Investigation on Correlation Between COVID-19 Spread and Government Restriction Methods in Greece and Prediction Modelling on Future Cases Using Machine Learning Algorithms.

A dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Science

By Alexandros C. Mitronikas

Department of Computer Science

City Unity College, Greece

University of Wales Institute, Cardiff

May 2021

ABSTRACT

This work has been conducted to investigate for any significant correlation between the spread of COVID-19 and the policy restriction and control methods taken by the administration in Greece, using machine learning algorithms and their predictive indications. It consists of the necessary research concerning the virus' nature, and strategically analyses through other machine learning approaches to best determine an efficient methodology for the issue. Using all available resources, the data is collected and further structured exclusively for this project's objectives and aims. Moving on, well known generalized but also specialized algorithms, models and figures are used to find significant insights concerning the structure of the collected data and potentially the optimal algorithm to make predictions with. Lastly, the results are discussed and critically analysed supporting the retrieved insights based on well-defined implementations, including the data pre-processing stages, hyperparameter tunings and feature selection methods.

TABLE OF CONTENTS

1. INTRODUCTION

Epidemics have been known as threats to humanity for centuries, having large population of numerous neighbourhoods, villages, cities and even whole countries corrupting, taking millions of lives in their paths. "*If anything kills over ten million people in the next few decades, it's most likely to be a highly infectious virus rather than a war."* (Gates, 2015). This matter has been emphasized by scientists as a topic with great need of research and systems investments subjected for the control, reduction of spread cases and more immediate treatment approaches.

"*The failure to prepare could allow the next epidemic to be dramatically more devastating than Ebola*" (Gates, 2015). Through the Ebola epidemic progression over ten thousand people died in only three West African countries. At the beginning, the health workers' ability to find victims and prevent more infections was efficient but ultimately the nature of the virus played a big role in controlling its spread. Ebola is not a disease that spreads through air and its symptoms are harsh and immediate while at the time, only few urban areas had been infected. Those characteristics prevented the virus from spreading to the rest of the world, which could have led to a global pandemic, a potentially huge risk in today's highly connected world. Global health systems and organizations should be more alerted and prepared for possible future pandemics.

COVID-19 stands for Coronavirus Disease firstly identified in 2019, in the city of China - Wuhan and is a disease that had spread across the globe within few months, leading to an ongoing pandemic during the writing of this study. The symptoms of this virus can be mild on some people while others may show no symptoms at all while infectious. Its transmission among people is expected when in close contact[1]. The disease's symptoms can vary, fever, cough, breathing difficulties, and loss of smell and taste are the most common ones. The virus can even become fatal to many, targeting mostly those who suffer from chronic diseases and elderly people (Sauer, 2020).

The community's health became the principal intention of the world, having multiple pharmaceutical companies such as "AbbVie", "Abbott Labs", "Johnson & Johnson", "Pfizer" and more rushing for accurate testing methods, potential treatments and prototype vaccines immunizing against the coronavirus (Speights, 2020). Meanwhile, all governments around the globe had started regulating their communities to reduce the spread in order to prevent the health care systems from

[1] A person is defined as "**contact**" if it encountered a confirmed COVID-19 case within a period that ranges from 48 hours before the onset of symptoms of the case up to 10 days after the onset of case symptoms (NPHO, 2020).

collapsing due to overcrowding. In addition, gaining time for treatment manufacturers to reach a possible therapy. Clearly, such regulations are far from simple for societies as today's people are highly depended on continuous flow of the market to cover their basic needs. These great responsibilities made an immense challenge for governments as they are unfamiliar to such pandemic incidents, examining new measures such as restrictions for unnecessary human mobilization while keeping the market active. It is rational for people to wonder whether these restriction policies or any other measures are effective enough to control the spread or further methods shall be tested.

Machine learning can become a great tool for monitoring, analysing and projecting the effectiveness of such measures, as it focuses on building applications that learn over past collected information and improve their predictive accuracy or decision-making over time. Data science has enormously contributed to a wide range of daily applications such as voice commands, music and product recommendations, various types of robotic intelligence have been beneficial in many sectors including the health care industry. In machine learning, algorithms are 'trained' to find similar patterns and features that are likely correlated in massive amounts of data and further decision making, as well as predictions and projections based on new input. A clear example of such predictive effort is the use of machine learning algorithms and models to give highly accurate results on the immediate diagnosis of Parkinson's Disease, using only the patient's voice sample.

Hopefully, and despite it being such a current global health crisis incident, the COVID-19 pandemic can be a unique opportunity for data collection and analyzation of a pandemic's "behaviour", potentially leading to new insights supporting the community's mental wellness and systems. This study approaches and allows for an investigation to further understand the nature for the ongoing COVID-19 pandemic, making a high-level accurate primary data collection, and finding their correlation to potentially determine the effectiveness of restriction policies against the spread of the virus in Greece. Significant correlation between spread control and the governmental policy regulations may be found through the algorithmic models. They can be examined and tried as intuition for various prediction models estimating upcoming confirmed COVID-19 cases. Such projections and overall identifications can prevent damages in analogous occasions in the future or act as a reference for expected changes. Could such system application be developed and become a tool assisting in finding accurate COVID-19 restriction policies for reducing the disease spread?

2. BACKGROUND

2.1 The COVID-19's nature

Before moving onto machine learning techniques of manipulating data for desired results, it is essential to further investigate and understand the nature of the virus SARS-CoV-2. The COVID-19's virion, is part of the coronavirus family including viruses responsible for (Heymann & Rodier, 2004) and MERS (Pollack, et al., 2013) infections. Scientists suggest it is a zoonotic disease, meaning it jumped from another species to human hosts, and infection between humans is expected to spread primarily through droplets of saliva or discharge from the nose from coughing or sneezing. Everyone's immune system responds differently in an effort to kill off the virus; some ingest and destroy the infected cells while others create antibodies to prevent host cell infection or even make chemicals that are toxic towards infected cells. Similarly, the virus affects different human systems in different ways; most commonly faced symptoms include fever, tiredness and dry cough while aches, pains, sore throat, conjunctivitis, headache, diarrhoea, loss of taste or smell, rash on skin or discolouration of fingers or toes are also likely to occur. In more severe cases the symptoms can become more serious with difficulty when breathing or shortness of breath, chest pains or pressure and loss of speech or movement. The World Health Organization (World Health Organization, 2020) reports that on average, symptoms take up to six days to show up from when firstly infected, however this number can increase up to 14 days for some or let others remain totally asymptomatic.

With this coronavirus being highly contagious, governments, with help from scientists and qualified epidemiologists have restricted mass public gatherings and implemented new policy measures to prevent overwhelming spread. Mainly, to remain protected it is suggested to wear a mask at all times while away from home, wash hands regularly, cough or sneeze in your bent elbow, avoid touching the nose, eyes and mouth, limit social gatherings or crowded destinations and avoid close contact with anyone sick. Frequent disinfection of objects and surfaces is great way to keep surroundings secure.

Understanding the nature of the virus on humans can lead to insights for a better machine learning approach and possibly result to higher accuracies. For instance, in the case of asymptomatic individuals the virus can still spread from them to other people because they are also contagious. Laboratory data suggests that in most common cases infected people appear to be most infectious about 2 days before they develop symptoms and early in their illness (WHO, 2020).

According to The Centre for Evidence-Based Medicine (CEBM), data is unclear on accurately determining the number of asymptomatic cases in ratio to symptomatic ones, while based on 21 analysed reports between 5% and 80% of people testing positive for SARS-CoV-2 may be asymptomatic (Heneghan, et al., 2020) (The Centre for Evidence-Based Medicine CEBM, 2021). Moreover, it is identified by the World Health Organization that some asymptomatic cases will become symptomatic over the next week making them "pre-symptomatic" and confirming the 14-day possibility estimation of showing symptoms.

The lack of information on this topic highly increases the need of wearing a mask at any public place, especially when in contact with another person, sick or not. Another matter to be further justified is that many infected and highly contagious people from the virus are likely to be unaware of carrying the virus, leading to more comfort when meeting others. This gap between the time of infection, the contagious period and the appearance of the first symptoms in any infected person, increase the chances of virion transmission in pre-symptomatic cases and mostly for contagiously asymptomatic. This supports the importance of being self-aware of the infection by regular and sufficient testing and tracing contacts[2] to inform in case of a positive result.

