COVID-19 Data Analysis and Predictions

Abstract:

"History repeats but science reverberates", Siddhartha Mukherjee. If history teaches us anything, it's that while pandemics may start small, their impacts can be as disastrous as wars or natural disasters. The difference today is that science gives us the ability to detect pandemics right at the very beginning and to take actions to mitigate their impacts before they spread too widely [1]. Over the past centuries, Humans have endured pandemics like Cholera and Spanish flu. Yet another contagious disease has taken over the world, The Chinese Coronavirus. Since December 2019, millions of patients have been diagnosed with the virus. The future is yet unknown. Predictions for the next couple of days are made by data scientists throughout the world so that the requirements for the necessary resources to deal with the outbreak are fulfilled in time. In this paper, we demonstrate the analysis of the trends in the COVID-19 datasets (global and local, both). We have applied techniques to train the supervised machine learning models effectively and predict the number of cases that are expected shortly. We found that rate of recoveries is faster than that of deaths and countries whose governments and communities have chosen to implement precautions strictly have halted the spread among their people.

1. Introduction:

Coronavirus disease (COVID-19) is an illness caused by a novel coronavirus now called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was first identified as a respiratory illness case in Wuhan City, China. It was initially reported to the WHO on the 31st of December, 2019. On January 30, 2020, the WHO declared the COVID-19 outbreak as a global health emergency and then later declared it as a global pandemic, its first such designation since declaring H1N1 influenza a pandemic in 2009. The illness caused by SARS-CoV-2 was termed COVID-19 by the WHO. The name COVID-19 was chosen to avoid stigmatizing the virus's origins in terms of populations, geography, or animal associations.

2. Background:

We live in an increasingly globalized world. Through jet travel, people and the diseases they carry can be anywhere in the world in a matter of hours. This respiratory virus isn't an outlier, it is the part of our interconnected viral village. It is a highly contagious virus and one sneeze is all it takes to spread throughout the community. By observing the rise in its spread, the World Health Organization (WHO) has declared it a pandemic [2], meaning that it's spreading worldwide.

Epidemics and Pandemics come in many shapes and forms. For instance, in 2010, a devastating earthquake hit Haiti, forcing thousands of people into refugee camps. Within weeks, the campers were infected

with cholera, a bacteria spread by contaminated water, which introduced a country-wide epidemic. Pandemics have occurred throughout the human history. However, the greatest pandemic that has affected the most is influenza. The first ever pandemic was observed in 1580. The 18th and 19th centuries saw at least 6 pandemics. In terms of mortality, the Great Flu pandemic of 1918 cannot be beaten. A death toll of 50 million people was observed worldwide.

A deadly contagious virus has hit the human race yet again. Coronavirus caused an epidemic of severe acute respiratory syndrome in China. In December 2019, the first case of this severely acute respiratory syndrome was observed in China and it then proliferated faster than public health measures could contain it. It became an international epidemic. Millions of people got affected. Out of the million patients, some couldn't survive, some have recovered and some are still fighting for their lives.

Fig 1: The world-map representing the total number of corona virus cases using Circle Markers.

3. Methods and Materials:

3.1. Libraries Used:

- For the analysis of the world-wide data, we have coded in Python 3.0 by using Anaconda Navigator 3 in Jupyter Notebook to process the data.
- The predominant libraries are matplotlib that is responsible for creating a figure concerning data provided in its function and other legends and labels to make it more readable **[3]**;
- Library seaborn that possesses a high-level interface for visualization of COVID'19 outbreak and its statistical information;
- Library sklearn which features multiple predictions models including SVM model by using SVR that is used to predict future peaks in our analyses.
- Library mpl_toolkits.mplot3d is used for 3d animation of the model.
- Datetime and time libraries are added to rearrange the date format.

