Using Machine Learning To Predict Changes In COVID-19 Related Deaths Using Policy Type

Linwood Hall halll05@students.ecu.edu
Ping Wang wangp19@students.ecu.edu
Brian Crawford crawfordbr20@students.ecu.edu

1 Abstract

Since December 2019, coronavirus disease 2019 (COVID-19) has evolved into pandemic, affecting nearly every human life on Earth. Within the United States, COVID-19 has caused state governments to enact policies mitigating the effects of this disease. Machine learning models are an important tool for predicting future events, such as deaths from COVID-19. Using an array of machine learning models, we successfully classified 67 percent of policy types, and successfully predicted 86 percent of COVID-19 related deaths in the United States.

2 Introduction

Coronavirus 2019 (COVID-19) was first recognized in Wuhan, China in late-December 2019 [3]. Since then, COVID-19 transmission has spread rapidly, becoming a pandemic. As of March 2021, 124 million people have been infected, with 2.73 million deaths related to COVID-19 worldwide[45]. This virus is thought to spread mainly through close contact, mostly within about 6 feet [3]. In response to the pandemic, the United States has enacted several policies to mitigate the effects of COVID-19. According to the Oxford COVID-19 Government Response Tracker (OxCGRT), on March 6, 2020 the U.S. had a Stingency index of 20.37 (on a scale of 100) on the COVID-19: Stringency Index [32]. Subsequently, the United States has enacted policies to slow the spread of COVID-19, such as masks and shelter in place orders[33] and as of March 7, 2021 the U.S. scored a 64.35.

The purpose of this paper is to use machine learning to predict deaths from COVID-19 for the United States. While previous attempts at predicting epidemiological features, such as COVID-19, have failed [22], machine learning uses data input from the surrounding environment [16]. Essentially, machine learning uses past data trends to predict future outcomes. To that end, we implemented an array of models to examine the effects of policy as well as changes in hospitalizations, number of cases, and infections to determine COVID-19 related deaths for the United States, as well as using classification models to predict policy type based on these same variables.

In section 3 of this paper, previous literature regarding machine learning for health related systems is reviewed. Also, variables affecting the change in COVID related deaths is discussed: various policy types(school closures, stay at home orders, gathering restrictions, etc.), COVID-19 containment, positive and negative COVID-19 tests, and number of cases. In section 4, information regarding our chosen datasets are presented. In section 5, our data methodology is discussed. This discussion will include our data processing methods, exploratory data analysis, and machine learning implementation parameters. The resulting accuracy of the machine learning model is presented. Finally, section 7 concludes the paper with overall discussion of the research, as well as future goals of this work.

3 Related Work

3.1 History of Pandemics

According to the World Health Organization (WHO), a pandemic ”occurs when simultaneous transmission takes place worldwide [4].” In addition, the most commonly used term for a pandemic is plague[20]. Some well known examples of plagues throughout history are: Black Plague, Spanish Flu, and Coronavirus Disease 2019.

3.2 History of Coronavirus Disease

The first scientific classification of coronavirus was published by two medical scientists, Tyrrell and Bynoe in 1966[46]. These two scientists took viral cultures from nasal washings of patients with cold like symptoms. The result showed that in 25 of 33 cultures, an "uncharacterized cold-producing agent" was found [46]. Around the same time, two American medical scientists found a similar virus of unknown origin. The virus was found in the human respiratory tract[19].

In the late 1960s, the "crown-like" appearance of the virus seen in surface projections resulted its name "corona" virus[24]. Further, the virus was later accepted as a new genus of viruses[47].

Throughout the following decades, many strains of coronavirus were studied. Several alarming trends were found. For example, two strains resulted in epidemics every 2-3 years[31]. Reinfection was found to be common in many strains[13]. Further, infections were found almost always in children[24].

In China (2002-2003), a new strain of coronavirus was found. The Severe Acute Respiratory Syndrome (SARS) was particularly virulent, spreading over much of the world[15]. During the 2002-2003 outbreak, SARS accounted for over 8000 infections and 774 deaths[6]. Ultimately, SARS was virulent and witnessed on a worldwide scale. Further, it would serve as a precursor to COVID-19.

3.3 Novel Coronavirus 2019

The first known case of coronavirus disease 2019 (COVID-19) appeared in Wuhan, China in late-December 2019[3]. The outbreak was epidemiologically linked to a seafood market in Wuhan[14]. Initially reported as pneumonia, the outbreak was identified as a novel (meaning new) coronavirus[14].

By March 2020, the COVID outbreak had become a pandemic[34]. Within two weeks of the first diagnosed cases, positive cases reached 1,000 infected persons. By the following week, positive cases exceeded 4,600[43] persons. This rising trend would continue through much of 2020.

According to the CDC, infected people are contagious for up to 14 days. Mild symptoms include (but are not exclusive to): fever, cough, and shortness of breath[3]. More severe symptoms include: confusion, trouble breathing, chest pressure[3]. Ultimately, severe symptoms can become fatal.

As of March 2021, COVID-19 has accounted for 2.73 million deaths worldwide[45]. Further, COVID-19 has held the highest death total of any flu season since the Spanish Flu pandemic of 1918[11; 39]. To alleviate these death totals, many governments, including the Unites States enacted policies to increase safety.

3.4 COVID-19 Policies

Governments around the world have been required to respond quickly to COVID-19 due to its rapid spread. Consequently, results have varied across the world [18]. Global cases currently stand over 109 million, with 27 million in the United States alone [48]. Responses to this rapidly growing pandemic range from geographic, such as school closings, to financial, such as income support [18]. Investigation into government responses to COVID-19 is important because public incident response can be tied to ethnic inequality, leading to over-representation of minorities in hospitalizations and deaths [40], as well as with elderly patients admitted to hospitals [17], or in nursing homes [44].

The United States is facing its most significant health related challenge since the Spanish Flu pandemic in 1918-1919 [42]. **Figure 1** shows the COVID-19 related deaths for North Carolina since March 2020. Further, COVID-19 related deaths peaked in January 2021.

With almost 500,000 deaths spread across all 50 states [48], the competitive nature of the U.S. health system has brought to light the difficulty of a national response to a disaster [42]. For example, many hospitals have trouble justifying trauma centers due to financial strain. However, a study conducted by the New England Journal of Medicine concluded that patient mortality rate was almost 2 percent higher in non-trauma centers vs. level 1 trauma centers [28]. Consequently, the need for a multi-stakeholder response to COVID-19 is necessary for government response in order to provide better decision support for policy making [36].