The test used to diagnose COVID-19 is a molecular PCR (Polymerase Chain Reaction) test and works by detecting genetic material from SARS-CoV-2. This genetic material can be found in the nose and upper throat of an infected individual and a sample is taken using a long swab. This

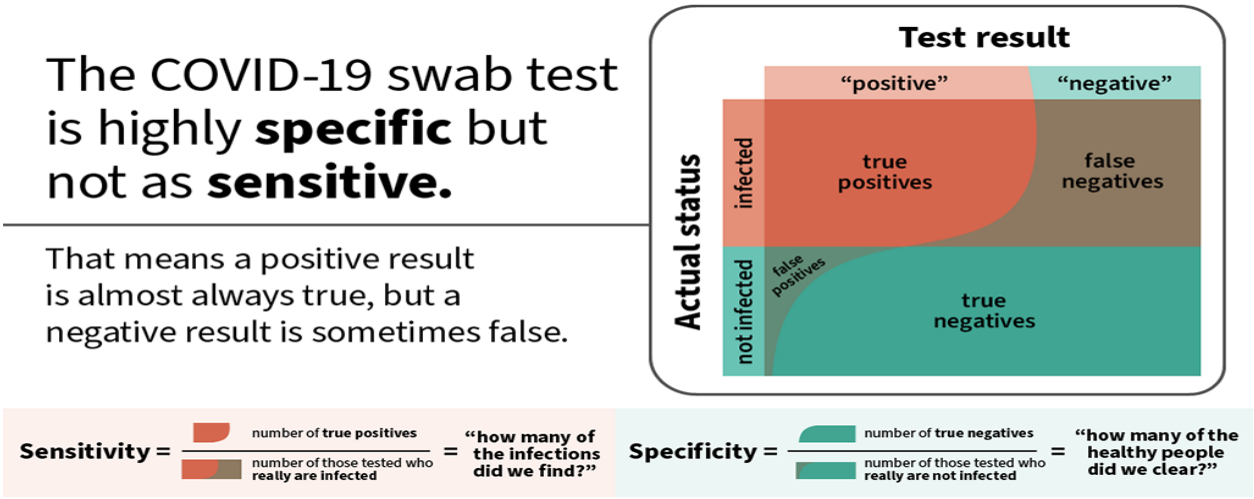

Figure 2.1.1 "COVID-19 Specificity & Sensitivity" (MIT Medical, 2020)

[2] According to healthcare protocols, contacts are categorized as **close contacts** (high risk exposure) or **contacts** (low risk exposure), and tracing is managed according to this categorization and the type of case identification, i.e. suspected, possible or confirmed case (NPHO, 2020).

genetic material cannot be confused with genetic material from other viruses making a false positive very unlikely. On the other hand, the test is not equally sensitive (Figure 2.1) meaning that an imperfect specimen collection, an early stage infection or already partially recovered sample might not contain enough viral material to be detected and turn out positive (MIT Medical, 2020).

Researchers interested in finding the variation of false-negative rate of COVID-19 testing concluded to a reduced probability of 67% for a false-negative result on day 4 of infection compared to the 100% on day 1. On the average day of symptoms (day 5), the median false-negative rate falls to 38% and is decreased to 20% by day 8. The false-negative testing rate begins to increase once again from 21% on day 9 to 66% on day 21 (Lauren M. Kucirka, 2020). This validates the likelihood of a false-positive result, potentially becoming source of false security, leading people to be less careful about social distancing or other safety measures.

There is currently an ongoing process for vaccinating the population at different stages for each country, based on their population, resources, vaccine provider and other factors. COVID-19 vaccines offer immunity two weeks after full vaccination which currently consists of two doses administered (Centers for Disease Control and Prevention, 2020). The first vaccines were available to the public by December 2020 and there already are countries with more than 50% of their population vaccinated but only slightly exceeding 5% in a global scale at the time of writing (April of 2021).

For the minority of people who are not able to recover at home and are being hospitalized or are at risk, a plethora of treatments are already available to curb the progression of COVID-19 in people developing severe illness. Many of the treatments have been authorized by the Food and Drug Administration (FDA) and have been categorized based on the patient's characteristics such as age, weight, chronic diseases etc. (Harvard Health Publishing, 2020).

Science has been working on a continuous effort to find new ways and uncover new data to benefit and support millions of lives. All tests, treatments, vaccines or researches have assisted in this global attempt against the spread of the virus to prevent as many diseased ones from disturbing symptoms and reduce the infected fatality rate to the minimum. Currently, the mortality rate may range from country to country while estimations suggest that only about 1% of infected individuals have fatal risks, a very small number compared to other epidemics, and will only continue to descent as science continues to reveal its secrets (WHO, 2020).

2.2 The role of Machine Learning

A team of researchers in India from various departments including Computer Science and Engineering, approached the growth and trend of the COVID-19 pandemic using machine learning and cloud computing for efficient high-speed computations. Cloud Data Centres that can run and feed continuously the proposed Machine Learning model are used to gain more insight on how the virus impacts the world's population and finally make predictions on the number of COVID-19 cases and the likely dates the pandemic shall end in various countries (Shreshth Tuli, 2020).

Through recent research by Data-Driven Innovation Laboratory, Singapore University of Technology and Design (SUTD, 2020), to estimate the number of cases over time, Gaussian distribution method was deployed, and the regression curves were drawn using the Susceptible-Infected-Recovered model. Many sources note that data corresponding to new cases over time have a large number of outliers and may or may not follow a standard distribution such as Gaussian or Exponential. In the mentioned studies, reporting for SATA-CoV-1, an earlier version of the virus, the data fits better on a Generalized Inverse Weibull (GIW) Distribution than a Gaussian distribution (Shreshth Tuli, 2020).

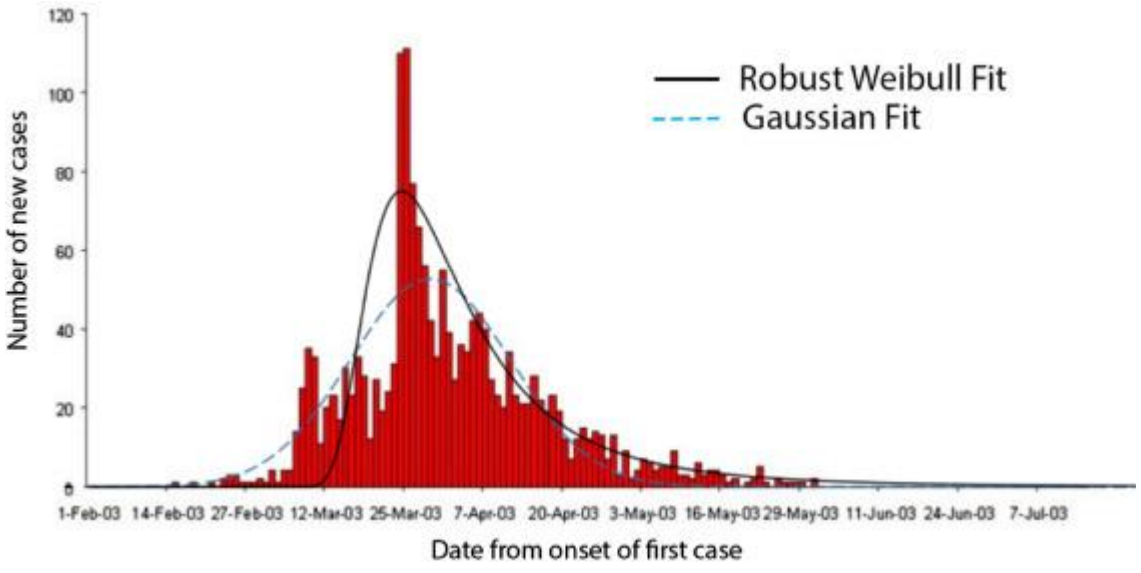

Figure 2.2 "Robust Weibull vs Gaussian" (Shreshth Tuli, 2020)

After running the algorithms using the dataset reported daily from the World Health Organization through the analysed and interpreted models for countries sustaining enough data, various predictions were made. Such data can be critical to prepare the healthcare services or required restriction applications in advance.

Specifically, the predictions of the baseline Gaussian model deployed by SUTD are concluded as overoptimistic which could potentially lead to premature uplifting of restrictions, causing adverse effect on the control and management of the epidemic. Furthermore, outbreaks of the corresponding diseases have received diverse responses from different countries. China, India, Australia and other similar countries have introduced partial or full nation-wide lockdowns while countries such as Sweden have introduced slight to no restrictions at all. These policies definitely affect the distribution of cases within an area and impact the curve parameters. The research suggests, having holistic models that can take indicators such as travel history of citizens and migrants and the restrictions specifically for each country shall lead to more accurate results as they incorporate the biases of the parameters. "*Such models can be explored in future.*" (Shreshth Tuli, 2020).

Another team of researchers and engineers from Greece from various departments of computer science, mainly focusing on data analysis and scientific modelling, decided to create an open access tool to better inform people about SARS-CoV-2 through real-time data and projections for various countries. Their goal is to provide insights and projections about the virus' progression in a fast, clean and simple design. With their tool publicly available online, they allow anyone to comprehend the huge amount of data and further favour the community's behaviour being one of the major factors for controlling the pandemic (Alkaios Sakellaris, 2020).

For the project's data visualization, multiple sources have been used and are classified in the two main categories of live data and time series data. Live data have been retrieved from Worldometer (Worldometer, 2021), BNO (BNO News, 2021) and Johns Hopkins University (Johns Hopkins University, 2021) and time series data from WHO (WHO, 2020), CDC (Centers for Disease Control and Prevention, 2020), ECDC (European Centre for Disease Prevention and Control, 2020) and John Hopkins University (Johns Hopkins University, 2021). Lastly, a governments' policy responses to the coronavirus, per country, dataset was used for analysis collected from University of Oxford (University of Oxford, 2021).

Up to this point, the team has managed to create projections for over ten countries including Greece, United Kingdom, Italy, parts of the United States and more, based on System Dynamics modelling making visualized simulations. It is also noted that there are more upcoming projections for various countries, but such simulations are not accurate forecasts neither future predictions.

System Dynamics (SD) is a method and mathematical modelling technique for studying the world around us, framing complex issues and problems. The COVID-19 pandemic being a high complexity problem makes SD a fitting tool for modelling its progression while enabling other non-disease-related features such as human behaviour, restriction policies and economy to integrate its methodology (System Dynamics Society, 2021).