3.2. Setting the Research Goal:

Provided with the abundant data of COVID'19 affected countries, we are determined to estimate the outbreak of this contagious virus for the next 10 days based on the reported data. Keeping in view that the official confirmed case number has surpassed all the previous predictions **[4]**, we would be applying two prediction models i.e. Linear Regression and SVM model on Global data and Fbprophet model on Local dataset to compare the results and to provide the mortality rate.

3.3. Retrieving Data:

Four datasets are used in the analyses and visualization of COVID'19 pandemic from the world's largest data science source, Kaggle to demonstrate the outbreak worldwide.

Global Dataset: The three global dataset possess the comprehensive information of country/region, province/state, longitude and latitude with their respective numbers of cases, deaths, and recoveries from all regions of the world ranging from 1/22/20 to 4/22/20. The covid19 confirmed global, deaths global, and recoveries global datasets are organized with rows representing the entries from different countries and the columns depict the date in which cases are registered accordingly.

Local Dataset: The local dataset demonstrates the date wise entries of confirmed cases, deaths and recoveries from different regions within Pakistan. It also shows the travel history, province and city name to illustrate the reason behind the high peaks of confirmed cases which is mainly due to local social contact and pilgrims from Tehran.

3.4. Data Preparation:

- Date and Time format is rearranged to bring uniformity.
- Null values from the datasets are removed using isnull() and remove() functions.
- In our analysis of the local dataset, the data is grouped with respect to date, city and province to illustrate

the number of confirmed cases, recoveries and deaths.

- Unique values are extracted on the basis of countries and dates to have a clear visualization of total cases, deaths and recoveries.
- Active cases from the local dataset are calculate by using the formula:

  Total active = Cases - (Deaths + Recovered)

3.5. Data Exploration:

Univariate, Bivariate, and Multivariate analysis are done to find the correlation among features in local dataset using heat maps. Additionally, a 3d model is built to demonstrate the number of cases, death, and recoveries within Pakistan. A useful of FacetGrid is implemented to exhibit the effect of travel history in the Covid'19 cases, carried out the function of sns.pairplot to get the histogram and scatterplot of individual integer type feature in the set. In the global dataset, illustrated the top 10 affected countries from this COVID'19 by country-wise summation, grouped by dates using plt.barplots. Also, a pie chart is built to exemplify the percentage of patients in each country.

3.6. Data Modeling:

For each of the city/province for active case visualization within Pakistan. To predict the outbreak for future, following models are used.

Support Vector Machine:

An algorithm for two-group classification problems which is categorized under the supervised machine learning approach. It is known as a Support Vector Machine. The model is used for categorizing and predicting new data after training the SVM model with sets of labeled data [5].

Features:

- This model works well with limited amount of data (thousands).
- Takes in data points and gives out a hyper-plane that separates the classes i.e. it draws a decision boundary between two sets of data.
- It treats linear and non-linear data separately.
- A simple linearly separable data works best in a 2D plane.
- For non-linear data, it uses the kernel trick to differentiate among the data tags.

The kernel trick:

SVM model predictions include the conversion of data into multi-dimensions for fair classification. Every step of transforming the data from a lower dimension to a higher one involves multiple steps of complicated calculation [6]. It becomes expensive when every vector of the dataset has to be transformed.

The kernel trick provides a cheaper solution and implements the dot product technique. The kernel function takes to input the data and transform it in the required form. Kernel functions are of different types:

- Linear

- Polynomial
- Radial-Bias function
- Sigmoid
- Non-Linear

Linear Regression Model:

It involves the fitting of linear equations to data and also models the relationship between two variables. It is not necessary for the variables to be dependent on each other but some significant association has to be there [7].

Linear regression line equation:

$Y = a + bX$

Where, Y = explanatory variable, X = dependent variable. The slope of line is b and a is the intercept.

Features:

- It is commonly used for predictive analysis.
- The predicted output values are continuous and form a slope.
- It offers extrapolation technique when the prediction has to be made outside the range of actual data.

