

4.1.1 Forecasting using statistic models

We will start forecasting using the statistic models. Is important that we can consider of the setup of the ARIMA and SARIMA models because those two use trend and seasonality.

The configuration of ARIMA is : order=(1, 1, 1)

The configuration of SARIMA is: order=(1, 1, 1), seasonal order=(1, 1, 1, 1)

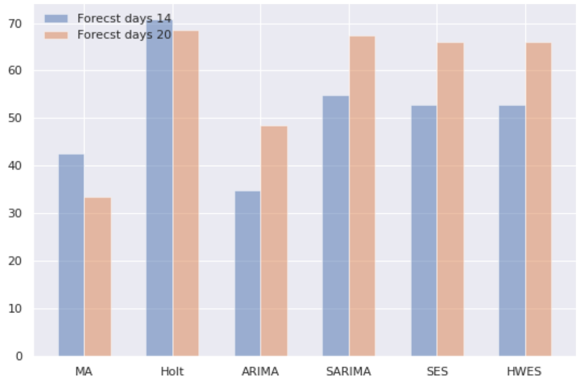


Figure 4.1: Bar plot of the statistical models

In the image 4.1 we can all the statistic methods that we defined above for the two cases. We can notice that ARIMA and SARIMA models due have better results in the long run.

Holt algorithm has the best results in both cases, with 70.73 % of success in the span of 14 days.

4.2 Forecasting using ML-NN models

The supervised algorithms and the neural networks have a variance in the results depending in the initialized state of the model/network, so for better statistical results, each algorithm in each state is being run 10 times and the mean of those results is used as the model's forecast. A more detailed summary of the neural networks models used in this section can be found in Appendix A.

In the list of the supervised algorithms described in the above sections the algorithm that has the best results across all the parameters is the k-nearest neighbors (kNN).

In the image 4.2 we can see the results of the kNN algorithm using all set of different parameters. In all the forecasts the model has better performance predicting 20 days into the future, except the case that the input of the model has historical reference of 12 days. The algorithm predicting 14 days into the future has double the percentage of accuracy of the other case.The best result is in the forecast of 14 days into the future and with 8 days of time step, historical reference, and with 86.52 % of success.

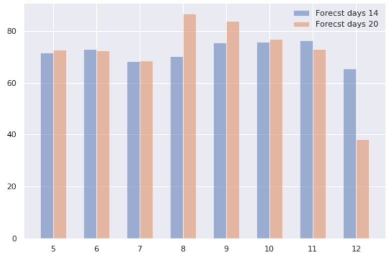
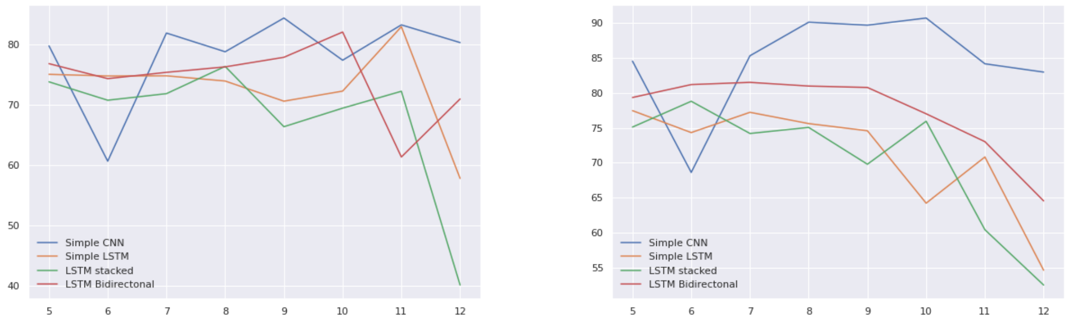


Figure 4.2: Bar plot of the kNN model

For our next forecast we are going to use the the deep learning algorithms that we mentioned above.

In the Figure 4.3 we can the accuracy of the models in all the set of parameters. The best algorithm in both forecast cases is the Convolutional Neural Network (CNN). In the case of 14 days forecast in the future and with 9 days of historical reference the model has accuracy of 84.4 %. In the case of 20 days forecast in the future and with 10 days of historical reference the model has accuracy of 90.7 %.

We can also notice that when we are feeding the models data with large historical reference of 11 and 12 days and try to predict further in the future we have poor prediction accuracy.



(α) 14 days forecast

(β) 20 days forecast

Figure 4.3: Deep learning models

Last but least, we tried the CNN-LSTM model. In this model more set of parameters than the other models have been tested, this is also the most complex model that we have tried. As a result the model didn't perform better neither from the machine learning models nor from the other deep learning models.