The core of their system mainly uses the Susceptible-Exposed-Infected-Recovered (SEIR) epidemiology model with slight changes in its basic structure, adding two new stocks “Infected Symptomatic” and “Infected Sick” replacing the stock “Infected”. Furthermore, it is assumed that part of asymptomatic cases exist within the stock “Exposed” as depicted in Figure 2.3. This replacement process is done to better illustrate the behaviour of SARS-CoV-2 based on its nature.

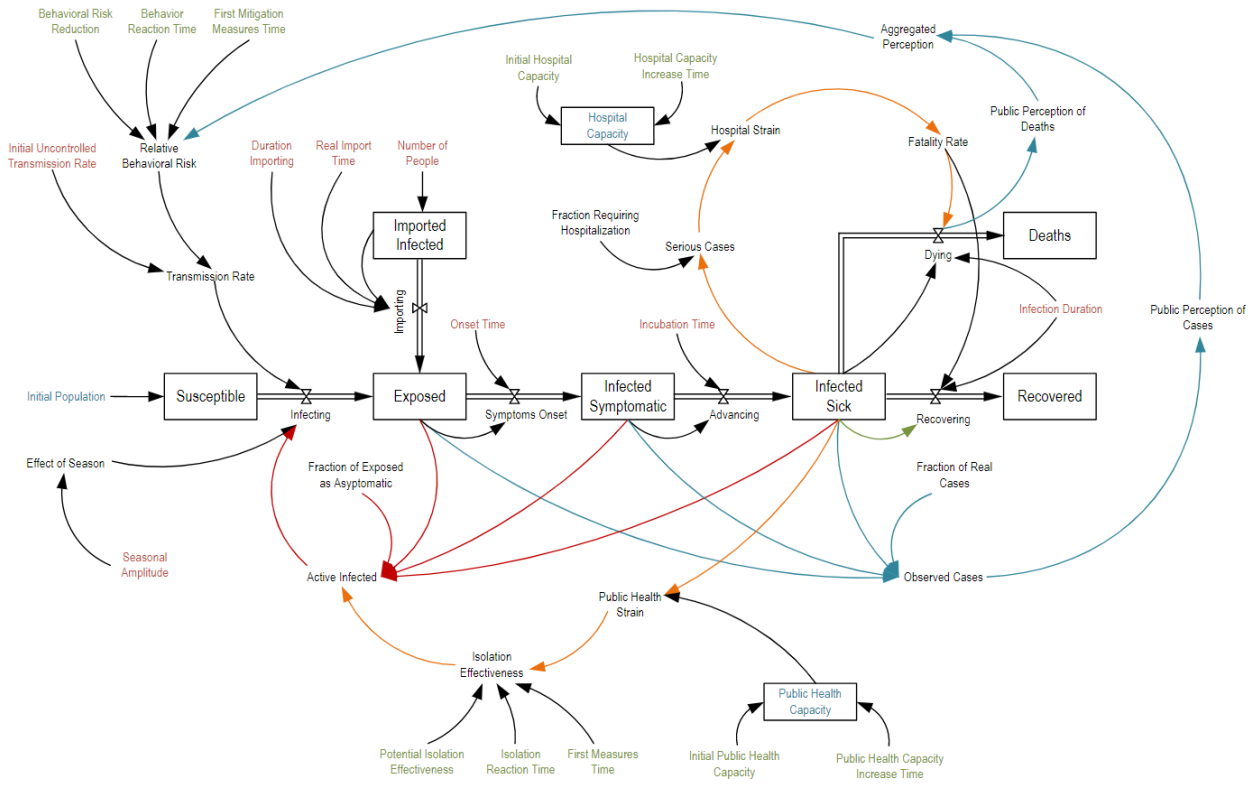

Figure 2.3 "Susceptible-Exposed-Infected-Recovered epidemiology model" (Alkaios Sakellaris, 2020)

The structure of this system allows interaction between changing variables including; red loop indicating the number of past infected people potentially infecting others, orange loop illustrates the public isolation mechanism established by the Public Health System and the hospital

capacity which if exceeded, fatality rate is increased. Lastly, blue loop changes the outcomes of imposed mitigation measures assuming that government's measures change the relative behaviour of the population after a specific time. Moreover, the variables controlling this simulation are also coloured, with red showing variables characterizing COVID-19 or are driven by its behaviour, blue shows country-depended variables and green shows variables controlling the transmission of the virus.

Overall, this model seems to aim towards a strong understanding of the virus' natural behaviour and the impact of population's behaviour towards the positivity rate. Through observation, it is identified how outsourced factors can greatly impact the system and drive to different results, possibly making more accurate runs when including such "out of the box" features. Certainly though, every model is wrong, and no one can model reality. There are still numerous variables that are not part of the system but do play important role in the community. For example, this model cannot recognize whether recovered people may now have immunity to the virus and what their ratio to the current population is. Additionally, effects and pressure from downgrade economies or new drugs lowering fatality rate are also excluded from this model's perspective, while rates of the disease are not based on age cohort. All outcomes from any model are based on its assumptions of the model structure given the quality and quantity of the data used to derive to assumptions.

More specifically, the analysis for COVID-19 progression in Greece includes data from positivity rate, number of daily tests and Mobility and Activity Index (MAI) provided from Google COVID-19 Community Mobility Reports (Google, 2021). This dataset provides information based on mobility trends for Grocery & pharmacy, parks, transit stations, retail and recreation, residential and workplaces, giving valuable insights towards the mobility and economic activity of a country.

With a glance, a huge drop of 75,4% in Mobility and Activity Index is seen (Figure 2.4) after the first lockdown Greece enforced in March 2020 while its correlation towards the government's restriction measures becomes more distinct. Through this graph the rise in mobilization is also identified once specific measures are lifted within the fifth month of the year and lastly a return on the baseline first hits in July when Greece gains back all its economic activity.

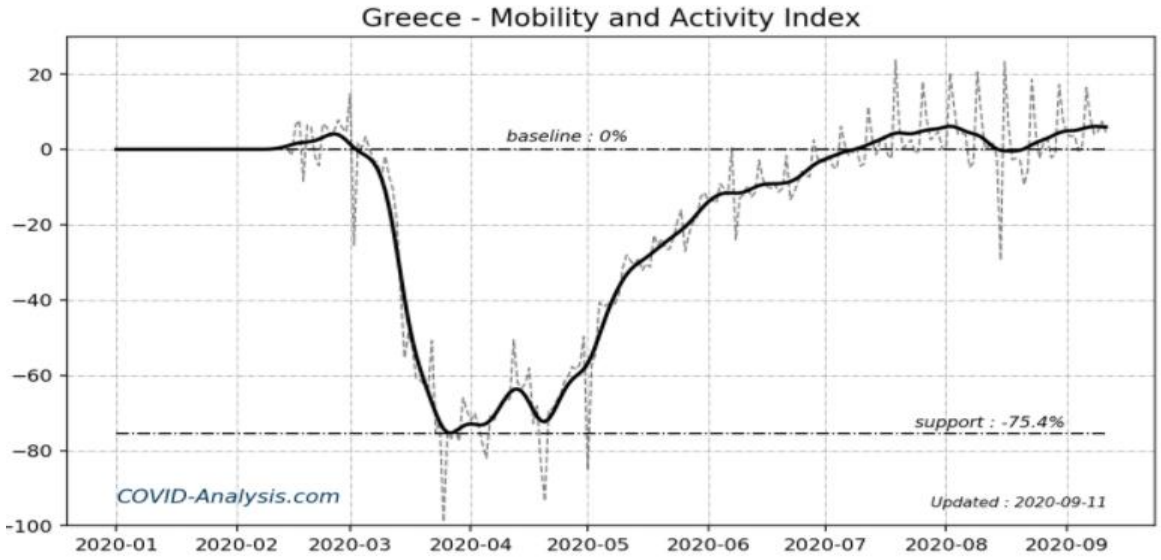

Figure 2.4 "Mobility & Activity Index" (Alkaios Sakellaris, 2020)

After further analyzation on data such as the number of daily tests and the positivity rate the team indicates that the epidemic in late 2020 for Greece is under control. Explicitly, for positivity rate it is suggested that any sharp increase exceeding 5% should conclude to new measures while capacity of testing should also increase.

Lastly, the mean absolute percentage error and median absolute percentage error (MAPE) on the cumulative cases and deaths (MdAPE), were used as evaluation metrics respectively allowing comparability of the model's performance with other COVID-19 forecasting models. The evaluation results show an overall error of less than 10% for cumulative cases models and 5% for cumulative deaths models, comparably little assuming the reported data show the reality of COVID-19 spread, which is very unlikely.

Also relying on the susceptible-infected-recovered (SIR) model, a journal published in Turkey further approached mobility datasets from Google and Apple and studied the role of social distancing policies in 26 countries, including Turkey, the United States, Belgium and more to analyse the transmission rate of the coronavirus disease over a timeline of 5 weeks using machine learning. Gradient boosted trees regression analysis showed that changes in mobility patterns resulting from social distancing policies explain approximately 47% of the variation in the disease transmission rates. Consistent with simulation-based studies, real cross-national transactional data confirms the effectiveness of social distancing interventions in slowing the spread of COVID-19 (Dursun Delen, 2020).

In most cases, the problem of the coronavirus disease is approached by time series models to better represent its functionality to the algorithm in terms of virus reproduction (r) over time. In this case, the reproductive ratio of the virus determines the average number of infected contacts per infected individual (Domingo, 2020). Such was the approach in a study aiming to model the spread of COVID-19 infection using a Multilayer Perceptron (MLP) with the use of infected, recovered and deceased patients' data. The dataset was transformed into a regression dataset and was used in a MLP artificial neural network. After cross-validation of the data the researchers showed high robustness of the deceased and confirmed patients' models while scoring lower on the recovered (Zlatan Car, 2020).