One major government decision is the idea of a national lockdown. According to a study of 149 countries performed by The British Medical Journal, an overall decrease of 13 percent of COVID incidence was seen [23]. In the United States, data acquired from mobile device locations found that an increase of people who stay at home of 8 per-

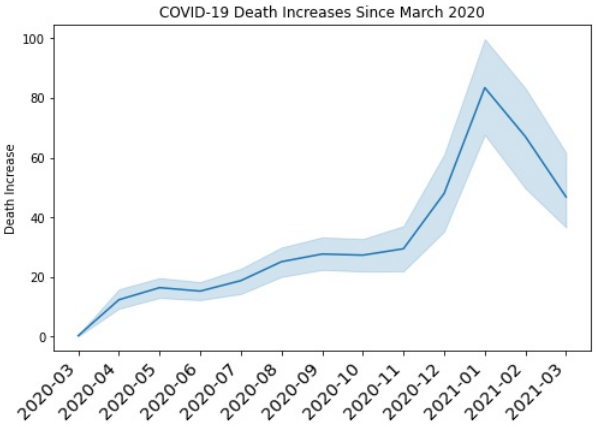

Figure 1: COVID-19 related death increases per month for North Carolina since March 2020. COVID-19 related death increases peaked in January 2021, and have decreased ever since.

cent occurred [12]. However, studies show that a targeted approach to disease mitigation generally outperform population-wide measures [26]. With cases fluctuating on a state-by-state basis, it is necessary to consider the geographical response as well due to locations of trauma centers [28], and population demographics [17].

3.5 The Relationship of Data and Epidemiology

The field of data science is integral in the global response to COVID-19[25]. For example, Ray et. al compared hypothetical lockdown duration's, resulting in proposing a longer lockdown period using data science techniques[41]. Alamo et al. reviewed different data-driven methodologies to be used in COVID-19 decision making[8]. Further, Luo et al. leveraged COVID-19 data into a "datathon" to encourage data exploration[27].

Binns et al. found that epidemiological studies of Lassa fever played a key role in diminishing severity[10]. COVID-19, however, showed an increase in the aspects of public health involved. In a study of China's epidemiological surveillance system, Idrovo and Manrique-Hernandez found that this system had good data quality [21].

One key aspect of COVID-19 study is forecasting the effects of the pandemic. Forecasting is especially important for healthcare systems that are pressed for resources. For example, Vinod and Prabaharan found the use of Artificial Intelligence (AI) useful in providing positive COVID-19 tests examined from X-Ray images. As a result, they were able to diagnose COVID-19 at an 88 percent precision[49]. Machine learning is another tool, that uses data trends to model and predict future outcomes.

3.6 Machine Learning

According to El Naqa and Murphy, machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment[16]. Mohri et al. define machine learning as computational methods using experience to improve performance or to make accurate predictions[30]. As I mentioned before, machine learning and prediction is especially important for healthcare systems with pressed resources [49].

In the past, predicting healthcare related issues has been difficult[35], and had failed in accuracy for the earlier stages of COVID-19[22]. However, as more data becomes available, techniques are getting better. Vinod and Prabarahan were able to diagnose coronavirus disease 19 using artificial intelligence on X-ray images[49]. Also, Devaraj et al. employed several machine learning techniques to predict death and other features at approximately 90 percent accuracy[7]. Further, a group of Hungarian medical scientists used hybrid machine learning to predict future COVID-19 outbreaks. This research yielded promising results[38].

Therefore, machine learning is a broad field with many applications. These applications have extended to the healthcare field. As a result, using machine learning techniques for COVID-19 analysis is a viable method to predict a wide spectrum of pandemic related effects.

4 Data

The dataset we will be using is from three sources:

1. The COVID Tracking Project
2. COVID-19 State and County Policy Orders
3. Oxford Covid-19 Government Response Tracker (OxCGRT)

The time frame examined for this research is between March 6, 2020 through March 7, 2021.

4.1 The COVID Tracking Project

The COVID Tracking Project is a volunteer based effort designed to collect and publish

COVID-19 data. Sponsored by *The Atlantic*, this organization is dedicated to providing accurate data through daily validation and cross-checking. This research began in early March 2020 based on the investigative efforts of Robinson Meyer and Alexis Madrigal. Around the same time, Jeff Hammerbacher (Founder and General Partner at **Related Sciences**), who was doing similar research, joined the two journalists in what became known as **The COVID Tracking Project**[1].

Since this project's inception, the data definitions have changed as a result of differences in the 56 jurisdictions that information is collected from. The data was then grouped into seven different categories: cases, PCR tests, antibody tests, antigen tests, hospitalization, outcomes, state metadata. All of the data was collected from the websites of state/territory public health authorities[1].

For the purposes of this study, information regarding the state name, date, and the daily increase in deaths were collected from **The COVID Tracking Project**. First, the state name is important when joining the datasets to ensure the geographical scope of the final dataset stays consistent. Next, the date is important when comparing this dataset to other datasets. This comparison ensures consistency between datasets in terms of the examined time frame. Finally, the daily increase in deaths is significant to this study in that it is a quantifiable value of the mortality related to COVID-19. The dependent variable for the machine learning models in this research is the daily increase in deaths. This data is based on the previous day's death total[1]. In general, forecasting COVID-19 effects is important[9]. While confirmed cases are important for policy making [29], predicting the daily increase in deaths could prove equally important.

4.2 COVID-19 State and County Policy Orders

The **COVID-19 State and County Policy Orders** from Data.gov, contains 4,220 records with 10 variables. The information contained within this dataset includes the policy type, as well as start/end dates. We took information regarding **state level** policies for North Carolina (Shelter In Phase, Phase 1, and Phase 2), and created a discretized variable (**policy**) within our processed dataset. **Figure 2** details the monthly death increases per policy type.

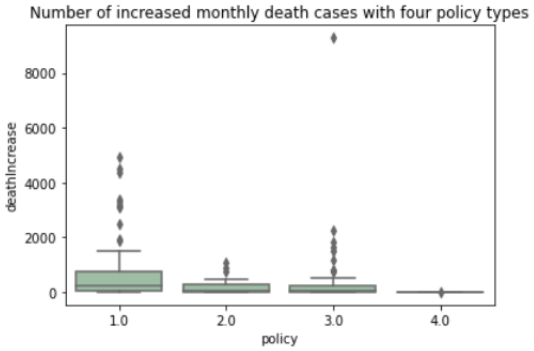

Figure 2: Boxplot of monthly death increase on four policy types. X-axis: policy types(1: Mandate Face Mask Use By All Individuals In Public Facing Businesses, 2:Mask Requirement, 3:Shelter in Place, 4:Mandate Face Mask Use By All Individuals In Public Spaces); y-axis:number of death increases. The boxplot shows the outliers, minimum, maximum, 1st quantile, median, and 3rd quantile of the death increases.