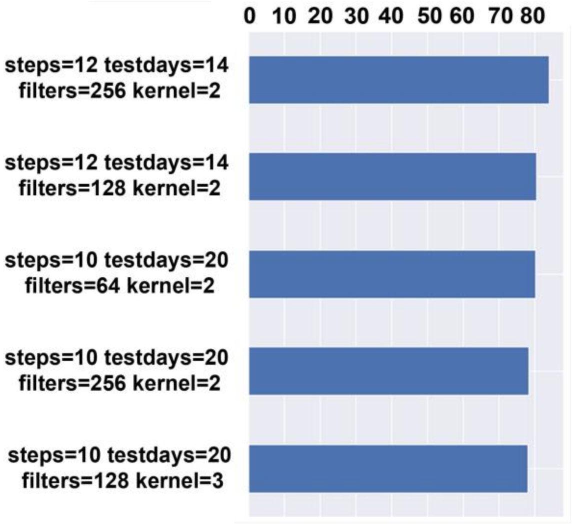


Figure 4.4: CNN-LSTM model

In the Figure 4.4 we can see a bar plot containing the 5 best set parameters for the model, we can also see the set of parameters that gave those results. The best accuracy that the model managed was 83.9 %.

Chapter 5

Conclusions

Seeing the Figure 5.1 of the daily infected patients we can see that is difficult to insert a line that fits the data. So it is clear that the linear models had a difficulty given precise predictions.

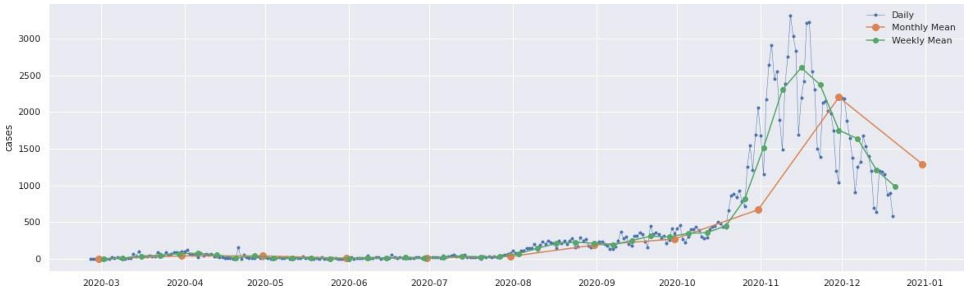


Figure 5.1: Daily cases

From the result of experiments in the previous section we can summarize that depending on the depth of time that you want to predict you have to provide the corresponding sequence of historical data. We can also conclude that the algorithm with the correct combination of training time and accuracy is the CNN model.

If we want to create a model that gives information about the spread of the disease in order to use this information in our advantage trying to stop the spread of the virus, is the SIR-D model.

Bibliography

[1] Landscape of covid-19 candidate vaccines. available. https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines. Ημερομηνία πρόσβασης:10-10-2020.

[2] Fatima Amanat and Florian Krammer. Sars-cov-22 vaccines: Status report. *Immunityy*, 52(4):583 – 589, 2020. doi: httpss://doi.org/10.1016/j.immuni.2020.03.007.

[3] Na Zhu, Dingyu Zhang, Wenling Wang, Xinwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, Peihua Niu, Faxian Zhan, Xuejun Ma, Dayan Wang, Wenbo Xu, Guizhen Wu, and George Gao. A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 382:727–733, 01 2020. doi :10.1056/NEJMoa2001017.

[4] Centers for disease control and prevention. 2019 novel coronavirus, wuhan, china: 2019-ncov situation summary. 2020. https://www.cdc.gov/csels/dls/locs/2020/outbreak-of-2019-novel-coronavirus-2019-ncov-in-wuhan-china.html.

[5] Taoo Zhang, Qunfu Wu, and Zhigang Zhang. Probablee pangolin origin of sars-cov-2 associated with the covid-19 outbreak. *Currentt Biology*, 30(7):1346 – 1351.e2, 2020. doi : httpss://doi.org/10.1016/j.cub.2020.03.022.

[6] Rambaut Andrew Lipkin W. Ian Holmes Edward C. Garry Robert F. Andersen, Kristian G. The proximal origin of sars-cov-2. *Nature Medicine*, 26:450–452, 2020. doi:10.1038/s41591-020-0820-9.

[7] Hogan Michael J. Porter Frederick W. Weissman Drew Pardi, Norbert. mrna vaccines — a new era in vaccinology. *Nature Reviews Drug Discovery*, 17:261–279, 2018. 10.1038/nrd.2017.243.