Through a comparative study analysis and forecasting of COVID-19 with use of data from Belgium, Denmark, France, Finland, United Kingdom, Germany, Turkey and Switzerland a team of researchers, aimed to find the most effective time series model to approach the problem. Auto-Regressive Integrated Moving Average (ARIMA), Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM) were all the tested methods while six model performance metrics were used to determine the most accurate of the models (MSE, PSNR, RMSE, NRMSE, MAPE and SMAPE). Due to their results, it can be said that LSTM is a great success in predictions of Finland, Switzerland and Germany and is clearly stated as the most successful prediction model overall (İsmail Kırbaş, n.d.).

Another research supporting this statement has been conducted in India in which LSTM data-driven estimations and curve fittings have been used to predict the number of COVID-19 cases in India 30 days ahead and further calculate the effect of preventive measures like lockdowns and social isolation on the spread of SARS-CoV-2. The researchers considered it a success as the value of r before lockdown was at 2.3 (i.e. an infected individual can infect 2.3 persons) and was reduced to 0.15 after the lockdown. The virion's reproduction rate is observed to be very much impacted by the preventive measures which worked well in containing this contagious virus in India (Anuradha Tomar, 2020).

The LSTM model does not only seem to handle directly COVID-19 data but also inputs such as mobility and activity indexes, restriction policy methods and not only. Results retrieved from a statistical exploration and univariate timeseries analysis on COVID-19 to understand the trend of disease, spreading and death, show performances of bidirectional LSTM models

outperforming multilayer LSTM models. In the study analysis environmental features were included using parameters such as the external temperature, sunshine, rainfall etc seeking for correlation within them (Ayan Chatterjee, 2020).

This recurrent neural network class seems to perform very well on learning long term dependencies in data such as for the COVID-19 pandemic and yet for investigation in correlation to the preventive measures of a country. LSTM is a great tool to model the scenario in Greece and investigate likely correlation between the COVID-19 cases and the restriction policies by concluding to more accurate predictions.

3. METHODOLOGY

While investigating for a design to approach the problem of COVID-19 pandemic in Greece, grant insights defining the effectiveness and correlation between Greek policies and the spread of SARS-CoV-2, many key factors guiding the method to follow have been identified. Through secondary research, the nature of the virus seems mysterious enough to assure us that any available data cannot be accurate enough, enhancing the statement of how simulating a complete reality can never be succeeded. On that note, machine learning is still a great tool to manipulate such huge amounts of data and still sustain high accuracy results despite its imperfections. Research also shows other works been implemented in different parts of the world that are considerably applicable giving desired results.

As this project focuses on the Greek territory, the data to be collected concerns the period of COVID-19 in Greece, where the virus was firstly identified on the 26th of February 2020. Multiple sources publicly available to choose from are there, while data concerning the COVID-19 is only a segment for the project's objectives. Moreover, an accurate dataset for the social distancing policies and restriction methods taken by the government in Greece is to be linked to the main dataset before any further handling by the algorithms. Additionally, the MAI provided by Google implies as a good indicator for estimating the mass mobilization within an area over time, making it a must-have dataset to include.

The designed dataset (Table 4.2.1) shall consist from a variety of attributes that could affect the matter of coronavirus in Greece, and by passing them through various models we grant a synchronized access to all information retrieving target results. The preparation or pre-processing stage for the dataset is necessary for it to compile through algorithms and give decisive outcomes.

Checks for imbalances, duplications or biases that could impact the learning process and mislead the results are altered.

Since the algorithms will try to learn from the data and give predictive results, it is essential to identify and settle the target attribute. In this case, the algorithms will be learning through the given data trying to determine the output of the upcoming reproduction rate number. Reproduction rate determines the average number of new infections per infected individual based on the given data.

Having already met epidemiologically well-fitting machine learning models on previous researches such as the LSTM time series model, it can be used as gesture of approach towards the finest machine learning model to use and retrieve accuracy results from. It might seem logical and coherent for a time series model to best represent such natural data in the real world, but the method of data collection and their structure are to determine the best option. To assure that indeed the LSTM model is the most fitting algorithm to use, multiple phases firstly using other basic machine learning algorithms including support-vector machines, random forests and simple neural networks will initially take place. Getting some first insights concerning the data, its structural needs and the responses of the models, will definitely support towards concluding to a well-fitting algorithm. Such insights could also come in handy in an alternative to another Recursive Neural Network model similar to LSTM, in case its output does not meet satisfactory levels.

4. DATA DESCRIPTION

4.1 Collection

For the models, the more the data to consider, the wider the range of possibilities. To back this statement up, features from three distinct datasets have been selected, merging them into a single data structure. The complete COVID-19 dataset from Our World in Data (OWD, 2021), the Google MAI dataset (Google, 2021), the Oxford COVID-19 Government Response Tracker (Univesity of Oxford, 2021) and few additional time measuring features were used for this dataset to be complete. Only data concerning the Greek matter were extracted in the time period between the 26th of February and the 10th of April, considerably a good enough period with over a year of data. Every feature in a dataset can carry valuable information to impact the algorithms' results, and for them to be effective proper formation in the pre-processing stage is required.

4.2 Features

Table 4.2.1 Dataset attributes

#	Attributes	Definition
1	date	Date & time
2	New cases	New cases per day
3	Total cases	The sum number of all new cases up to date
4	New deaths	New deaths per day
5	Total deaths	The sum number of all new deaths up to date
6	New cases / million	New cases per day per million
7	Total cases / million	The sum number of all new cases up to date per million
8	New deaths / million	New deaths per day per million
9	Total deaths / million	The sum number of all new deaths up to date per million
10	Reproduction rate	The average number of infected contacts per infected individual
11	New tests	New tests per day
12	Total tests	The sum number of all new tests up to date
13	New tests / thousand	New tests per day per thousand
14	Total tests / thousand	The sum number of all new tests up to date per thousand
15	Tests per case	The average number of tests per new case
16	Positivity rate	The rate of positive outcomes from new tests
17	People vaccinated	Number of people having taken the first vaccine dose
18	People fully vaccinated	Number of people having both vaccine doses taken
19	New vaccinations	Number of new vaccinations per day

20	Total vaccinations	The sum number of all new vaccinations up to date
21	% Total vaccinations	The percentage of total vaccinations to the country's population
22	% People vaccinated	The percentage of first dose vaccinations to the country's population
23	% People fully vaccinated	The percentage of second dose vaccinations to the country's population
24	stringency_index	Number representing the stringency index of a country (0/100)
25	retail_and_recreation_percent _change_from_baseline	Google Mobility and Activity Index for retail and recreation areas
26	grocery_and_pharmacy_percent _change_from_baseline	Google Mobility and Activity Index for grocery and pharmaceutical areas
27	parks_percent_change_from _baseline	Google Mobility and Activity Index for parks
28	transit_stations_percent_change _from_baseline	Google Mobility and Activity Index for transit stations
29	workplaces_percent_change _from_baseline	Google Mobility and Activity Index for workplace areas
30	residential_percent_change _from_baseline	Google Mobility and Activity Index for residential areas
31	C1_School closing	School closure policies
32	C1_Duration	School closure policy duration
33	C2_Workplace closing	Workplace closing policies
34	C2_Duration	Workplace closing policy duration
35	C3_Cancel public events	Cancelation of public events policies
36	C3_Duration	Cancelation of public events policy duration
37	C4_Restriction on gatherings	Restriction on gatherings policies
38	C4_Duration	Restriction on gatherings policy duration
39	C5_Close public transport	Closure of public transportation policies
40	C5_Duration	Closure of public transportation policy duration
41	C6_Stay at home requirements	Stay at home requirement policies
42	C6_Duration	Stay at home requirement policy duration
43	C7_Restriction on internal movement	Internal movement policies
44	C7_Duration	Internal movement policy duration
45	C8_International travel controls	International travel control policies
46	C8_Duration	International travel control policy duration
47	H1_Public information campaigns	Public information campaigns
48	H1_Duration	Public information campaign duration

49	H2_Testing policy	Testing policies
50	H2_Duration	Testing policy duration
51	H3_Contact tracing	Contact tracing policies
52	H3_Duration	Contact tracing policy duration
53	H6_Facial Coverings	Facial Covering policies
54	H6_Duration	Facial Coverings policy duration
55	H7_Vaccination policy	Vaccination policies
56	H7_Duration	Vaccination policy duration
57	H8_Protection of elderly people	Protection of elderly people policies
58	H8_Duration	Protection of elderly people policy duration
59	GovernmentResponseIndex	Number representing the Government's responses to the pandemic
60	ContainmentHealthIndex	Number representing the public containment health ratio

4.3 Data justification

A total of 60 attributes including information for over a year span, covering variables that immediately matter the coronavirus pandemic in Greece, mobility indexes for respective areas and restrictive implemented actions for reducing death and reproduction rate taken by the administration. All features contain information in numerical form simplifying the machine learning pre-processing procedure and giving more efficient results.

Some major changes have been implemented to the final dataset, mostly within Government Response Tracker attributes based on the algorithmic needs, to enhance the approach towards the project's objectives and reduce the likelihood of misinformation by removing unnecessary parts. Features concerning the economical movements of the government are dropped from the table and columns representing the duration of each implemented response are added for tracking.