4.3 COVID-19 Stringency Index

The University of Oxford has provided a stringency index based on government response to COVID-19. This response index includes nine response indicators of government response including school and workplace closures, and travel bans [32].

For the state level dataset, information was collected once-a-week and is still continually maintained by the University of Oxford. **Figure 3** shows the average Stringency Index per state between the time frame we examined.

In comparison to the national level stringency tracker, the state level dataset includes information on closings for school and work, as well as travel restrictions. **Figure 4** shows the ten most stringent states for government response to COVID-19. To place a personal geographic marker on this research, North Carolina was the tenth most stringent state in the dataset [2].

The stringency index is the most important field from this dataset. However, other data is useful as well. Not only does **Figure 4** provide information about the stringency index, it also shows that there is a clear differentiation in the political affiliations of the governor in each state.

Finally, the data is provided daily at the state level. The stringency index is an overall indicator of government response, and is used in this research as a predictor of COVID-19 related deaths.

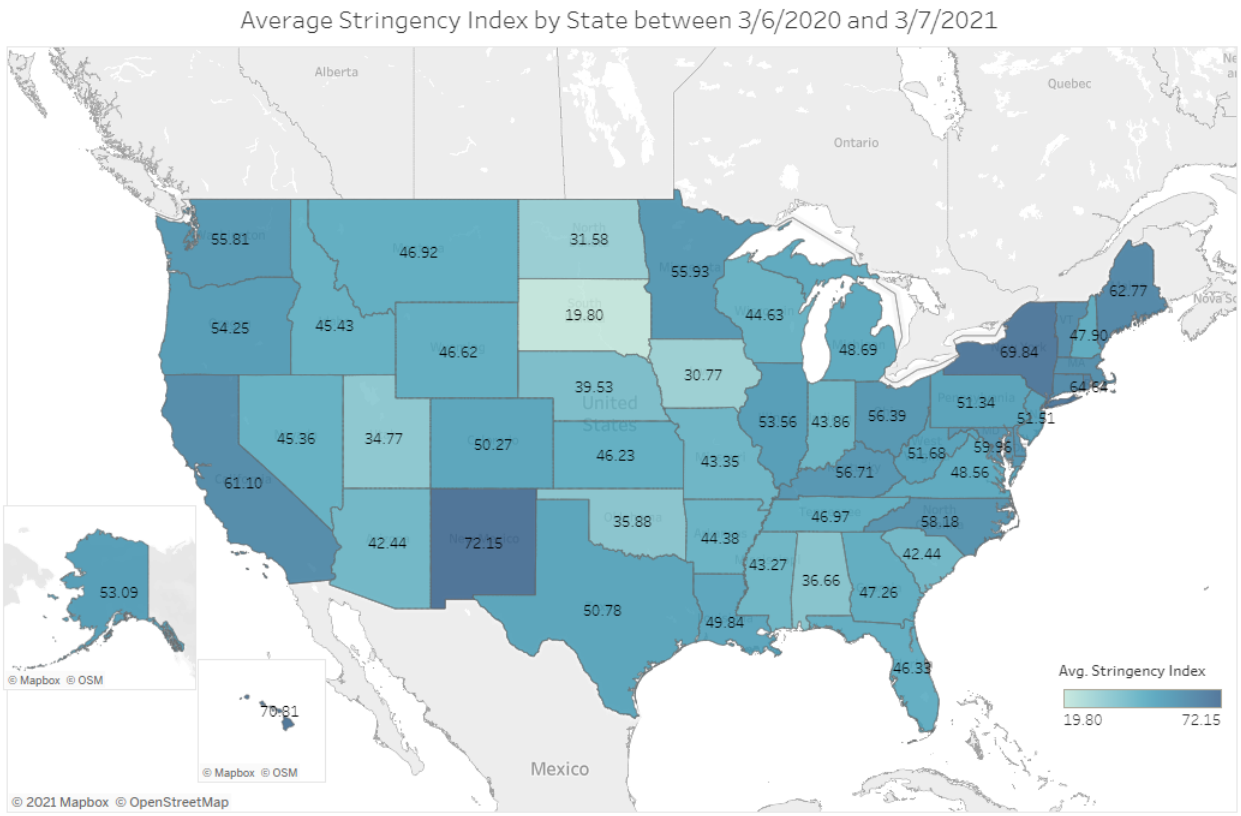

Figure 3: Average Stringency Index by State between March 6, 2020 and March 7, 2021.

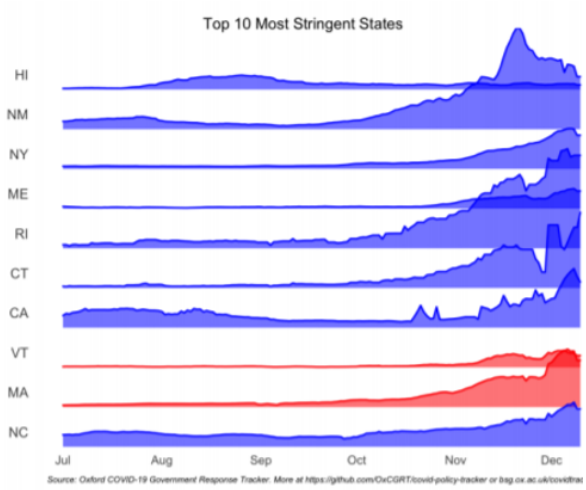

Figure 4: The top ten most stringent states for COVID-19 government response. Interestingly, North Carolina is the tenth most stringent state for government response. The colors in the map represent the political affiliation of the governor for each state. Only two republican (red) states, Vermont and Massachusetts, are present. Source: https://www.bsg.ox.ac.uk/sites/default/files/2020-12/BSG-WP-2020-034-v2.pdf

5 Method

This research required us to gather data from three websites: The COVID Tracking Project, COVID-19 State and County Policy Orders dataset, and the COVID-19 Stringency Index. Subsection 5.1 explains the data preprocessing steps. Next, data analysis is discussed in subsection 5.2. Subsection 5.3 details assumptions made about the data. Subsection 5.4 discusses the feature selection methods. Finally, subsection 5.5 documents the models chosen for this research.

5.1 Data Preprocessing

Collecting the data from multiple sites created a unique issue for this research. The data is time sensitive, so the date field for each dataset is important for data consistency and integrity.

First, each dataset was compared to see if a date field was present. This field was then compared among the three datasets. Further, the date ranges were compared. As a result, the ultimate date range of the research was chosen based upon the smallest time duration within the datasets. In this case, approximately one year of data was cropped from each dataset.

