Team D

Literature Review

---

1) Satu, M.S.; Howlader, K.C.; Mahmud, M.; Kaiser, M.S.; Shariful Islam, S.M.; Quinn, J.M.W.; Alyami, S.A.; Moni, M.A. Short-Term Prediction of COVID-19 Cases Using Machine Learning Models. Appl. Sci. 2021, 11, 4266. https://doi.org/ 10.3390/app11094266

Short-Term Prediction of COVID-19 Cases Using Machine Learning Models

The paper discusses the case in its infancy for Bangladesh, where limited awareness and various socio-economic factors, combined with very high population density can lead to explosive growth. Such growth is very different from global trends, and should be recognized and evaluated as such, for eg. By taking factors like population density, GDP, HDI, portion of population in urban areas, etc into account.

In the paper, we see models such as Naive-Bayes, Support Vector Machines, Linear Regression, Decision Trees, LSTM, and Random Forest applied to data from various countries. Many advanced models such as LASSO, RNN, GRU, and VAE also see use, to compare the effectiveness of the modelling.

Comparing the result from various models, and pose the hypothesis that short-term models may be more capable of modelling the virus' growth and spread in B'desh, as opposed to more generalized predictors for long-term evaluation. The team makes the model for a 7 day (week-long) period, which is neither so short that prediction, even accurate, may not be useful; nor a period so long that lookahead planning may be futile due to various unpredictable conditions.

In further clarification, "[. . .], it is important to note that predictions of infection levels are sensitive to non-linear changes of parameters so that long term prediction tends to give poor results. For this reason, we have focused on implementing short-term forecasting models where accuracy is more likely to be achieved."

The paper uses algorithms and models which have successfully predicted curves for numerous other SARS based viruses, or different epidemics such as Dengue, Swine Flu, etc. Furthermore, it also explores the possibilities of using cloud-based services to exploit the power of large-scale processing of data.

Various metrics for evaluating the results of various algorithms have been used, in which the authors clarify that RMSE is the most widely used and recognized, due to effectiveness in analyzing individual regression models. Metrics such as MAE and R-Squared can also be utilized, which give non-directional average error, and the degree of relation between dependent and independent variables.

Tests were conducted over 35 different periods of rounds and the results were fairly consistent across the rounds. The results showed that Prophet algorithm made incredibly consistent and good predictions across the majority of the rounds, with some difference in the earlier tests due to tuning of growth factors going on.

In the initial phase of the tuning, Linear Regression showed the best prediction versus the actual statistics recorded during the same period of 9 March 2020 through 9 April 2020. Due to the nature of Linear Regression, there were obvious unconformities owing to the straight line nature of the curve.

In the second round of testing data, SVR showed the best results, but the predictions continued to be slightly off from actual. In this case, they were on average higher than the actual rate owing to higher tuned growth rate. The 3rd algorithm onwards, the prediction is almost spot on, with the PR algorithm and Poly-MLR giving the best results in round 3 and 4 of the iterations.

Since the 5th iteration upto the 34th, Prophet algorithm gives the best results and confirms correctly to the real world data.

The paper further discusses how various models have been made to predict the cases within B'desh with some success, but that none of them had done such a comprehensive test with applying over 10+ different algorithms on such a scale and perform various testing metrics to get the correct predictions.

The paper states that such models can help various organizations minimize the losses from Covid through the application of Machine Learning on a cloud-based processing service.

Usherwood, T., LaJoie, Z. & Srivastava, V. A model and predictions for COVID-19 considering population behavior and vaccination. Sci Rep 11, 12051 (2021). https://doi.org/10.1038/s41598-021-91514-7

A model and predictions for COVID-19 considering population behaviour and vaccination

A whole host of Covid-related papers have been published in the past year or two, with a majority of them dealing with modeling the spread, including models involving segmentation and compartmentalizing of infected populace into sections of Susceptible, Infected, and Recovered (SIR Models).

Furthermore, models have also been made by taking into account SIR models across age ranges, to optimally distribute vaccinations to facilitate lowest infected and death counts, including vaccine effectiveness across age ranges and such.

Contrary to most others, this paper introduces functions to classify ‘levels of caution’ and ‘sense of safety’, culminating in a time dependant infectious disease transmission rate function.