For the government response attributes a specified coding method is seen to justify changes within topic and differentiate the nature of the fields. "C" code stands for containment and closure policies and "H" is for health system policies. Within the time period of the pandemic, changes within policies are seen, making at times a more restrictive or loosening approach based on the circumstances. These changes are represented by the code values within each government response attribute listed in Table 4.3.1.

Table 4.3.1 COVID-19 Government Response Tracker dataset codebook (University of Oxford, 2021)

available at: [covid-policy-tracker/codebook. covid-policy-tracker · GitHub]

ID	Name	Description	Coding
C1	C1_School closing	Record closings of schools and universities	0 - no measures 1 - recommend closing or all schools open with alterations resulting in significant differences compared to non-Covid-19 operations 2 - require closing (only some levels or categories, eg just high school, or just public schools) 3 - require closing all levels Blank - no data
C2	C2_Workplace closing	Record closings of workplaces	0 - no measures 1 - recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operation 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors) Blank - no data
C3	C3_Cancel public events	Record cancelling public events	0 - no measures 1 - recommend cancelling 2 - require cancelling Blank - no data
C4	C4_Restriction on gatherings	Record limits on gatherings	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data
C5	C5_Close public transport	Record closing of public transport	0 - no measures 1 - recommend closing (or significantly reduce volume/route/means of transport available) 2 - require closing (or prohibit most citizens from using it)

			Blank - no data
C6	C6_Stay at home requirements	Record orders to "shelter-in-place" and otherwise confine to the home	0 - no measures 1 - recommend not leaving house 2 - require not leaving house with exceptions for daily exercise, grocery shopping, and 'essential' trips 3 - require not leaving house with minimal exceptions (eg allowed to leave once a week, or only one person can leave at a time, etc) Blank - no data
C7	C7_Restriction on internal movement	Record restrictions on internal movement between cities/regions	0 - no measures 1 - recommend not to travel between regions/cities 2 - internal movement restrictions in place Blank - no data
C8	C8_International travel controls	Record restrictions on international travel Note: this records policy for foreign travellers, not citizens	0 - no restrictions 1 - screening arrivals 2 - quarantine arrivals from some or all regions 3 - ban arrivals from some regions 4 - ban on all regions or total border closure Blank - no data
H1	H1_Public information campaigns	Record presence of public info campaigns	0 - no Covid-19 public information campaign 1 - public officials urging caution about Covid-19 2- coordinated public information campaign (eg across traditional and social media) Blank - no data
H2	H2_Testing policy	Record government policy on who has access to testing Note: this records policies about testing for current infection (PCR tests) not testing for immunity (antibody test)	0 - no testing policy 1 - only those who both (a) have symptoms AND (b) meet specific criteria (eg key workers, admitted to hospital, came into contact with a known case, returned from overseas) 2 - testing of anyone showing Covid-19 symptoms 3 - open public testing (eg "drive through" testing available to asymptomatic people) Blank - no data
H3	H3_Contact tracing	Record government policy on	0 - no contact tracing

		contact tracing after a positive diagnosis Note: we are looking for policies that would identify all people potentially exposed to Covid-19; voluntary bluetooth apps are unlikely to achieve this	1 - limited contact tracing; not done for all cases 2 - comprehensive contact tracing; done for all identified cases
H6	H6_Facial Coverings	Record policies on the use of facial coverings outside the home	0 - No policy 1 - Recommended 2 - Required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing not possible 3 - Required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible 4 - Required outside the home at all times regardless of location or presence of other people
H7	H7_Vaccination policy	Record policies for vaccine delivery for different groups	0 - No availability 1 - Availability for ONE of following: key workers/ clinically vulnerable groups (non elderly) / elderly groups 2 - Availability for TWO of following: key workers/ clinically vulnerable groups (non elderly) / elderly groups 3 - Availability for ALL of following: key workers/ clinically vulnerable groups (non elderly) / elderly groups 4 - Availability for all three plus partial additional availability (select broad groups/ages) 5 - Universal availability
H8	H8_Protection of elderly people	Record policies for protecting elderly people (as defined locally) in Long Term Care Facilities and/or the community and home setting	0 - no measures 1 - Recommended isolation, hygiene, and visitor restriction measures in LTCFs and/or elderly people to stay at home 2 - Narrow restrictions for isolation, hygiene in LTCFs, some limitations on external visitors and/or restrictions protecting elderly people at home

			3 - Extensive restrictions for isolation and hygiene in LTCFs, all non-essential external visitors prohibited, and/or all elderly people required to stay at home and not leave the home with minimal exceptions, and receive no external visitors Blank - no data

5. DESIGN, TESTING, IMPLEMENTATION

5.1 Pre-processing

The final dataset is to be designed from all three retrieved datasets and merged into a single one using "pandas" for python, the pre-processing stage is then to be implemented. Firstly, missing values are identified in order to be adjusted or removed. For the COVID-19 dataset's attributes, all missing values prior to each first true value for every attribute have been set as 0 based on the attribute's nature (ex. People fully vaccinated, values before first full vaccination are set to 0). For the Google MAI dataset, no missing values were found, and the data was merged to the newly designed dataset aligned by date. Next, the COVID-19 Government Response Tracker contained duplicate attributes with similar information such as the stringency index which was removed. Moreover, parameters used as "flags" for geographic scope were replaced by "duration" periods to measure the scale of effectiveness over time and approach one of the main objectives finding the most optimal time duration of each restriction. Lastly, features concerning the economical movements of the government are dropped from the table and the dataset is also attached to the main design. Since each row of data represents a single day in data, all other rows containing even a single missing value were completely removed. The table now consisting of 345 total rows and 60 columns is ready for scaling and further optimization.

To establish the output column, the reproduction rate attribute is copied and shifted one cell up, indicating the next day's reproduction rate number labelled as rr+1. This way, we are implementing a predictive method for the algorithm to train and give results targeting the following day. Based on research, we have identified that the fifth day is the average day of symptoms, when also the median false-negative test result falls to 38% chance compared to 100% probability of false-negative results on day one of infection. Thus, to make accurate results based on the virus' nature and its connection to the confirmed cases, another reproduction rate output is added and

labelled as rr+5, consisting of the reproductive number found five days later. This creates another predictive path for the algorithm to train and give results targeting the fifth following day.

5.2 Random Forests' Regression

For the first iterations, the random forests regression algorithm will be used as a generally used technique to predict values, in this scenario to retrieve reproduction rate predictive results. Using this ensemble learning technique, taking multiple algorithms multiple times to make predictions based on the decision trees, we are hoping to find insights concerning the specified dataset.

For this model, scaling the data is unnecessary and thus the dataset is split into training and testing segments right before fitting it in. The model also consists of an imbedded method for choosing the best features towards the best result, making the feature selection stage unnecessary. Instead of an algorithmic feature selection stage based on correlation, the implementation of the model will run in two base cycles for each prediction, consisting of manually selected attributes and excluding the features related to government policies to begin. Lastly, to get accurate model results, hyperparameter tuning is valuable and can make significant changes. Accuracy results will be taken from an average of all iterations for each cycle while this process will be repeated for predicting both rr+1 and rr+5, the first and fifth day's prediction.

Table 5.2.1 Random Forests Regression Prediction Results

RFR	NO POLICIES	WITH POLICIES
Day 1	98,38%	98,55%
Day 5	96,69%	97,31%

While tuning the hyperparameters, a change in prediction accuracy is met once the "maximum-features" attribute is about half of the actual's dataset size. This suggests that some of the attributes are not correlated or unusable for this model. Parameters "n-estimators" and "max-depth" make no significant difference. Moreover, it has become observable that prediction values when including the restriction policy attributes are slightly higher, suggesting for possible correlation in restriction policies towards the outcome of reproduction rate.

5.3 Support Vector Regression and Grid Search

For the next model, the approach will differ based on its nature and needs for accurate results. Support Vector Regression (SVR) is a supervised learning technique, based on the concept of support vectors. It aims at reducing the error and minimising the range between the predicted and observed values.

For the SVR model, the problem is approached from a different perspective. Due to its vast hyperparameter options, the so-called Grid Search method will be used, basically making a loop to efficiently test numerous possible tunings at one run. The most important hyperparameter for the SVR model is the "kernel" option used by the algorithm through each iteration, indicating the method the data is to be processed by it, having to choose between the "linear", "sigmoid", "poly" or "rbf" kernel. For this model, the data is intentionally rescaled between 0 and 1 to improve the mathematical capacities of the machine. The Grid Search method gives the ability to then print the best hyperparameter options of each model based on its R squared coefficient (R^2) accuracy score. The coefficient R2 is defined as (1−u/v), where u is the residual sum of squares and v is the total sum of squares. The most optimum possible score is 1.0 while results can be negative.

Table 5.3.1 SVR Grid Search best parameter / kernel best result (day 1 prediction)

GRID SEARCH ALL FEATURES OPTIMAL SVR (day 1)				
Kernel / Parameter	poly	linear	sigmoid	rbf
C	1	0.1	1	1000
degree	1	1	1	1
gamma	0.1	1	0.01	0.0001
Best R2	0.9223085324381037	0.922409481949082	0.9229933947889384	0.9222760718137747

Table 5.3.2 SVR Grid Search best parameter / kernel best result (day 5 prediction)

GRID SEARCH ALL FEATURES OPTIMAL SVR (day 5)				
Kernel / Parameter	poly	linear	sigmoid	rbf
C	1000	0.1	1000	100
degree	1	1	1	1
gamma	0.0001	1	0.0001	0.001
Best R2	0.8816548557145835	0.8814451820314761	0.8814634852708154	0.8928804333498668

While testing, ‘sigmoid’ kernel gives the most efficient R^2 coefficient for the single day prediction while ‘rbf’ kernel is best for the fifth’s day prediction. Overall, all kernels retrieve lower results for the fifth’s day prediction as expected, indicating that direct correlation of the data towards day 5 is likely reduced.