Classifier	Accuracy Score
K-Nearest Neighbors	0.67
Logistic Regression	0.49
Decision Tree	0.65
Support Vector Machine	0.52
Gaussian Naive Bayes	0.37
Linear Discriminant	0.53

Table 1: Classifier models ran against the base dataset for this COVID-19 data analysis. Early model runs were inaccurate due to noisy and inconsistent data. The data inconsistency was a result of inconsistent data collection methods between states.

Next, each dataset was combined on the date field. Once combined, the datasets could be examined as one whole dataset. There was no unexplained NULL data within these datasets. Each table had been curated to ensure data consistency.

5.2 Data Analysis

With the combined datasets, data analysis is important to see what variables had the most impact on COVID-19 deaths. The combined dataset has 18,667 records and 31 features. These features are a mixture of nominal data for geographic location, dates, policy information, and various indices.

As part of the data analysis, six classification models were constructed to predict policy type based on COVID-19 related deaths, number of positive test results, number of hospitalized patients and number of recovered patients. As shown in **Table 1**, K-Nearest Neighbor Classifier and Decision Tree Classifier have the highest accuracy scores–0.67 and 0.65 respectively. The result shows that policy type has influence on the infection rate. However, noise within the dataset was present, and thus, needed to be removed.

At first glance, some of the more important features in the dataset are the policies. The policy types in the dataset include Stay-At-Home orders, gathering restrictions, and closures for schools and workplaces. This data is discrete and are categorized in increasing severity. For example, a 0 index would be considered as though no measures were in place. On the other hand, the highest index for the policy type group would be the most severe. For example, for a policy type group of school closures, a designation of 3 would be for all schools to close completely [32]. **Figure 4** illustrates the OxCGRT Indicator Codes for the three policy type groups [32]. These policy type groups are a part of the overall government response to COVID-19.

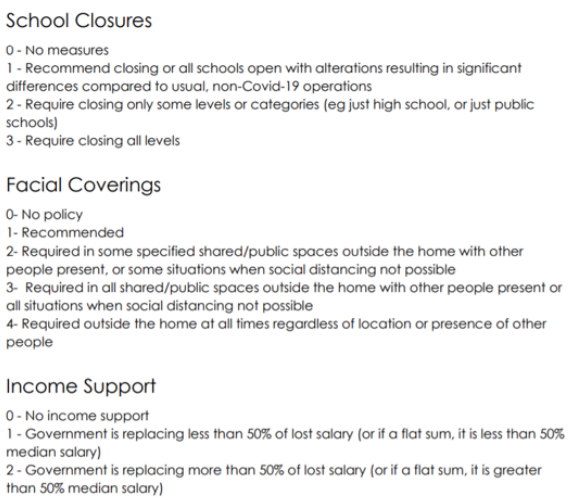

Figure 5: The OxCGRT Indicator Codes for each policy type grouping. The three policy type groups are: school closures, facial coverings, and income support. Each code increases in severity as the categorical value increases. Source: OxCGRT Indicator Codes

The stringency index is the government response to COVID-19, in the form of lockdown procedures [32]. Essentially, the more stringent a state is, the more restrictions were put in place in response to COVID-19. Thus, as government response increases, so do the number of policy implementations. However, other factors play a role in the stringency index as well.

Another variable in this dataset are the government response index [32]. This index is similar to stringency. However, as expected, the government response index is the overall state-level response to the pandemic. In essence, it shows the variation in how each state government responded to the pandemic.

Other variables to note would be the containment health and economic support indices. These indices illustrate how well the government responded to outbreaks, and financial support for the pandemic, respectively [32].

5.3 Data Assumptions

From the data analysis, assumptions on the data can be made. First, policy should have a negative effect on COVID-19 related deaths. Next, government response should have a positive effect on the policy implementations, as well as COVID-19 deaths. Finally, policy implementations have a positive effect on the containment index.

Policy implementations indicate an overall government response to COVID-19. For example, if a stay-at-home order is issued by a state government, inherently people cannot spread the virus. Therefore, the more stay-at-home orders should indicate a drop in deaths over time. Of course, the number of deaths would increase at first, hence the policy being implemented. However, over time, a steady decline in deaths should occur.

In parallel, overall government response should have a positive effect on COVID-19 policy implementations. Further, the higher the level of government response, the more policy implementations there will be. In contrast, the higher the level government response, COVID-19 related deaths should lower.

Finally, if there is a higher the level of policy implementations, the result should be an increase in the containment index. In essence, if more COVID-19 restrictions are put into place, there should be fewer outbreaks of COVID-19 under normal circumstances.

5.4 Feature Selection

The chosen dataset holds 31 features. Using all of these features can become noisy, and can have a negative impact on the machine learning models. Therefore, feature selection is necessary to simplify the number of features that impact modeling.

To simplify the features on the model, the **SelectKBest** feature selection method was used. This package can be found in the feature selection library of the scikit-learn machine learning stack. It uses univariate testing methods to determine the 'k' highest scores based on the linear dependency two features have with one another [37].

Using SelectKBest, four features were chosen based on their proposed influence on the machine learning models. The selected features were based on increases in death, the government response index, facial covering policy, and vaccination policy.

5.5 Models

Machine learning was chosen to solve this research question. Since the response variable is COVID-19 related deaths, and is therefore a continuous datatype, regression was the best method. Further, various regression models were implemented using the scikit-learn machine learning library for the Python programming language [37]. **Figure 6**, illustrates a method of choosing machine learning models to implement [5]. For this research, we will present several models and their results. This presentation of various models is a result of the exploratory nature of this research to predict a singular value.

This research attempted to predict COVID-19 related deaths. Using **Figure 6** as a template, the following models were implemented: ridge linear regression, elastic net, support vector regression, and ensemble bagging regressor.

This research also attempted to classify policy type based on COVID-19 related deaths as well. Using **Figure 6** as a template, the following models were implemented:

5.5.1 Ridge Linear Regression

Using Ridge Linear Regression, COVID related deaths were predicted based on the selected features. Ridge linear regression uses a penalty on the size of the coefficients. This results in solving some of the issues related to Ordinary Least Squares [37]. For this model, the data was split into a 70-30 train/test split. Further, the data was cross validated 5 times to prevent overfitting on the data, as it is time series dependent.

5.5.2 Elastic Net

Elastic Net Regression uses the combined L1 and L2 regression regularization methods found in ridge and lasso regression. In a sense elastic net regression is the combination of ridge and lasso regression. This model is useful when multiple features are correlated with one another [37]. For this model, the data was split into a 70-30 train/test split. Further, the data was cross validated 5 times to prevent overfitting on the data, as it is time series dependent.