Function : $\beta = \beta 0\ fI\ fV$

$fI = e^{(-dI\ I)}$

$fV = 1/fI + (1 - 1/fI)^{(e-dV\ V)}$

The Level of Caution function is related to the various countermeasures and precautions taken by various people in the populace and how common they are, as a measure of modeling the degree of prevention. It includes things such as social distancing, personal protective hygiene, and government schemes (eg. Covid-0 policy in China).

The input factor depends on various states of the population which change throughout a period of time such as awareness, fatigue caused by the pandemic, changes in seasons which may affect daily routines, and changes in government policies such as mandatory lockdowns or not.

The level of caution is inversely related to e to the power of the above factor, so we can see that as the population is affected with more and more of the above changes in state (factor approaches 1), the level of caution correspondingly approaches 0, showing that there is closer to 0 chance of transmission between any two persons.

As the factor responsible for the Sense of Safety function approaches 0 (no drop in safety precautions despite high sense of safety; i.e, increasing vaccinations), the Sense of Safety function approaches a value of 1, signalling that safety measures continue to be adopted appropriately.

If the populace feels that higher vaccination can allow for worse safety measures, then the Sense of Safety value neutralizes the Level of Caution, leading to no change in transmission rate from the base transmission rate.

The base transmission rate is calculated for each location based on various standards and taking into account a multitude of factors such as population density, pre-pandemic behaviour, and stipulated contact rates. As such, the paper finds that bustling metropolises like New York have a higher base transmission rates compared to more rural areas such as South Dakota.

The model, after careful calculation of the various responsible factors, gives a terrifyingly accurate prediction of cases per day. The model shows that the behaviour of the population in the sample states can be clearly checked off as High Caution and Low Sense of Safety, an ideal state to prevent greater spread of the disease.

Further classifying the population's response to the rollout of vaccinations, the paper classifies that the population shows no changes in its state as a result of the rollout of the vaccination, preserving

the high alertness level, as opposed to having falling precautions and alertness with the rollout of the vaccines.

As opposed to the previous paper we discussed, this follows a SIRDV model, for Susceptible, Infectious, Recovered, Deceased, and Vaccinated; so to fit their model well.

Muhammad, L.J., Algehyne, E.A., Usman, S.S. et al. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN COMPUT. SCI. 2, 11 (2021). https://doi.org/10.1007/s42979-020-00394-7

Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset

Contrary to our previous discussed works, the current paper accounts for predicting whether an individual is infected or not based on symptoms experienced and various different categorization techniques, whereas the previous papers focussed on predicting the overall trend across a subset of the populations with machine learning concepts.

The dataset which was being worked on was rehashed to only contain the Age and Sex, along with eight other indicators, the likes of which included Diabetes, Pneumonia, Tobacco consumers, etc.

The authors continue on to apply specific supervised learning algorithms such as naive bayes, Logistic Regression, and Decision Tree. The authors also use Artificial Neural Networks, and analyze the Coefficient of Correlation between the various independent variables (Age, Sex, Pneumonia, Diabetes, Tobacco consumption, etc) and the dependent variable (Positive RT-PCR test result for Covid-19 in the tested patients).

The paper explains the concepts behind the various methodologies used. The authors showcase the Bayes theorem of probability and how it related to classification based Machine Learning Algorithms.

Furthermore, Logistic Regression application and formula is explained; how the ever-versatile Decision Tree algorithm, suitable for all kinds of data, works and how to de-noise the dataset for it; a representation of SVM is also given, explaining how a hyperplane is built to maximize distance between classes.

A brief look on the complex topic of ANNs is also given, touching on how bias and transfers between layers simulate the connection of neurons inside the brain. Lastly, the analysis of regression coefficient can show the degree of relatedness amongst the various independent decider

variables and also their relation to the actual output variable, the dependent variable that is the frequency of occurrence of Covid-19 cases amongst the population.

Post analysis, the paper shows that their results have indicated a weak positive correlation between any independent variable and a dependent variable.

They further show the Specificity (True Negative Rate), Sensitivity (True Positive Rate), and the overall Accuracy of the various methods applied.

The 3 base algorithms of Decision Tree, Linear Regression, and Naive Bayes showed close to a 95% accuracy with ~85% sensitivity, ~90% specificity. Whereas, Artificial Neural Network and Support Vector Machine show closer to 90% accuracy, with ~92% sensitivity and ~80% specificity.