Fbprophet:

It is a machine learning algorithm for time series with an input of dataframe with 2 columns named 'ds' for dates and 'y' for values against them, specifically designed by the Facebook data scientist group **[8]**. It is a predictive model, frequently used with nonlinear trends, and possesses the ability to handle outliers with seasonal effects. It provides a practicable forecast with minimal effort.

Prophet.make_future_dataframe function is used to future forecasting. This model is not encapsulated `within a built-in library in Anaconda, rather needs to be installed explicitly using PyPI.

Features:

- Fast and Accurate
- Robust to missing data.
- Automatic, requires less efforts to be implemented

To predict the outbreak for the next 10 days, we have used SVM and Linear Regression model with test_split ratio of 0.15 and shuffle being False as the trend grows exponentially. Below are the best suitable values for SVM features.

- kernel = ['poly', 'sigmoid', 'rbf']
- c = [0.01, 0.1, 1, 10]
- gamma = [0.01, 0.1, 1]
- epsilon = [0.01, 0.1, 1]
- shrinking= [True, False]

SVM runs the model and chooses the best parameters out of those provided according to the data it gets. Larger C tends to give an optimization result. Both values for shrinking is provided to reduce the kernel dimensions depending upon the model. Svm_estimator plays a major role in this infectious disease forecast as it returns the highest score of accuracy. We trained the model using the original dates reported in the dataset and passed the next 10 days to the model to predict the number of cases for the future. Similarly, we fit the same x_train and y_train_confirmed n the linear regression model with the

same splitting ratio to get the linear prediction of the future for the next 10 days to have a comparative analysis.

In the local dataset, we have executed fbprophet model to predict the trend of this outbreak in the next 10 days. Dates from the dataset are considered as ‘ds’ and the total number of cases on each day is the ‘y’ in this model.

Initially, Pystan library is installed to download fbprophet library to get the exclusive functions of fit and Prophet.make_future_dataframe. The advantage of this model over others is that it predicts the range I.e. yhat_upper and yhat_lower rather than a number.

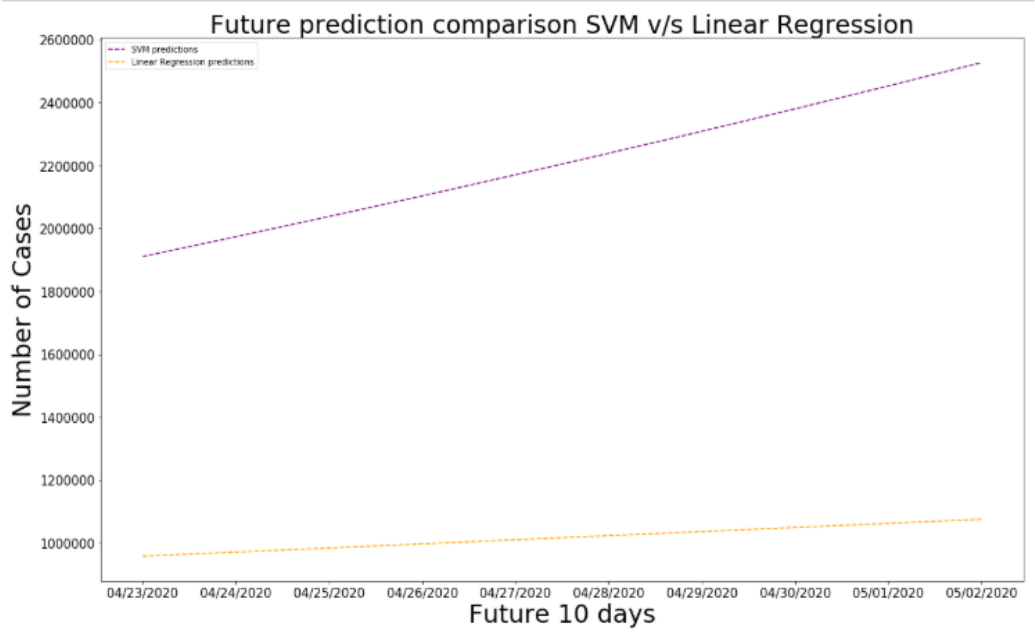

Fig 2 depicts the comparison between the two models predictions on Global dataset.