5.5.3 Support Vector Regression

Support Vector Regression (SVR) is a part of the support vector machine (SVM) library within scikit-learn. SVR is particularly useful for high dimensional data. However, these models a weak against datasets where the number of features are greater than the number of samples [37].

5.5.4 Bagging Regressor

The Bagging Regressor model is part of the ensemble methods within scikit-learn. Ensemble methods are useful in that they prepare a well performing model by combining several weaker models. Bagging models work well against complex datasets due to its method of reducing the chance of overfitting.

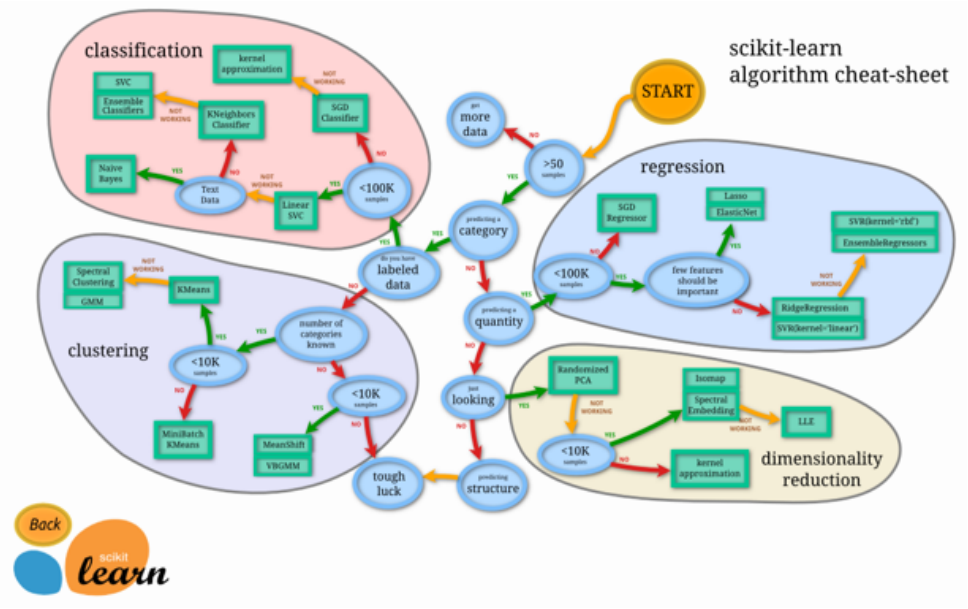

Figure 6: An simple illustrated path for choosing the best machine learning model. Some of the decisions along the path relate to sample size, the response variable datatype, as well as what alternate models should be implemented if a proposed model does not provide the desired results. Source: Analytics Vidhya Improving Machine Learning Results

5.5.5 KNN classifier

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN algorithm is used both in classification and regression problems. KNN is based on feature similarity.

5.5.6 Logistic regression classifier

Logistic regression is usually used when the target variable is categorical variable. It predicts the probability of each category of the class attribute. Threshold is specified to indicate at what value the examples will be classified as on class vs. the other classes.

5.5.7 Decision tree classifier

Decision tree classifier predict the class of target variables represented by the leaves of the tree structure. It can also be used to visualize the decision in decision making process.

5.5.8 Support Vector Classifier

Support vector classifier(SVC) distinctly classify data points by finding a hyperplane in an N-dimensional space.

5.5.9 Gaussian Naive Bayes classifier

Gaussian Naive Bayes classification is one of the Naive Bayes classifier family which is based on the Bayes Theorem. Gaussian Naive Bayes classifier is probability based classifier and can be applied when features follow Gaussian distribution.

5.5.10 Linear Discriminant Analysis classifier

Linear Discriminant Analysis(LDA) classifier is related to analysis of variance(ANOVA). But unlike ANOVA, which uses categorical independent variables to detect continuous dependent variable, LDA uses continuous independent variables to classify examples to dependent categories(ie. classes).

6 Results

Machine learning models were built to gauge the predictability of COVID-19 deaths, and the classification of enacted policies in the United States. The first subsection details the prediction of COVID-19 related deaths for the United States. The second subsection details the classification of state level policy types in the U.S. related to COVID-19.

6.1 Predicting COVID-19 Deaths With Linear Regression Models

6.1.1 Ridge Linear Regression

Using Ridge Linear Regression, COVID-19 related deaths were predicted based on the selected

features of: increases in death, the government response index, facial covering policy, and vaccination policy. Using this, as well as the RepeatedKFold cross validation technique (5-fold). This model resulted in a 32% accuracy. One reason for this low accuracy is the number of data samples, as well as the ridge model's susceptibility to outlier data. With COVID-19 deaths, especially around April 2020, some states experienced over 300+ deaths in one day, whereas, other states experienced less than 100, or even 50 deaths in one day. For the purposes of a linear relationship, this proved detrimental to the model accuracy.

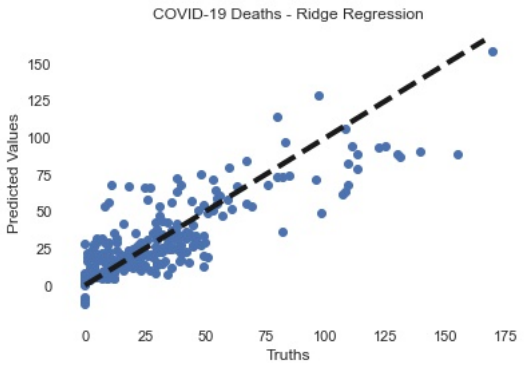

Figure 7: Ridge regression model detailing the relationship between the predicted and truth values. Higher predicted values for deaths indicate a lower accuracy.

6.1.2 Elastic Net Regression

Elastic Net Regression takes into account both penalties on the L1 and L2 regularization techniques found in Lasso and Ridge regression, respectively. Using the selected features (as well as using RepeatedKFold for 5 folds and a train/test split of 70-30) this model performed at a 32% accuracy. One of the weaknesses of the Elastic Net model is that it can struggle if one regularization parameter is not performing well. For this instance, the Ridge Regression model seen in section 6.1.1 did not perform well, and had a negative effect on the Elastic Net model. **Figure 8** shows the predictions related to the Elastic Net model. The blue dots are the original values from the test dataset. The red line shows the predicted values. This regression model has difficulty in predicting the outlier values.