Rather than using extremely advanced concepts or having revolutionary results, it shows the very foundation of machine learning and how it can be applied to day-to-day tasks. Compared to the previous two papers discussed, we can also gain some life lessons such as to contribute to a larger cause, one does not need to make headways and can also do what they are able to do.

Malki, Zohair, et al. "The COVID-19 pandemic: prediction study based on machine learning models." Environmental science and pollution research 28.30 (2021): 40496-40506.

The COVID-19 pandemic: prediction study based on machine learning models

This paper used a machine learning approach to predict the spread of the virus in several selected countries. However, the nature of the virus is more or less the same everywhere, so the same approach can be applied to predict the spread of COVID-19 infections in other countries. A machine learning model is presented using statistical visualization graphs to better predict the spread of the COVID-19 pandemic.

They used the quality and density of the WHO data collected to determine the predictive value of the technique. Data used in this study were collected from official data repositories such as the official website of Johns Hopkins University, WHO and Worldometer. The data show the daily total number of confirmed COVID-19 positive cases, daily and total deaths, and total and daily recoveries.

This study is primarily based on a decision tree algorithm for global realtime data of COVID-19. The core idea is to use supervised machine learning algorithms for time series forecasting. The algorithms proposed for this work, namely decision tree algorithms and linear regression, are powerful models for predicting problems related to sequence and time series data. Using machine

learning, they developed a predictive model using available data from COVID-19 found on a well-known data storage website.

WHO uses historical data to represent the proportion of confirmed cases, mortality, recovery and growth rates. Formulas were developed to calculate confirmed rates of change and mortality, patient recovery rates, and pandemic growth rates. They proposed methods predicted possible confirmed cases in the next seven days in the United States. Experimental results showed an exponential increase in confirmed cases from a few hundred thousand to nearly 2.5 million.

So the prediction was found to be suboptimal. This is a difficult problem for training deep learning models on small datasets. Given the distinction between predicted and actual values, the results were fairly accurate. To validate the performance of the proposed method, they used the root mean square error for each of his three attributes: confirmed cases, recovery, and death. They showed the final predictions of the proposed model for all attributes. They conducted a comparative study using state-of-art methods.

The proposed model was compared with various state-of-the-art models (random forest, ARIMA, deep learning) and the accuracy of the machine learning model on the training data set was measured by root mean square error (RMSE) and mean absolute error (MAE). . The machine learning models were developed to predict an estimate of the spread of his COVID-19 infection in many countries and how long the virus could be expected to be stopped thereafter. Their findings predicted a sharp decline in his COVID-19 infections worldwide in the first week of September 2021.

Rahimi, Iman, Fang Chen, and Amir H. Gandomi. "A review on COVID-19 forecasting models." Neural Computing and Applications (2021): 1-11.

A review on COVID-19 forecasting models

This paper presents an overview and brief analysis of the leading machine learning predictive models for COVID-19. The work presented in this study consists of two parts. In the first section, a detailed Scientometric analysis presents an influential tool for literature analysis performed on his COVID-19 data from the Scopus and Web of Science databases. Keywords and topics are covered in the above analysis, and later in the work, classification of machine learning prediction models, criteria evaluation, and comparison of solution approaches are described.

Putra and Khozin Mu'tamar used a particle swarm optimization (PSO) algorithm to estimate the parameters of a susceptibility, infection and recovery (SIR) model. The results show that the proposed method is accurate and has sufficiently small error compared with other analytical methods. Mbuvha and Marwala [2] adjusted her SIR model for reported cases in South Africa after considering different reproduction number (R0) scenarios for reporting transmission and medical resource estimates. . Qi and Xiao suggest that both daily temperature and relative humidity can

affect COVID-19 outbreaks in Hubei and other provinces. Salgotra and Gandomi developed his two his COVID-19 prediction models based on genetic programming and applied these models to India. The results of a study by show that the genetic evolutionary programming model of COVID-19 cases in India has proven to be highly reliable.

Mahalle and Kalamkar used WHO and social media communications as datasets to classify predictive models as mathematical models and machine learning techniques. Important parameters such as mortality, measurement parameters, quarantine duration, medical resources, and mobility were also examined.

Naudé outlined the contribution of artificial intelligence (AI) to COVID-19. Some areas of AI contributing to COVID-19 have been identified as early warning and alerting, tracking and prediction, data dashboards, diagnosis and prognosis, treatment and cure, and social control.