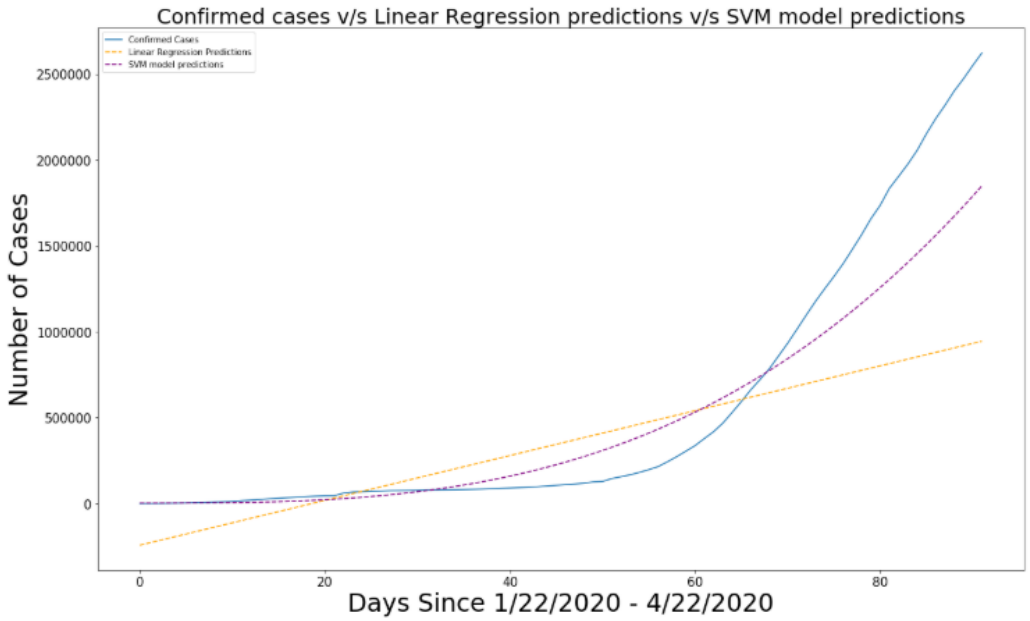

Fig 3 depicts the comparison between the Confirmed cases, linear regression model predictions and SVM predictions.