[8] Who (world health organization). (2020) world health organization media brief- ing. director general's opening remarks at the media briefing on covid-19 - 11 march 2020. `https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020`. Ημερομηνία πρόσβασης: 15-10-2020.

[9] The johns hopkins coronavirus resource center (crc), updated source of covid-19. `https://coronavirus.jhu.edu/map.html`. Ημερομηνία πρόσβασης: 15-12-2020.

[10] Who coronavirus disease (covid-19) dashboard. `https://covid19.who.int/`. Ημερομηνία πρόσβασης: 15-12-2020.

[11] Jeffrey Seow, Carl Graham, Blair Merrick, Sam Acors, Kathryn Steel, Oliver Hemmings, Aoife O'Bryne, Neophytos Kouphou, Suzanne Pickering, Rui Galão, Gilberto Betancor, Harry Wilson, Adrian Signell, Helena Winstone, Claire Kerridge, Nigel Temperton, Luke Snell, Karen Bisnauthsing, Amelia Moore, and Katie Doores. Longitudinal evaluation and decline of antibody responses in sars-cov-2 infection, 07 2020. MedRxiv 2020. [CrossRef].

[12] Javier Ibarrondo, Jennifer Fulcher, David Goodman-Meza, Julie Elliott, Christian Hofmann, Mary Hausner, Kathie Ferbas, Nicole Tobin, and Otto Yang. Rapid decay of anti–sars-cov-2 antibodies in persons with mild covid-19. *New England Journal of Medicine*, 383, 07 2020.

[13] Heshu Sulaiman Rahman, Masrur Sleman Aziz, Ridha Hassan Hussein, Hemn Hassan Othman, Shirwan Hama Salih Omer, Eman Star Khalid, Nusayba Abdulrazaq Abdulrahman, Kawa Amin, and Rasedee Abdullah. Thee transmission modes and sources of covid-19: A systematic review. *Internationall Journal of Surgery Open*, 26:125 – 136, 2020. doi : "https://doi.org/10.1016/j.ijso.2020.08.017.

[14] Deepro Pasha, Alex Lundeen, Dilruba Yeasmin, and M. Pasha. An analysis to identify the important variables for the spread of covid-19 using numerical techniques and data science. *Case Studies in Chemical and Environmental Engineering*, 3:100067, 06 2021.

[15] Covid-19 to slash global economic output by $8.5 trillion over next two years. `https://github.com/lisphilar/covid19-sir`. Ημερομηνία πρόσβασης: 15-12-2020.

[16] Δεδομένα για τον κορωνοϊό στην Ελλάδα. `https://covid19.gov.gr/covid19-live-analytics/`. Ημερομηνία πρόσβασης: 15-12-2020.

[17] Eva Ostertagova and Oskar Ostertag. Forecasting using simple exponential smoothing method. *Acta Electrotechnica et Informatica*, 12:62–66, 12 2012. doi: 10.2478/v10198-012-0034-2.

[18] Sourabh Shastri, Amardeep Sharma, Vibhakar Mansotra, Anand Sharma, Arun Bhadwal, and Monika Kumari. A study on exponential smoothing method for forecasting. *International Journal of Computer Sciences and Engineering*, 6:482–485, 04 2018.

[19] Time series analysis for financial data iv— arma models. `https://medium.com/auquan/time-series-analysis-for-finance-arma-models-21695e14c999`. Ημερομηνία πρόσβασης : 3-12-2020.

[20] Holt’s forecasting model. In P. M. Swamidass, editor, *Encyclopedia of Production and Manufacturing Management*, page 274. Springer US, 2000. doi:10.1007/1-4020-0612-8409.

[21] Robin John Hyndman and George Athanasopoulos. *Forecastingg: Principles and Practice*. OTextss, Australiaa, 2nd edition, 2018.

[22] Mofeng Yang, Jiaohong Xie, Peipei Mao, Chao Wang, and Ye Zhirui. Application of the arimax model on forecasting freeway traffic flow. 07 2017.

[23] Albert Ling, G. Darmesah, Khim Chong, and chong mun Ho. Application of arimax model to forecast weekly cocoa black pod disease incidence. *Mathematics and Statistics*, 7:29–40, 09 2019.

[24] Nari Arunraj, Diane Ahrens, and Michael Fernandes. Application of sarimax model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems*, 7:1–21, 04 2016.

[25] Subramaniam Dhanabal and Chandramathi SA. A review of various k-nearest neighbor query processing techniques. *Int. J. Comput. Appl.*, 3, 01 2011.

[26] Lior Rokach and Oded Maimon. *Decision Trees*, volume 6, pages 165–192. 01 2005.