The paper presented a keyword-based network visualization, and the top keywords researchers focused on, including coronavirus, forecast, epidemic, humans, statistical analysis, quarantine, hospitalization, mortality, and weather.

A detailed analysis of the total number of works cited and the number of records vs. affiliation is also presented.

This paper aims to review the main predictive models for COVID-19 and provides a brief analysis of the published literature. This article used keyword analysis to highlight the most important subject areas. In addition, several criteria were identified that will help researchers in their future work. The paper also acknowledges the most useful models researchers have used to predict this pandemic.

Additionally, this paper will help researchers identify key gaps in the research field and develop new machine learning models for predicting COVID-19 cases. A detailed scientometric analysis has been conducted as an influential tool for use in bibliographic analysis and review.

Xiang, Yue, et al. "COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models." Infectious Disease Modelling 6 (2021): 324-342.

COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models

In light of the COVID-19 global pandemic, governments around the world rely on mathematical predictions to support decision-making during the epidemic. Extrapolation of epidemic impacts from early stages of growth is affected by model, data, and behavioral uncertainties. Although the complexity and methodologies vary, simple models can still be used as a reference during the early growth stages of the epidemic and provide the basis for more complex models of infection to understand the course of disease development.

George E.P. “All models are wrong, but some are useful,” says British statistician Box. We need to work closely together to identify accurate epidemiological data. It helps to accurately assess the impact of time-related variables and include them in different models. At the same time, in combination with monitoring results from serological studies, it helps modeling researchers to accurately calibrate epidemiological models based on real-world situations.

On the other hand, for differences in predictions between different public health strategies, the most important impact was travel restrictions. Various studies have been conducted on the impact of contact tracing and social isolation, but it has been argued that improved quarantine and reporting rates, as well as the wearing of face masks, are essential for epidemic prevention and control.

Various countries should take appropriate public health measures to control the progression of the epidemic. In particular, there is evidence that recessive infections are non- or poorly contagious, suggesting that existing COVID-19 epidemiological models may overestimate epidemic risk. The input epidemiological parameters of the prediction models show significant differences in predicting the severity of epidemic spread.

Therefore, preventive control organizations should be cautious when basing public health strategies on the predictive outcomes of mathematical models.

Chu, Xu, et al. "Data cleaning: Overview and emerging challenges." Proceedings of the 2016 international conference on management of data. 2016.

Data Cleaning: Overview and Emerging Challenges

Industry and academics have recently shown a surge in interest in several facets of data cleansing, including innovative methods. These have included techniques such as crowdsourcing, abstractions, and new approaches towards scalability. Xu Chu et. al focus on not only looking into these innovative approaches but also give a breakdown into good practices, and different approaches.

The authors outline and evaluate two statistical approaches to cleaning up qualitative data, where methods either evaluate how cleaning will affect either upcoming numerical queries or employ machine learning techniques to increase accuracy, or efficiency. They place these methods within a single, comprehensive listing of data cleansing, enabling the demonstration of how numerous qualitative methods can lend themselves to such statistical analysis.

They move on to provide the full overview, which includes a breakdown of various aspects of cleaning such as error detection based on What, Where, How; repairing errors with the same What, Where, How. What lends itself to targetting and types, Where towards Automation, and How towards intelligence and modelling.

In the next section the authors introduce to the readers what data cleaning means from the perspective of statistics. Prominent problems like deduplication, repairing missing values, and presence of incorrect values is discussed. Furthermore, they also touch upon how to remove irrelevant and erroneous data.

The paper showcases its real value here as it provides a trove of detailed papers which check through and investigate each aspect of such errors in detail.

Next, they move on to discussing emerging challenges in cleaning, such as livestreamed data and collection for edge computing, privacy concerns with increasing realization of the extent of data collection in the world, and problems which may arrive with file formats such as JSON which contain semi-structured, or text documents which contain unstructured data.

Austin, Peter C., et al. "Missing data in clinical research: a tutorial on multiple imputation." Canadian Journal of Cardiology 37.9 (2021): 1322-1331.

Missing Data in Clinical Research: A Tutorial on Multiple Imputation

Missing data refers to the variables which may be absent from the data being used for our purposes. This paper goes into detail about how we can utilize Multiple Imputation properly using R packages such as MICE.