Differing from the Random Forests model, the SVR model does not have a way to choose the best features for the best results, instead it uses complex diagrams to indicate results based on all of its given data. Thus, a similar run for both prediction days is undertaken after a solid feature selection process.

For feature selection, a correlation heatmap (Figure 5.3.1) is calculated using ‘Pearson Correlation’ seeking for correlation between all attributes and more specifically towards each outcome. For day 1 prediction, the features; ‘New deaths / million’, ‘Reproduction rate’, ‘New vaccinations’, ‘% Total vaccinations’, ‘% People vaccinated’ and ‘stringency-index’ were found having the strongest correlation of all. Similarly, for day five prediction, ‘New deaths / million’, ‘Reproduction rate’, ‘% People fully vaccinated’, ‘New vaccinations’ and ‘stringency-index’ have the strongest correlation.

Another insight by the feature selection mechanism towards the correlation of restriction policies and the reproduction rate, is the strong correlation of stringency index indicating the overall regulatory stringency number, and the effect of vaccination attributes aiming for better accuracy results.

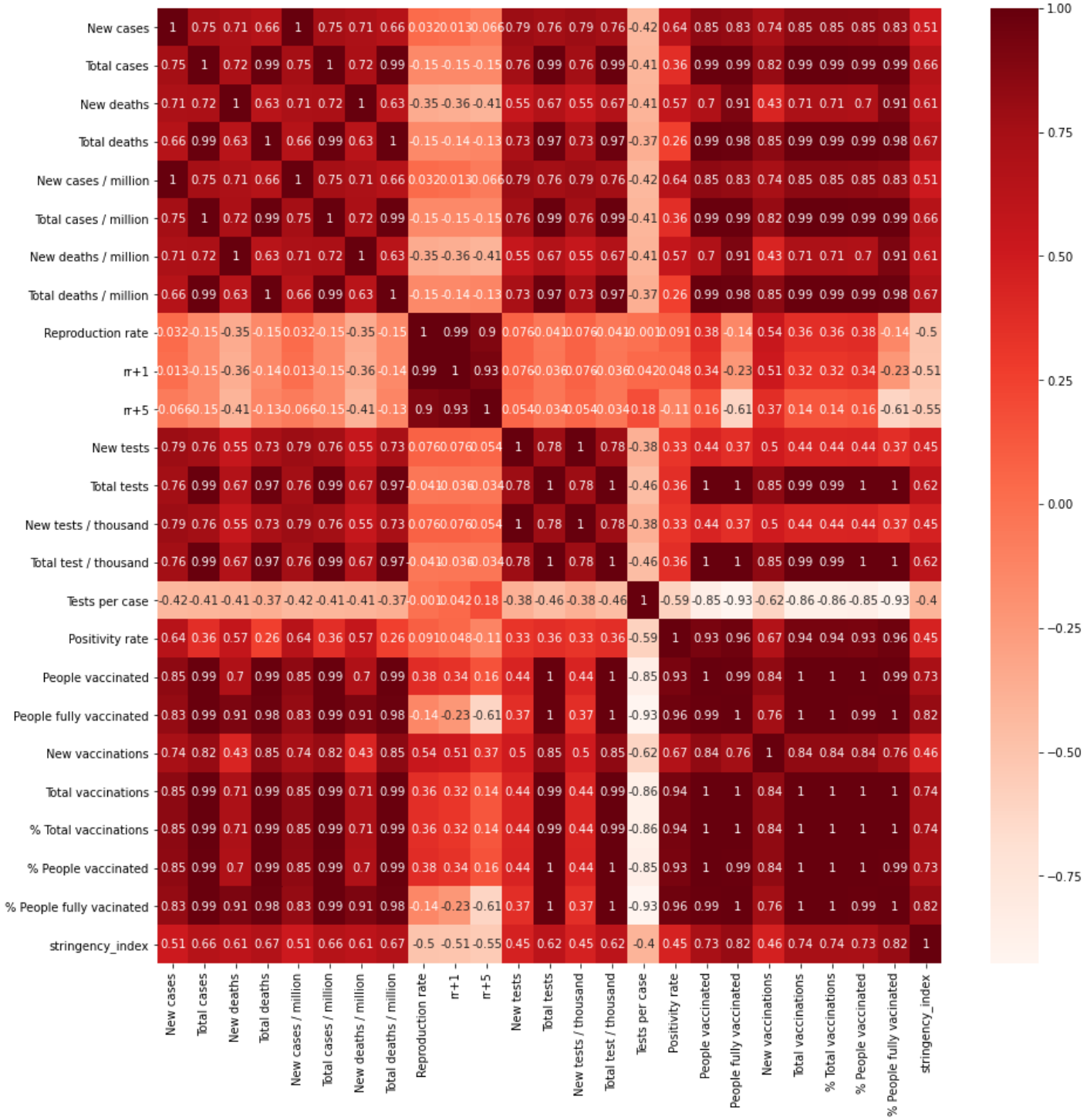

Figure 5.3.1 Correlation Heatmap

Once again, having selected only the correlated features, we find the best parameters for SVR towards the best R^2 score (Table 5.3.3, Table 5.3.4) for both day 1 and day 5 predictions using the Grid Search method. After finding the best parameters for both day 1 and day 5 predictions, we observe the given below outcomes.

Table 5.3.3 Selected Features & SVR Grid Search best parameter / kernel best result (day 1 prediction)

GRID SEARCH FEATURES SELECTED OPTIMAL SVR (day 1)				
Kernel / Parameter	poly	linear	sigmoid	rbf
C	0.1	0.1	1	100
degree	1	1	1	1
gamma	1	1	0.1	0.001
Best R2	0.9491448375726022	0.9491448375726022	0.958385734526878	0.9506865686920833

Table 5.3.4 Selected Features & SVR Grid Search best parameter / kernel best result (day 5 prediction)

GRID SEARCH FEATURES SELECTED OPTIMAL SVR (day 5)				
Kernel / Parameter	poly	linear	sigmoid	rbf
C	0.1	1	1000	100
degree	1	1	1	1
gamma	1	1	0.001	0.1
Best R2	0.8000783344716345	0.7974586419462579	0.798084423346126	0.8445727011937022

Even though results vary, a pattern is identified based on the kernel choice taken by the Grid search algorithm. Again, for the first day's prediction we see the 'sigmoid' kernel holding the best R^2 score and 'rbf' kernel for day 5 respectively. Another possible indication for an accurate feature selection stage, the strong correlation of the selected features succeeded a better result for day 1 reproduction rate prediction, while a lower day five prediction suggests that values are now likely less corelated.

5.4 Long-Short Term Memory

The LSTM model is the most popular model in the time series domain, it was also found to be practical by numerous other scientists using it for epidemiological concerns and similar other projects suggested by research.

Since the Recurrent Neural Networks operate differently as well, the approach will differ for this run. Feature selection is unusual for these models, as they find the best features to work with autonomously. Thus, similarly to the Random Forests Regression process, the data will be split dividing the policy tracker attributes from the rest of the dataset which still includes the stringency index. Furthermore, the Grid Search method used for Support Vector Regression model will be used in a similar but more simplistic way for finding the optimal hyperparameter. The results will be given by Mean Absolute Error (MAE), Mean Squared Error and the R^2 coefficient. In some cases, the R^2 coefficient has high results just like other algorithms, but the mean squared and absolute error might also be high. Such results may be indicating high accuracy for R^2 but they are misleading since the mean error is so over the optimal. Key to finding a fitting model is low mean errors and good accuracy.

LONG-SHORT TERM MEMORY			
NO POLICIES IN DATA			
TUNING	MSE	R_SQ	MAE
1	0,648615	0,686318	-0,856206
10	0,930791	0,790952	-1,66374
100	1,32531	0,858417	-2,79277
200	3,62311	1,34958	-9,36859
300	1,23302	0,784428	-2,52867
500	1,54199	0,885535	-3,41288
700	1,09395	0,830577	-2,13068
1000	0,183167	0,333823	0,475813
AVERAGE	1,322494125	0,81495375	-2,78471538

Table 5.4.1 LSTM model results, no policy dataset

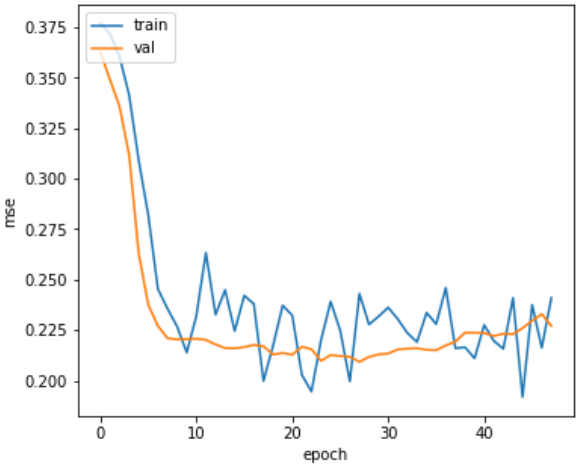

Figure 5.4.1 LSTM model results, no policy dataset

As shown in Table 5.4.1 and Figure 5.4.2, the results from the LSTM model are not the best despite it being the most popular for such type of information. The singular tuning is giving low MSE, MAE and R^2 but that would suggest using only one layer which is not practical for LSTM. Through the graph, we can identify points where actual values intercept with the predicted ones. This visualization helps to understand the algorithmic intensions, showing a relatively good prediction overall, but based on the numbers previously met models had better success. But in the overall process, the mean error is too high to even consider the accuracy result.