6.1.3 Support Vector Regression

Support Vector Regression works better in a higher dimension dataset. However, these models tend to struggle with larger datasets, and with

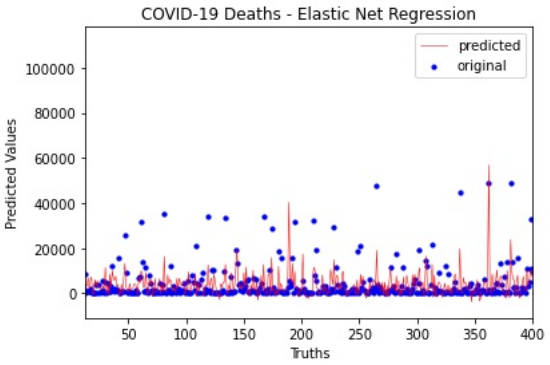

Figure 8: Elastic Net Regression model detailing the relationship between the predicted and truth values. Higher predicted values for deaths indicate a lower accuracy. This model had trouble predicting the outliers, as shown by the blue dots (original) and the red line (predicted values).

features that are highly correlated. As a result, this model performed at a less than 10% accuracy. One issue with this dataset is that if there is a lot of correlation between the variables, this model will have a lower performance. **Figure 9** shows the predicted vs. actual values for the SVR model. This model had a great deal of trouble predicting the outlier ground truth values. Therefore, it did not perform well with the COVID-19 Stringency dataset.

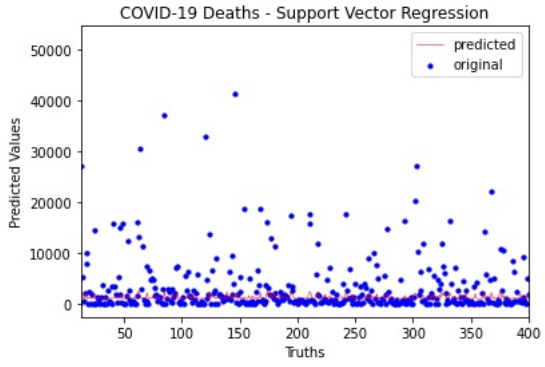

Figure 9: Support Vector Regression model detailing the relationship between the predicted and truth values. This model had trouble predicting the outliers, as shown by the blue dots (original) and the red line (predicted values).

6.1.4 Bagging Regressor

The Bagging Regressor model is part of the ensemble regressors in scikit-learn. Bagging generally does well to prevent overfitting, which is important for the time-based data seen in the COVID-19 dataset. For this dataset, the Bagging

Regressor correctly predicted COVID-19 related deaths at 86% accuracy. This performance was the best model run on this dataset. Part of the success of this model is that it takes the weaker models and combines them into one "better" model. It also tends to work well with datasets under 100,000 samples, as seen in the path from **Figure 6**.

Figure 10 shows the model predictions versus the actual ground truth values. This model performed particularly well. Since bagging models tend to do well at reducing overfitting, this model was successful at predicting the outlier values that the other models could not.

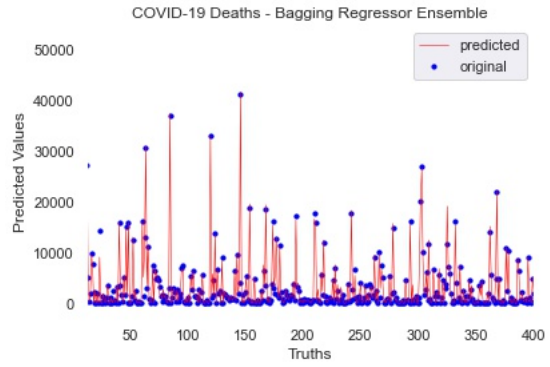

Figure 10: Bagging Regressor ensemble model detailing the relationship between the predicted and truth values. This model, as expected from figure 5, works well with datasets that are small or large. Bagging models excel at reducing overfitting, therefore this result is the best model result for regression.

6.2 Predicting Policy Types With Classification Models

Like in the preliminary analysis, we built classification models to classify policy types. But for the combined dataset, we have different stringency levels for each of the eight policy orders–School closing, Workplace closing, Cancel public events, Restrictions on gatherings, Close public transport, Stay at home, Restrictions on internal movement, and International travel controls. Instead of classifying the simple index of policy type(0,1,2,3), we built six classification models to classify different stringency level for the eight policy orders. K-Nearest Neighbors classifier, Logistic classifier, Decision tree classifier, Support vector classifier, Gaussian Naive Bayes classifier and Linear Discriminant classifier are applied. Selected features for classification include 'ConfirmedCases', StringencyIndex', 'GovernmentResponseIndex', 'ContainmentHealthIndex', and 'deathIncrease'.

Tables 2 through 9 below show the classification results of eight policy orders. The highest accuracy scores for classifying School closing, Workplace closing, Cancel public events, Restrictions on gatherings, Close public transport, Stay at home requirements, Restrictions on internal movement, and international travel controls are 0.72, 0.76, 0.83, 0.68, 0.67, 0.73, 0.71 and 0.72 respectively. The accuracy scores indicate that policy stringency can be classified which means that policy types have influence on death increase.

Classifier	Accuracy Score
K-Nearest Neighbors	0.70
Logistic Regression	0.70
Decision Tree	0.67
Support Vector Machine	0.69
Gaussian Naive Bayes	0.67
Linear Discriminant	**0.72**

Table 2: Classification accuracy scores of stringency category of "School closing" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.50
Logistic Regression	0.52
Decision Tree	0.63
Support Vector Machine	0.50
Gaussian Naive Bayes	0.67
Linear Discriminant	**0.76**

Table 3: Classification accuracy scores of stringency category of "Workplace closing" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.62
Logistic Regression	0.62
Decision Tree	0.68
Support Vector Machine	0.57
Gaussian Naive Bayes	0.66
Linear Discriminant	**0.83**

Table 4: Classification accuracy scores of stringency category of "Cancel public events" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.49
Logistic Regression	0.46
Decision Tree	0.63
Support Vector Machine	0.46
Gaussian Naive Bayes	0.61
Linear Discriminant	**0.68**

Table 5: Classification accuracy scores of stringency category of "Restrictions on gatherings" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.54
Logistic Regression	0.54
Decision Tree	0.63
Support Vector Machine	0.57
Gaussian Naive Bayes	0.60
Linear Discriminant	**0.67**

Table 6: Classification accuracy scores of stringency category of "Close public transport" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.59
Logistic Regression	0.52
Decision Tree	0.57
Support Vector Machine	0.52
Gaussian Naive Bayes	0.62
Linear Discriminant	**0.73**

Table 7: Classification accuracy scores of stringency category of "Stay at home" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.49
Logistic Regression	0.43
Decision Tree	0.60
Support Vector Machine	0.43
Gaussian Naive Bayes	**0.71**
Linear Discriminant	0.68

Table 8: Classification accuracy scores of stringency category of "Restrictions on internal movement" policy using six classifiers

Classifier	Accuracy Score
K-Nearest Neighbors	0.66
Logistic Regression	0.65
Decision Tree	0.62
Support Vector Machine	0.43
Gaussian Naive Bayes	**0.72**
Linear Discriminant	0.68

Table 9: Classification accuracy scores of stringency category of "International travle controls" policy using six classifiers

7 Discussion and Future Work

One of the bigger issues with this study is the inconsistency in which data collection occurs on a state level. Although the datasets were largely aggregated, the consistency between states' collection methods was piecemeal. This is a result of state departments holding different requirements for the usage of their data.