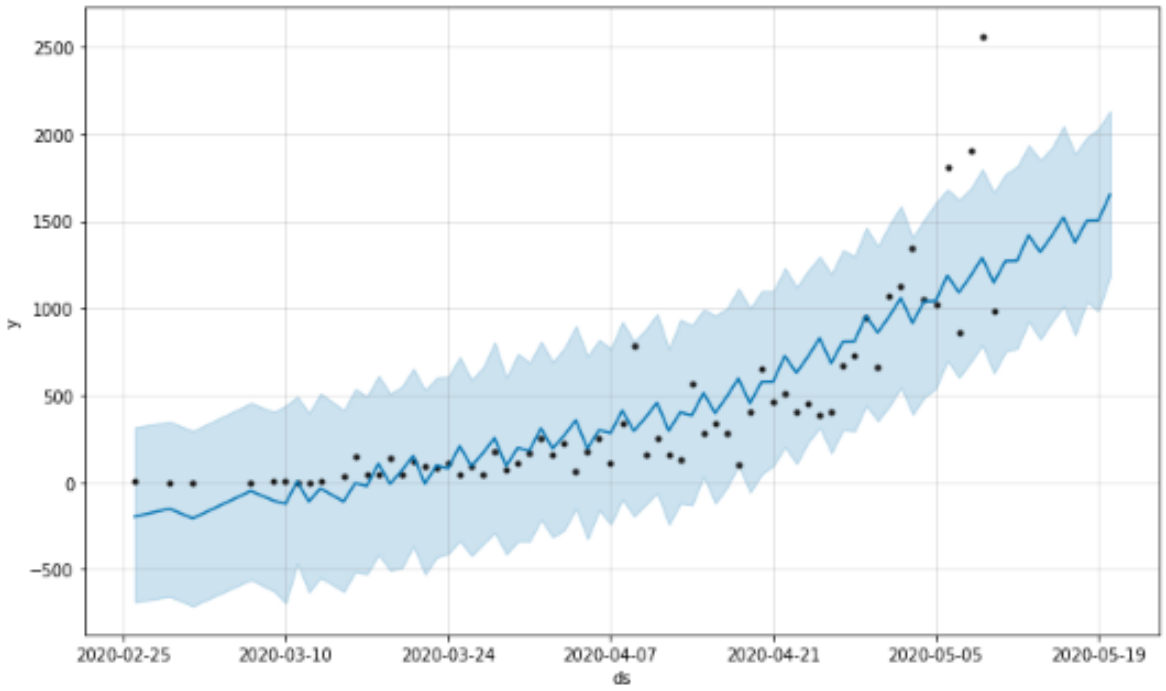

Fig 4 depicts the predictions of total cases in Pakistan ranging from February 25th, 2020 to May 10th, 2020 plus the future 10 days using the Fbprophet model.

4. Data and Results:

The fetched datasets had data from all over the globe from January 22nd, 2020 to April 22nd, 2020. A future prediction was made using two supervised machine learning models. Both the models predicted the total number of cases in the next 10 days i.e. from April 23rd, 2020 to May 2nd, 2020. According to Fig 2, SVM model predictions are way much larger than linear regression model.

Predictions of data from January 22nd, 2020 to April 22nd, 2020 were also made and comparison among the three i.e. Actual data, Linear regression predictions and SVM model predictions was also visualized to see the differences. In Fig 3**,** the difference among the three kinds of predictions can clearly be seen.

According to Fig 4, it can clearly be observed that the prediction is quite accurate and the fbprophet model has fit perfectly on local dataset

5. After effects of COVID 19 on economy:

The impacts of COVID 19 have led to Governments around the world preparing panic plans, and aid packages to sustain their economies [9]. All in all, global supply chains and organizations have stopped working, affecting the economic condition of communities across the globe. Millions of people could lose their jobs over the coming months. Every day new companies keep shutting down operations, revising estimates, or announcing layoffs, adding more tension to the already uncertain and declining economic situation. Consumers have also changed their consumption patterns, resulting in shortages of many goods in supermarkets around the world. Uncoordinated governmental responses in many countries, and unplanned and sudden lockdowns, have led to a disruption in the supply chain [10]. The travel restrictions imposed by governments around the world subsequently led to the reduction in the demand for all forms of travel leading to the tourism industry losing over $200 billion globally [11]. Stock markets collapsed in March 2020. Most stock indices around the world have registered their biggest one-day falls on record. In conclusion, COVID-19 has affected businesses and organizations globally, both directly and indirectly have an effect on the financial markets and the global economy.

6. Future Aspects of COVID in ML

COVID-19 has now taken over more than half of the world. All the scientists and researchers are putting most of their efforts in searching for its cure, meanwhile, it is likely impossible to have it done without Machine Learning and Artificial Intelligence. Machine learning (ML), the present type of AI, works by recognizing designs in chronicled training information [12]. Currently, Deep Learning is used to predict the structure of proteins and their interactions with chemical compounds to facilitate new antiviral drugs or vaccines. We use a deep learning-based method, DFCNN [13]. The present solutions cost highly and are inefficient due to the rapidly increasing cases. Future work on creating, hosting, and benchmarking COVID-19 related datasets is essential because it will help to accelerate discoveries useful for tackling the disease. Repositories for this goal should be

created following standardized protocols and allow researchers and scientists across the world to contribute to and utilize them freely for research purposes [14]. Among the published works, the use of AI deep learning techniques for COVID-19 diagnosis based on radiology imaging data appears to be dominant. Accordingly, there is a demand for future work on developing a benchmark framework to evaluate and compare the existing methods [15].