LONG-SHORT TERM MEMORY			
WITH POLICIES IN DATA			
TUNING	MSE	R_SQ	MAE
1	0,31533	0,46777	0,11685
10	1,59683	1,06141	-3,56981
100	1,17208	0,865926	-2,35425
200	1,92477	0,97791	-4,56529
300	1,69097	0,902876	-3,83922
500	0,747206	0,684156	-1,13835
700	1,00864	0,797982	-1,88652
1000	0,316355	0,46353	0,0946554
AVERAGE	1,096522625	0,777695	-2,142741825

Table 5.4.1 LSTM model results, dataset with policies

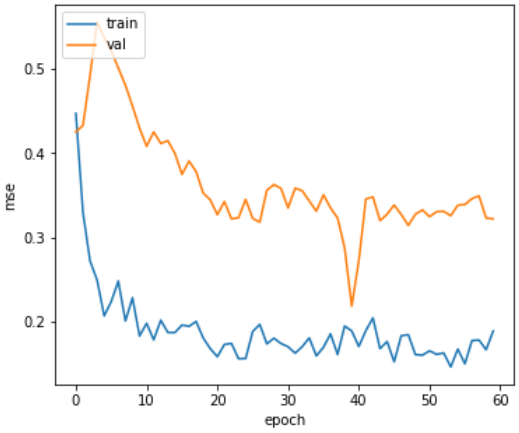

Figure 5.4.2 LSTM model results, dataset with policies

Unexpectedly, the results for the data including policy tracking information seem to be even more inaccurate. Through the visualization, a similar to the actual values pattern is identified but the line is far from the original, creating a gap that immensely increases the mean absolute error.

Now having results for the suggested LSTM model, the results overall turned out to be at lower scores than previously met ones by other algorithms. That does not necessarily mean that this algorithm is poor, but it could be making these less accurate predictions due to the shortage of data. An even lower result is scored on the second cycle where the policy tracking data were included, possibly due to the unique style of the dataset for tracking policies which did not help the algorithm.

This investigation using the Recursive Neural Networks has revealed some insights concerning the dataset and its nature. The best path to support the outcomes, is by implementing and testing other new recursive models.

5.5 Deep Neural Networks

The Deep Neural Networks (DNN) is a typically special kind of Neural Networks due to its distinct structure of data manipulation. It consists of additional hidden layers which allow data to flow from the input layers to the output layers without coming back.

Starting, as DNN model has similar foundation to the LSTM, we will be feeding the data in the same manner, splitting the COVID-19 Policy Tracker dataset from the full table at first. Moreover, following the same procedure for scaling the data and hyperparameter tuning using the Grid Search method is to be conducted once again. As mentioned, the DNN model has more layers than LSTM, multiplying the number of possible parameters to tune and making the Grid Search method even more valuable.

Table 5.5.1 DNN model results, dataset without policies / day 1

DEEP NEURAL NETWORK			
NO POLICIES IN DATA | DAY 1			
ITERATION	MSE	R_SQ	MAE
1	0,237793	0,381946	0,319484
2	0,11027	0,251454	0,684431
3	0,0907114	0,23696	0,740402
4	0,111125	0,265267	0,681983
5	0,118829	0,271639	0,659936
6	0,116433	0,268035	0,666792
7	0,0856453	0,202632	0,7549
8	0,130316	0,26639	0,629648
9	0,0807141	0,29613	0,769013
10	0,126012	0,285226	0,639379
AVERAGE	0,1074732	0,261535	0,692757

Table 5.5.2 DNN model results, dataset with policies / day 1

DEEP NEURAL NETWORK			
WITH POLICIES IN DATA | DAY 1			
ITERATION	MSE	R_SQ	MAE
1	0,117688	0,280939	0,663201
2	0,136068	0,288814	0,6106
3	0,102807	0,25199	0,705787
4	0,139401	0,285614	0,601062
5	0,08878	0,230226	0,745926
6	0,116045	0,264598	0,667902
7	0,125657	0,282594	0,640397
8	0,11205	0,281323	0,678892
9	0,128704	0,284698	0,631675
10	0,094587	0,299269	0,45991
AVERAGE	0,113504	0,272539	0,641444

The DNN model seems to generally understand the data structure from the information fed in it correctly. This is because all outputs seem to be more stable between parameterization and each iteration. Moreover, the scale between MSE and MAE is more logical and enhances how the R^2 accuracy responds. Through it we can identify an average accuracy of 26% on predicting the first day of reproduction rate with small errors scored relatively to the LSTM. Additionally, we do observe a change when running the data with the use of policy tracking. That could possibly be due to their correlation towards the desired outcome, but the change is too slight to confirm. Overall,

the Deep Neural Network seem able on identifying its given data and do make logically accurate results based on its outputs.

Table 5.5.3 DNN model results, dataset with policies / day 5

DEEP NEURAL NETWORK			
NO POLICIES IN DATA | DAY 5			
ITERATION	MSE	R_SQ	MAE
1	0,297575	0,417097	0,281776
2	0,2824466	0,411917	0,318242
3	0,285824	0,433237	0,310138
4	0,33099	0,44404	0,201127
5	0,313837	0,463841	0,242527
6	0,237981	0,372905	0,425612
7	0,282277	0,418102	0,318698
8	0,219997	0,371456	0,469018
9	0,351398	0,474799	0,151869
10	0,507379	0,570526	-0,224604
AVERAGE	0,316210375	0,44361325	0,236798125

Table 5.5.4 DNN model results, dataset with policies / day 5

DEEP NEURAL NETWORK			
WITH POLICIES IN DATA | DAY 5			
ITERATION	MSE	R_SQ	MAE
1	0,333991	0,448948	0,193882
2	0,535786	0,584877	-0,293167
3	0,344968	0,458808	0,167389
4	0,366034	0,464037	0,116545
5	0,376534	0,487123	0,0912008
6	0,34556	0,461593	0,16596
7	0,267366	0,423322	0,354688
8	0,240488	0,374915	0,41956
9	0,230235	0,380545	0,444307
10	0,367494	0,465436	0,113021
AVERAGE	0,317334875	0,439472375	0,23408385

Interestingly enough, we see a major accuracy improvement and a lower mean absolute error when predicting the fifth day of reproduction rate. This output supports the day five of false-negative accuracy reduction of the COVID-19 molecular test, indicated by the research (Lauren M. Kucirka, 2020). Moreover, since day 5 is both an accurate testing period and the average symptom day, it can be suggested that the algorithm has efficiently located such patterns within its data giving more optimum values.

6. DISCUSSION

Through research, multiple key suggestions were identified guiding the overall process and approach towards the COVID-19 pandemic in Greece and its correlation to the restrictive policies. Insights concerning the nature of the virus directed that predicting a single date should not be considered enough. Moreover, other works identifying the relativity to the spread and the mobilization of the community basically suggested the further use of restrictive measures for a better approach to the algorithms since they are part of the cause.

The data structure has always been a very significant pillar for the algorithmic processes, given that this specific style of collection did cooperate with the majority of the chosen models, it is evident to say that the given results could carry significant insights for future use.

The Random Forests Regression and Support Vector Regression are two basic machine learning algorithms that can make some great decisions, having given such high accuracy scores. That though, does not directly mean the results are at their best, these relatively high accuracy scores retrieved could be due to other factors such as very similar or linear data found between attributes. Moreover, there are many other models to consider for this type of information, and specifically for time series predictions.

Another attractive area of the project is the correlation method used for finding relativity between all attributes themselves and singularly towards the outcomes. The correlation heatmap firstly identified the reproduction rate as highly correlated, which was logical and very similar to the outcome, as it is the exact next day of the attribute's input. Moreover, strong correlation was also found within attributes concerning the death rate, total vaccinations and stringency index. We can understand how the number of total vaccinations affects the spread, as the number of immunized individuals rises the total number of likely infectious reduces. Moreover, death rate is acknowledged to be proportional to the reproduction rate, the higher rate of infections, the higher the rate of deaths. Lastly, the stringency index seemed rationale since the beginning, while the correlation outcome could only enhance this suggestion. In a similar way at which the Mobility and Activity Index showed correlation from other researchers, the restrictive measures were expectedly correlated alike.

The Neural Network algorithm is an artificial structure with layers of connected neurons. Its structure is inspired by the biological Neural Networks and consists a combination of various algorithms altogether, solving complex problems based on its data and tunings. Subclass of neural networks specifically dealing with temporal data are the Recurrent Neural Networks, and amongst its most popular ones, the LSTM model. Despite its popularity and the researchers' suggestions, the LSTM model did not qualify as the best to use for the given dataset. However, this gave the opportunity to further explore the Neural Network group of models, looking for alternative algorithms that could potentially process the given information more efficiently. Specifically, the DNN model seemed to cooperate well enough with the data giving logical predictions at first. The

extended dataset including policy tracking information did not give significant improvements by any singular attribute, while the overall outcome only had a change too slight to make assumptions despite the fact of 'stringency-index' being among the highly correlated to the output feature. However, the significant improvement in accuracy scorings for predicting day five, strongly supported this analyzed action for choosing an additional date for outcomes and additionally the choice of the DNN model.