Initial model runs are accurately predicting at approximately 65-70 percent. For the rest of this project, we will be focusing on data visualization, as well as improving some of the models. To this end, research will be done on specific models that have been run against COVID-19 data.

The techniques described in this paper can be expanded in multiple ways. First, we could expand our study to include data throughout the entire world. If taken as a whole, the increase in data records would allow the models to train on more examples, which should result in higher accuracy values. Second, we could expand this research to predict other values, such as hospitalization increases. This prediction could help healthcare professionals with the awareness of when outbreaks will occur, as well as how to manage the number of beds within each hospital. Finally, this research could be expanded to examine extreme COVID-19 outbreaks in deaths, as well as, positive case increases.

References

[1] URL https://covidtracking.com/about-data/sources.

[2] URL https://www.bsg.ox.ac.uk/research/publications/variation-us-states-responses-covid-19.

[3] Coronavirus (covid-19) frequently asked questions. URL https://www.cdc.gov/coronavirus/2019-ncov/faq.html.

[4] WHO | The classical definition of a pandemic is not elusive. URL https://www.who.int/bulletin/volumes/89/7/11-088815/en/.

[5] Dec 2015. URL https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/.

[6] Sars, Dec 2017. URL https://www.cdc.gov/sars/about/fs-sars.html.

[7] *Results in Physics*, 21:103817, Feb 2021. ISSN 2211-3797. doi: 10.1016/j.rinp.2021.103817.

[8] T. Alamo, D. G. Reina, and P. Millán. Data-driven methods to monitor, model, forecast and control covid-19 pandemic: Leveraging data science, epidemiology and control theory. *arXiv:2006.01731 [physics, q-bio]*, Jun 2020. URL http://arxiv.org/abs/2006.01731. arXiv: 2006.01731.

[9] N. Altieri, R. L. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y. S. Tan, et al. Curating a covid-19 data repository and forecasting county-level death counts in the united states. *arXiv preprint arXiv:2005.07882*, 2020.

[10] C. Binns, W. Y. Low, and L. M. Kyung. The covid-19 pandemic: Public health and epidemiology. *Asia Pacific Journal of Public Health*, 32(4): 140–144, May 2020. ISSN 1010-5395, 1941-2479. doi: 10.1177/1010539520929223.

[11] R. R. Britt. The Coronavirus Has Killed More Americans Than Any Flu Season Since 1918, Oct. 2020. URL https://elemental.medium.com/us-covid-19-deaths-compared-to-diseases-pandemics-wars-2a7495a43280.

[12] A. Brzezinski, G. Deiana, V. Kecht, and D. Van Dijcke. The covid-19 pandemic: government vs. community action across the united states. *Covid Economics: Vetted and Real-Time Papers*, 7:115–156, 2020.

[13] K. A. Callow, H. F. Parry, M. Sergeant, and D. A. J. Tyrrell. The time course of the immune response to experimental coronavirus infection of man. *Epidemiology and Infection*, 105(2): 435–446, Oct. 1990. ISSN 0950-2688, 1469-4409. doi: 10.1017/S0950268800048019. URL https://www.cambridge.org/core/product/identifier/S0950268800048019/type/journal_article.

[14] M. Ciotti, S. Angeletti, M. Minieri, M. Giovannetti, D. Benvenuto, S. Pascarella, C. Sagnelli, M. Bianchi, S. Bernardini, and M. Ciccozzi. COVID-19 Outbreak: An Overview. *Chemotherapy*, 64(5-6):215–223, 2019. ISSN 0009-3157, 1421-9794. doi: 10.1159/000507423. URL https://www.karger.com/Article/FullText/507423.

[15] C. Drosten, S. Günther, W. Preiser, S. van der Werf, H.-R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R. A. M. Fouchier, A. Berger, A.-M. Burguière, J. Cinatl, M. Eickmann, N. Escriou, K. Grywna, S. Kramme, J.-C. Manuguerra, S. Müller, V. Rickerts, M. Stürmer, S. Vieth, H.-D. Klenk, A. D. M. E. Osterhaus, H. Schmitz, and H. W. Doerr. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *The New England Journal of Medicine*, 348(20): 1967–1976, May 2003. ISSN 1533-4406. doi: 10.1056/NEJMoa030747.

[16] I. El Naqa and M. J. Murphy. What is machine learning? *Machine Learning in Radiation Oncology*, page 3–11, 2015. doi: 10.1007/978-3-319-18305-3_1.

[17] S. Garg, L. Kim, M. Whitaker, A. O'Halloran, C. Cummings, R. Holstein, M. Prill, S. J. Chai, P. D. Kirley, N. B. Alden, et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—covid-net, 14 states, march 1–30, 2020. *Morbidity and mortality weekly report*, 69(15):458, 2020.

[18] T. Hale, N. Angrist, E. Cameron-Blake, L. Hallas, B. Kira, S. Majumdar, A. Petherick, H. T. Toby Phillips, and S. Webster. Variation in government responses to covid-19. *Blavatnik School of Government working paper*, 32(7):2020–11, 2020. doi: https://www.bsg.ox.ac.uk/sites/default/files/2020-09/BSG-WP-2020-032-v7.0.pdf.

[19] D. Hamre and J. J. Procknow. A new virus isolated from the human respiratory tract. *Experimental Biology and Medicine*, 121(1):190–193, 1966. doi: 10.3181/00379727-121-30734.

[20] D. Huremović. Brief History of Pandemics (Pandemics Throughout History). In D. Huremović, editor, *Psychiatry of Pandemics: A Mental Health Response to Infection Outbreak*, pages 7–35. Springer International Publishing, Cham, 2019. ISBN 9783030153465. doi: 10.1007/978-3-030-15346-5_2. URL https://doi.org/10.1007/978-3-030-15346-5_2.