It illustrates what Multiple Imputation is: unlike what we first thought it is not imputing multiple variables at once, but rather imputing multiple values for a single missing variable based on multivariate analysis.

They first describe the various ways that data can be missing -> Missing at Random, Missing Completely at Random, Missing Not at Random. They discuss what each means and what impact it can have on the imputation methods we can use.

The authors go into detail about the various methods which have been used for dealing with missing data. They start with “complete case” analysis, which makes it so that only the rows which have no missing values are used. They discuss that bias can occur when the data is not missing at random.

Next, mean-value imputation is discussed, and the dangers associated with it. For example, how it artificially brings down the standard deviation in the data, or how it will completely ignore multivariate interactions, eg. mean value imputation of blood pressure can ignore characteristics such as height, age, etc.

The natural successor to this is the regression-fit based approach of imputing values, which is better than the previous methods in regards to how the imputation affects the dataset, but still have to contend with the effect of assigning one particular final value to a truthfully missing cell.

Multiple Imputation is discussed next, and about how it is essentially regression-fit based imputation but with multiple different values for each missing cell. This makes it so that we have many different “full” datasets after the Multiple imputation is complete, and this represents the true uncertain nature of the missing data, which was inherently missing itself in the previous imputation methods.

Infact, the popular MICE package for R stands for Multiple Imputation with Chain Equations. It furthermore includes various methods such as predictive mean-matching, logistic regression, Bayesian regression, and odds based approaches.

The paper ends with a comparison between the analysis of the dataset with various different imputation methods discussed throughout, and how the resultant analysis is affected.

[Villavicencio, C.N.; Macrohon, J.J.E.; Inbaraj, X.A.; Jeng, J.-H.; Hsieh, J.-G. COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. Algorithms 2021, 14, 201. https://doi.org/10.3390/a14070201]

COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA

Early diagnosis is essential to prevent the development of a disease that can cause danger on human lives. COVID-19, which is an infectious disease that has mutated into several variants, has become a global pandemic that requires it to be diagnosed as soon as possible. Using technology, the available information about COVID-19 is increasing every day useful information from massive data can be extracted using data mining. Many models have been proposed which can predict the presence of coronavirus in a person and one such published paper consists of some models predicted by some machine learning algorithms and comparing which one of those algorithms produces the best model, the one having less bias and less variance

It’s not always possible to avoid meeting other fellow people and we being inter-dependable upon one other have to meet and interact for our survival, which brings us to the wearing of masks by everyone which can slowdown transmission, but not all people can wear masks and masks are also not that effiecient to prevent transmission as social distancing, which brings us to the idea of separating only some group of individuals from a given environment who may be a carrier (Not all people are carriers, the virus must be present in one’s body to a significant extent for a person to become a carrier). For this we have to identify those individuals who are covid positive (who have this virus in their body) where the prediction models come handy.

The paper first gives the statistics of covid- 19, the countries it has affected maximally, number of cases of recovery, number of cases, number of deaths, etc. Then it goes on to explain the symptoms of the disease, how the virus spreads, the parts of the body the disease affects, common prevention measures of covid- 19, how humanity is going through a failed attempt to control this disease as the number of cases everyday has been rising despite thorough human intervention through quarantine, vaccination and other measures. Then it says how prediction models help us in preventing the disease to a great extent.

The machine learning algorithms J48 Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours and Naïve Bayes algorithms were applied through WEKA machine learning software. Each model was evaluated using 10 fold cross validation and compared according to major accuracy measures, correctly or incorrectly classified instances, kappa, mean absolute error, and time taken to build the model. The results show that Support Vector Machine using Pearson VII universal kernel outweighs other algorithms by attaining 98.81% accuracy and a mean absolute error of 0.012.

There are other algorithms, eg deep learning algorithm called encoder which uses original patient features to get reconstructed patient features, which the algorithm itself generates and is different from the original patient features, which predict if a person having those particular features is covid-19 susceptible or not. Although these provide accurate results, the accuracy of SVM which is 98.81% for this data which is more than enough, also that the cost of implementation of encoders are also very high, it is more plausible to with SVM’s taking cost factors as well into consideration. In [4] a different dataset is used and it turns out that the naive- bayes classification turns out to be the best algorithm for the dataset, it actually depends which algorithm suits which dataset, one algorithm may produce the highest accuracy in one dataset while another may prove to be the best dataset for another algorithm.