7. CONCLUSION

Overall, there are many interesting insights to grasp from this algorithmic journey and to judge. Most definitely, more ideal models are there yet to be found, potentially giving new, better, and more interesting results. Seeking for other methods to predict such kind of information is always necessary to enhance the supported statements.

The circumstances that could affect the COVID-19 spread in the real world are endless, while models such as the DNN only use a specified number of 'circumstances' standing as attributes. The models can be used as tools of predictive gesture to efficiently get outcomes when using huge amounts of data, while implementing actions based on algorithmic assumptions should concern the community as a whole. However, as the technology in our days evolves rapidly, it is safe to say that such tools could definitely get improved and possibly become practical tools used by the administration to assist in such circumstances like epidemics in the very near future.

For this project, the dataset structure is unique based on the methodology of its data collection when it comes to government response information, based on the algorithms there was no significant difference from particular government responses other than the highly correlated stringency index. Having more analytic information of the restriction changes in Greece or any other country could potentially improve the algorithmic patterns towards a more accurate result.

With suitable tunings, the Greek government could make use of such indications for future reference, using such tool and finding fitting restrictive measures and additionally predict their impact in the society for the upcoming dates. Though, this tool and the current pandemic state are yet at a very early stage to efficiently suggest major changes and predict their impact with high certainty, while only previously met restrictions may be applied. Overall, the data collected from

such occurrences is correlated to the value of such algorithmic tools since the data fed in them becomes example and interconnects our reality with complex mathematical equations.

REFERENCES

Alkaios Sakellaris, K. M. D. S. N. T., 2020. *Covid-analysis.* [Online]
Available at: https://covid-analysis.com/
[Accessed 2021].

Anuradha Tomar, N. G., 2020. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science Direct,* Volume 728, p. 138762.

Ayan Chatterjee, M. W. G. S. G. M., 2020. Statistical Explorations and Univariate Timeseries Analysis on COVID-19 Datasets to Understand the Trend of Disease Spreading and Death. *mdpi,* 20(11), p. 3089.

BNO News, 2021. *BNO News.* [Online]
Available at: https://bnonews.com/
[Accessed 2021].

Centers for Disease Control and Prevention, 2020. *Centers for Disease Control and Prevention.* [Online]
Available at: https://www.cdc.gov/coronavirus/2019-ncov/index.html
[Accessed 2021].

Domingo, E., 2020. Long-term virus evolution in nature. *ScienceDirect.*

Dursun Delen, E. E. B. D., 2020. No Place Like Home: Cross-National Data Analysis of the Efficacy of Social Distancing During the COVID-19 Pandemic. *JMIR Public Health and Surveillance,* 6(2).

European Centre for Disease Prevention and Control, 2020. *European Centre for Disease Prevention and Control.* [Online]
Available at: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
[Accessed 2021].

Gates, B., 2015. *The next outbreak? We're not ready | Bill Gates.* s.l.: TED.

Google, 2021. *COVID-19 Community Mobility Report.* [Online]
Available at: https://www.google.com/covid19/mobility/
[Accessed 2021].

Harvard Health Publishing, 2020. Treatments for COVID-19. *Harvard Health*, Issue https://www.health.harvard.edu/diseases-and-conditions/treatments-for-covid-19.

Heneghan, C., Brassey, J. & Jefferson, T., 2020. *COVID-19: What proportion are asymptomatic?*. [Online]
Available at: https://www.cebm.net/covid-19/covid-19-what-proportion-are-asymptomatic/
[Accessed 10 May 2021].

Heymann, D. L. & Rodier, G., 2004. Global Surveillance, National Surveillance, and SARS. *Emerging Infectious Diseases*, 10(2), pp. 173-175.

İsmail Kırbaş, A. S. A. D. T. F. Ş. K., n.d. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Science Direct*, Volume 138, p. 110015.

Johns Hopkins University, 2021. *Johns Hopkins University CORONAVIRUS INFORMATION*. [Online]
Available at: https://covidinfo.jhu.edu/
[Accessed 2021].

Lauren M. Kucirka, S. A. L. O. L., 2020. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. *ACP Journals.*

MIT Medical, 2020. *How accurate are the laboratory tests used to diagnose COVID-19?*. [Online]
Available at: https://medical.mit.edu/covid-19-updates/2020/06/how-accurate-diagnostic-tests-covid-19
[Accessed 2021].

Mitronikas, A., 2021. *Prediction and Classification of Parkinson's Disease using Machine Learning*, s.l.: s.n.

OWD, 2021. *Our World In Data.* [Online]
Available at: https://ourworldindata.org/coronavirus-source-data
[Accessed 2021].

Pollack, M. P., Pringle, C., Madoff, L. C. & Memish, Z. A., 2013. Latest outbreak news from ProMED-mail: Novel coronavirus – Middle East. *International Journal of Infectious Diseases,* 17(2013), pp. e143-e144.

Sauer, L., 2020. *hopkinsmedicine.* [Online]
Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus
[Accessed 2021].

Shreshth Tuli, S. T. R. T. S. S. G., 2020. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Science Direct,* Volume 11.

Speights, K., 2020. *The Motley Fool.* [Online]
Available at: https://www.fool.com/investing/2020/04/07/here-are-all-the-companies-working-on-covid-19-vac.aspx
[Accessed 2021].

SUTD, 2020. *Data-Driven Prediction of COVID-19 Pandemic End Dates.* [Online]
Available at: https://ddi.sutd.edu.sg/
[Accessed 2021].

System Dynamics Society, 2021. *SystemDynamics.* [Online]
Available at: https://systemdynamics.org/what-is-system-dynamics/
[Accessed 2021].

The Centre for Evidence-Based Medicine CEBM, 2021. *COVID-19: What proportion are asymptomatic?.* [Online]
Available at: https://www.cebm.net/covid-19/covid-19-what-proportion-are-asymptomatic/
[Accessed 2021].

University of Oxford, 2021. *COVID-19 Government Response Tracker.* [Online]
Available at: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker
[Accessed 2021].

Univesity of Oxford, 2021. *COVID-19 GOVERNMENT RESPONSE TRACKER.* [Online]
Available at: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-

response-tracker
[Accessed 2021].

WHO, 2020. *World Health Organization Novel Coronavirus situation reports.* [Online]
Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
[Accessed 2021].

World Health Organization , 2021. *Estimating mortality from COVID-19.* [Online]
Available at: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19
[Accessed 2021].

World Health Organization, 2020. *World Health Organization Coronavirus.* [Online]
Available at: https://www.who.int/health-topics/coronavirus#tab=tab_3
[Accessed 2021].

Worldometer, 2021. *Worldometer.* [Online]
Available at: https://www.worldometers.info/coronavirus/
[Accessed 2021].

Zlatan Car, S. B. Š. N. A. I. L. V. M., 2020. Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. *Computational and Mathematical Methods in Medicine.*

BIBLIOGRAPHY

Johnson, J. (n.d.). *What's a Deep Neural Network? Deep Nets Explained*. [online] BMC Blogs.
Available at: https://www.bmc.com/blogs/deep-neural-network/.

SPRH LABS (2019). *Understanding Deep Learning: DNN, RNN, LSTM, CNN and R-CNN*. [online] Medium.
Available at: https://medium.com/@sprhlabs/understanding-deep-learning-dnn-rnn-lstm-cnn-and-r-cnn-6602ed94dbff.

Cross Validated. (n.d.). *What is the difference between a neural network and a deep neural network, and why do the deep ones work better?* [online]

Available at: https://stats.stackexchange.com/questions/182734/what-is-the-difference-between-a-neural-network-and-a-deep-neural-network-and-w.

Xia, Y. (2020). *Tune the hyperparameters of your deep learning networks in Python using Keras and Talos*. [online] Medium.
Available at: https://towardsdatascience.com/tune-the-hyperparameters-of-your-deep-learning-networks-in-python-using-keras-and-talos-2a2a38c5ac31 [Accessed 17 May 2021].

Python (2019). *Welcome to Python.org*. [online] Python.org.
Available at: https://www.python.org/.

www.tutorialspoint.com. (n.d.). *Time Series - LSTM Model - Tutorialspoint*. [online]
Available at:
https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm [Accessed 17 May 2021].

Ma'amari, M. (2018). *Deep Neural Networks for Regression Problems*. [online] Medium.
Available at: https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33.

Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras(2019). [online] Machine Learning Mastery.
Available at: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/.

Analytics Vidhya. (2019). *RNN From Scratch | Building RNN Model In Python*. [online]
Available at: https://www.analyticsvidhya.com/blog/2019/01/fundamentals-deep-learning-recurrent-neural-networks-scratch-python/ [Accessed 17 May 2021].

Scikit-learn.org. (2019). *scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation*. [online]
Available at: https://scikit-learn.org/stable/index.html.

Girgin, S. (2019). *Support Vector Regression in 6 Steps with Python*. [online] Medium. Available at: https://medium.com/pursuitnotes/support-vector-regression-in-6-steps-with-python-c4569acd062d.

DataSklr. (n.d.). *Feature selection methods with Python*. [online] Available at: https://www.datasklr.com/ols-least-squares-regression/variable-selection [Accessed 17 May 2021].

Tokuç, A.A. (2021). *Normalization vs Standardization in Linear Regression | Baeldung on Computer Science*. [online] www.baeldung.com. Available at: https://www.baeldung.com/cs/normalization-vs-standardization [Accessed 17 May 2021].