[27] Theodoros Evgeniou and Massimiliano Pontil. Support vector machines: Theory and applications. volume 2049, pages 249–257, 01 2001.

[28] Bayes' theorem definition. `https://www.investopedia.com/terms/b/bayes-theorem.asp`. Ημερομηνία πρόσβασης : 11-12-2020.

[29] B. M. Gayathri and C. P. Sumathi. An automated technique using gaus-sian naïve bayes classifier to classify breast cancer. *International Journal of Computer Applications*, 148:16–21, 08 2016.

[30] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 08 1996. doi: 10.1007/BF00058655.

[31] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues(IJCSI)*, 9, 09 2012.

[32] An intuitive explanation of random forest and extra trees classifiers. `https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b`. Ημερομηνία πρόσβασης : 7-12-2020.

[33] Ioannis Tsamardinos Stefanos Fafalios, Pavlos Charonyktakis. Gradient boosting trees. Technical report, Gnosis Data Analysis PC, April 2020.

[34] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 12 2013. doi :10.3389/fnbot.2013.00021.

[35] Stanislaw Osowski, Krzysztof Siwek, and T. Markiewicz. Mlp and svm networks - a comparative study. volume 46, pages 37– 40, 02 2004. doi:10.1109/NORSIG.2004.250120.

[36] Tim Menzies, Ekrem Kocagüneli, Leandro Minku, Fayola Peters, and Burak Turhan. *Chapter 24. Using Goals in Model-Based Reasoning*. 12 2015. doi : 10.1016/B978-0-12-417295-1.00024-2.

[37] Maryam Abo-Tabik, Nicholas Costen, John Darby, and Yael Benn. Towards a smart smoking cessation app: A 1d-cnn model predicting smoking events. *Sensors*, 20:1099, 02 2020. doi :10.3390/s20041099.

[38] The promise of recurrent neural networks for time series forecasting. `https://machinelearningmastery.com/promise-recurrent-neural-networks-time-series-forecasting/`. Ημερομηνία πρόσβασης: 16-12-2020.

[39] Ashutosh Kumar Singh Jitendra Kumar, Rimsha Goomer. Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters. *Procedia Computer Science*, 125:676–682, 2018. doi: `https://doi.org/10.1016/j.procs.2017.12.087`.

[40] Carlos Iturrino, Francesco Grasso, Antonio Luchetta, Maria Piccirilli, Libero Paolucci, and Giacomo Talluri. A comparison of power quality disturbance detection and classifica- tion methods using cnn, lstm and cnn-lstm. *Applied Sciences*, 10:6755, 09 2020. doi : 10.3390/app10196755.

[41] Covsirphy. `https://github.com/lisphilar/covid19-sir`. Ημερομηνία πρόσβασης: 18-12-2020.

Appendix A

NN models summary

```
Model: "Simple CNN"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 9, 64)             320
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 4, 64)             0
_________________________________________________________________
flatten (Flatten)            (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 50)                12850
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 51
=================================================================
Total params: 13,221
Trainable params: 13,221
Non-trainable params: 0
_________________________________________________________________
Model: "Simple LSTM"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 50)                10400
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 51
=================================================================
Total params: 10,451
Trainable params: 10,451
Non−trainable params: 0
_________________________________________________________________
Model: ”LSTM Stacked”
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 12, 50)            10400
_________________________________________________________________
lstm_2 (LSTM)                (None, 50)                20200
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 51
=================================================================
Total params: 30,651
Trainable params: 30,651
Non−trainable params: 0
_________________________________________________________________
Model: ”LSTM Bidirectonal”
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional (Bidirectional (None, 100)               20800
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 101
=================================================================
Total params: 20,901
Trainable params: 20,901
```

```
Non−trainable params: 0
—————————————————————————————————————————————————————————————————
Model: ”CNN−LSTM”
—————————————————————————————————————————————————————————————————
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 10, 64)            256
—————————————————————————————————————————————————————————————————
conv1d_2 (Conv1D)            (None, 8, 64)             12352
—————————————————————————————————————————————————————————————————
dropout (Dropout)            (None, 8, 64)             0
—————————————————————————————————————————————————————————————————
max_pooling1d_1 (MaxPooling1 (None, 4, 64)             0
—————————————————————————————————————————————————————————————————
flatten_1 (Flatten)          (None, 256)               0
—————————————————————————————————————————————————————————————————
dense_7 (Dense)              (None, 100)               25700
—————————————————————————————————————————————————————————————————
dense_8 (Dense)              (None, 1)                 101
=================================================================
Total params: 38,409
Trainable params: 38,409
Non−trainable params: 0
————————————————————————————————————
```