7. Conclusion:

Covid-19 is a pandemic that has affected the whole world in different ways. Countries are facing serious economic and health crises [16]. People are dying due to unemployment and malnutrition. Hospitals are out of resources and the contagious virus is not slowing down the spread. But the good news is that every nation is trying to stop the spread by social distancing and obeying the laws imposed by the government. By the Grace of God, many countries have taken control of the situation but some are still struggling.

- Summary of the findings:

  The information and technology sector is playing its part in making new machines and resources for the help of people. Data scientists are busy analyzing the trends and forecasting the future. In this paper, a study of global and local data sets has been displayed. Also, a prediction of the future 10 days has also been made using supervised machine learning models.

  The data of confirmed cases has been trained on two different models i.e. Support Vector machine model and linear regression model. It has been observed that the prediction made by two models vary too much from the actual data. A huge rise was observed in the SVM predictions and the linear regression predictions were much lower than the actual data.

- Limitations of the project:

  The prediction was made only of the future 10 days due to less amount of training data.

  The results of the linear regression model were flawed and not much could be deduced out of them.

8. References

[1] L. L. K. J. Y. &. G. X. Jia, "Prediction and analysis of Coronavirus Disease 2019," *arXiv preprint arXiv:2003.05447.*, 2020.

[2] C. A. Z. O. N. K. M. K. A. A.-J. A. .. &. A. R. Sohrabi, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19).," *International Journal of Surgery.*, 2020.

[3] W. McKinney, "Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.," *O'Reilly Media, Inc.*, 2012.

[4] C. R. L. T. A. &. S. C. Anastassopoulou, "Data-based analysis, modelling and forecasting of the COVID-19 outbreak," *PloS one*, vol. 3, no. 15, 2020.

[5] G. F. &. J. E. M. Smits, "Improved SVM regression using mixtures of kernels.," *In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02*, vol. 3, no. 02CH37290, pp. 2785-2790, 2002.

[6] X. M. A. C. J. &. R. K. Song,

"Comparison of machine learning techniques with classical statistical models in predicting health outcomes.," *In Medinfo,* pp. 736-740, 2004.

[7] G. A. &. L. A. J. Seber, Linear regression analysis, John Wiley & Sons, 329.

[8] T. T. Mengistie, "COVID-19 Outbreak Data Analysis and Prediction Modeling Using Data Mining Technique.," *International Journal of Computer,* vol. 1, pp. 37-60, 2020.

[9] N. Fernandes, "Economic Effects of the Coronavirus Outbreak (COVID_19) on the World Economy.," 2020.

[10] "The socio-economic implications of the coronavirus pandemic (COVID_19)".

[11] P. K. A. T. Ozili, "Impact on the Global Economy," *Spillover of COVID-19,* 27 March 2020.

[12] R. C. J. &. H. Sujath, "A.E. A machine learning forecasting model for COVID-19 pandemic in India," *Stoch Environ Res Risk Assess,* 2020.

[13] L. L. C. Y. e. a. Zhang H, "IVS2vec: A Tool of Inverse Virtual Screening Based on word2vec and Deep Learning Techniques," *Methods,* 2019.

[14] L. Q. X. Z. Y. Y. W. K. B. C. Li, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT.," *Radiology,* 2020.

[15] T. t. Nguyen, "Artificial Intelligence in the Battle against Coronavirus (COVID-19)," *A Survey and Future Research Directions.*.

[16] C. K. T. d. M. C. P. M. L. I. D. A. S. d. O. N. J. V. A. S. J. S. d. S. R. I. .. &. N. M. L. R. Lima, "The emotional impact of Coronavirus 2019," *Psychiatry research,* no. 112915, 2020.