[21] A. J. Idrovo and E. F. Manrique-Hernández. Data quality of chinese surveillance of covid-19: Objective analysis based on who's situation reports. *Asia Pacific Journal of Public Health*, 32(4):165–167, May 2020. ISSN 1010-5395, 1941-2479. doi: 10.1177/1010539520927265.

[22] J. P. Ioannidis, S. Cripps, and M. A. Tanner. Forecasting for covid-19 has failed. *International Journal of Forecasting*, 2020. doi: 10.1016/j.ijforecast.2020.08.004.

[23] N. Islam, S. J. Sharp, G. Chowell, S. Shabnam, I. Kawachi, B. Lacey, J. M. Massaro, R. B. D'Agostino, and M. White. Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. *bmj*, 370, 2020.

[24] J. S. Kahn and K. McIntosh. History and recent advances in coronavirus discovery. *Pediatric Infectious Disease Journal*, 24(11), 2005. doi: 10.1097/01.inf.0000188166.17324.60.

[25] S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, and et al. Leveraging data science to combat covid-19: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, 1(1):85–103, Aug 2020. ISSN 2691-4581. doi: 10.1109/TAI.2020.3020521.

[26] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438 (7066):355–359, 2005.

[27] E. M. Luo, S. Newman, M. Amat, M.-L. Charpignon, E. R. Duralde, S. Jain, A. R. Kaufman, I. Korolev, Y. Lai, B. D. Lam, and et al. Mit covid-19 datathon: data without boundaries. *BMJ Innovations*, 7(1), Jan 2021. ISSN 2055-8074, 2055-642X. doi: 10.1136/bmjinnov-2020-000492. URL https://innovations.bmj.com/content/7/1/231.

[28] E. J. MacKenzie, F. P. Rivara, G. J. Jurkovich, A. B. Nathens, K. P. Frey, B. L. Egleston, D. S. Salkever, and D. O. Scharfstein. A national evaluation of the effect of trauma-center care on mortality. *New England Journal of Medicine*, 354(4):366–378, 2006.

[29] M. Maleki, M. R. Mahmoudi, M. H. Heydari, and K.-H. Pho. Modeling and forecasting the spread and death rate of coronavirus (covid-19) in the world using time series models. *Chaos, Solitons Fractals*, 140:110151, 2020. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2020.110151. URL https://www.sciencedirect.com/science/article/pii/S0960077920305476.

[30] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. Adaptive computation and machine learning. The MIT Press, second edition edition, 2018. ISBN 9780262039406.

[31] A. S. Monto. Medical reviews. Coronaviruses. *The Yale Journal of Biology and Medicine*, 47(4): 234–251, Dec. 1974. ISSN 0044-0086.

[32] B. S. of Government. Covid-19 government response tracker. URL https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker.

[33] T. D. of Health and H. Services. North carolina's strategy to combat covid-19. URL https://covid19.ncdhhs.gov/slow-spread/north-carolinas-strategy-combat-covid-19.

[34] W. H. Organization. Timeline: Who's covid-19 response. URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline.

[35] T. Panch, P. Szolovits, and R. Atun. Artificial intelligence, machine learning and health systems. *Journal of Global Health*, 8(2). ISSN 2047-2978. doi: 10.7189/jogh.08.020303. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6199467/.

[36] S. Panneer, K. Kantamaneni, R. R. B. Pushparaj, S. Shekhar, L. Bhat, and L. Rice. Multistakeholder participation in disaster management—the case of the covid-19 pandemic. In *Healthcare*, volume 9, page 203. Multidisciplinary Digital Publishing Institute, 2021.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[38] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen. Covid-19 pandemic prediction for hungary; a hybrid machine learning approach. *Mathematics*, 8(6):890, Jun 2020. doi: 10.3390/math8060890.

[39] S. D. Pitlik. COVID-19 Compared to Other Pandemic Diseases. *Rambam Maimonides Medical Journal*, 11(3), July 2020. ISSN 2076-9172. doi: 10.5041/RMMJ.10418. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7426550/.

[40] L. Platt and R. Warwick. Are some ethnic groups more vulnerable to covid-19 than others? *Institute for Fiscal Studies*, May 2020. doi: http://www.ifs.org.uk/inequality/wp-content/uploads/2020/04/Are-some-ethnic-groups-more-vulnerable-to-COVID-19-than-others-V2-IFS-Briefing-Note.pdf.

[41] D. Ray, M. Salvatore, R. Bhattacharyya, L. Wang, J. Du, S. Mohammed, S. Purkayastha, A. Halder, A. Rix, D. Barker, and et al. Predictions, role of interventions and effects of a historic national lockdown in india's response to the covid-19 pandemic: data science call to arms. *Harvard data science review*, 2020(Suppl 1), 2020. ISSN 2688-8513. doi: 10.1162/99608f92.60e08ed5. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7326342/.

[42] K. N. Remick, B. Carr, and E. Elster. Covid-19: Opportunity to re-imagine our response to a national medical crisis. *Journal of the American College of Surgeons*, Feb 2021. doi: 10.1016/j.jamcollsurg.2021.01.013.

[43] A. Spinelli and G. Pellino. COVID-19 pandemic: perspectives on an unfolding crisis. *British Journal of Surgery*, 107(7):785–787, June 2020. ISSN 0007-1323. doi: 10.1002/bjs.11627. URL https://doi.org/10.1002/bjs.11627.

[44] D. G. Stevenson and A. K. Cheng. Nursing home oversight during the covid-19 pandemic. *Journal of the American Geriatrics Society*, 2021.

[45] T. N. Y. Times. Coronavirus in the u.s.: Latest map and case count, Mar 2020. URL https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.

[46] D. Tyrrell and M. Bynoe. Cultivation of viruses from a high proportion of patients with colds. *The Lancet*, 287(7428):76–77, 1966. doi: 10.1016/s0140-6736(66)92364-6.

[47] D. Tyrrell, J. Almeida, C. Cunningham, W. Dowdle, M. Hofstad, K. McIntosh, M. Tajima, L. Zakstelskaya, B. Easterday, A. Kapikian, and et al. Coronaviridae. *Intervirology*, 5(1-2):76–82, 1975. doi: 10.1159/000149883.

[48] J. H. University and Medicine. Covid-19 map. URL https://coronavirus.jhu.edu/map.html.

[49] D. N. Vinod and S. Prabaharan. Data science and the role of artificial intelligence in achieving the fast diagnosis of covid-19. *Chaos, Solitons Fractals*, 140:110182, Nov 2020. ISSN 09600779. doi: 10.1016/j.chaos.2020.110182.