A REPORT

ON

Comparative Study of Machine Learning Models for Forecasting the Spread of Covid-19

BY

Name of the Student	I.D. No.s
VEDANT BORDIA	2018A4PS0088G
TUSHAR MALOO	2018A4PS0848G
AYUSH KUMAR	2018A8PS0823G
ABHIVANDAN GANDHI	2018A8PS0831G

Prepared in partial fulfillment of the Machine Learning Course Nos. BITS F464

Under the supervision of Dr. Ashwin Srinivasan and Tirtharaj Dash

AT

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE-PILANI,

K.K. BIRLA GOA CAMPUS

APRIL 2020

Comparative Study of Machine Learning Models for Forecasting the Spread of Covid-19

Vedant Bordia[1], Tushar Maloo[1], Ayush Kumar[2], Abhivandan Gandhi[2]

[1] Department of Mechanical Engineering, BITS-Pilani, K.K. Birla Goa Campus

[2] Department of Electrical and Electronics Engineering, BITS-Pilani, K.K. Birla Goa Campus

ABSTRACT**:** The Covid-19 pandemic has affected the world immensely at all fronts. Economies of countries have collapsed while medical services and facilities are facing severe stress. With the onset of vaccination and the second wave across the globe, it is now even more important and necessary to forecast the spread of covid-19 so the whole world can be prepared to face the upcoming dire situations. The aim of this report is to do a comparative study of various machine learning models to get the best forecasting possible. The models were trained with 9 input parameters like confirmed cases, fatalities, etc. A total of five models are compared, details of which will be discussed in this report, amongst which the best one is Holt's Winter model with a mean absolute percentage error less than 1%. The study clearly points out a very good forecasting of cases has been achieved, along with the fact that vaccinations did reduce the active case numbers as expected by all.

KEYWORDS: Machine Learning, Linear Regression, Polynomial Regression, SVM, Holy's Linear model, Holt's Winter Model

1. Introduction

The SARS-CoV-2 virus causing novel coronavirus disease 2019 (COVID-19) pandemic continues to pose a critical and urgent threat to global health. This pandemic continues to strain medical systems around the world in a variety of ways, including sharp rises in demand for hospital beds and critical shortages of medical supplies, while many healthcare staff have been contaminated. As a result, the ability to make immediate clinical decisions and use healthcare services effectively is critical. Alongside this the global economy took a huge dip, which is being compared to the 2008 financial crisis. The COVID-19 pandemic has led to a total of 18,762,976 positive cases, while fatalities upto 208,330 as recorded on 29th April 2020. With such huge number with are increasing quite rapidly, there is an urgent need to predict these numbers with utmost accuracy.

There are numerous techniques related to modeling, statistics, data mining, artificial intelligence (AI), and machine learning that are being used to analyse the data collected till date to forecast future trends. The various stages involved in such analysis and predictions include defining the task, collecting related data from various sources, analyzing the data, implementing various models to forecast the numbers and finally, the quantitative and qualitative analysis of the results obtained from the best performing model. Prediction models combine several input features to estimate the spread of covid-19.

In this paper, we aim to implement various machine-learning models that predicts the confirmed cases, recovered cases and fatalities, and then do a comparative study to get the best possible methods. The models were trained on the previously available data of confirmed cases, recovered cases and fatalities, along with different data indicating the mobility of the local population. For our study we focused on the continent of Europe, but the proposed model will be applicable globally.

2. Data Collection

The data used in this study were taken from various resources. Data for no. of confirmed cases, no. of deaths and no. of recovered cases over the time were taken from John Hopkins University GitHub directory. The data present here gets updated regularly so care was taken such that whenever the code is run, the updated data is automatically downloaded. Data corresponding to the mobility change was taken from google and in a manner similar to the previous care was taken to use updated data as much as possible. The data downloaded from Google was then cleaned and unnecessary columns were removed. Lastly, data corresponding to the vaccination drive was taken from Kaggle (https://www.kaggle.com/gpreda/covid-world-vaccination-progress) and was cleaned to match the study requirements.

3. Methodology

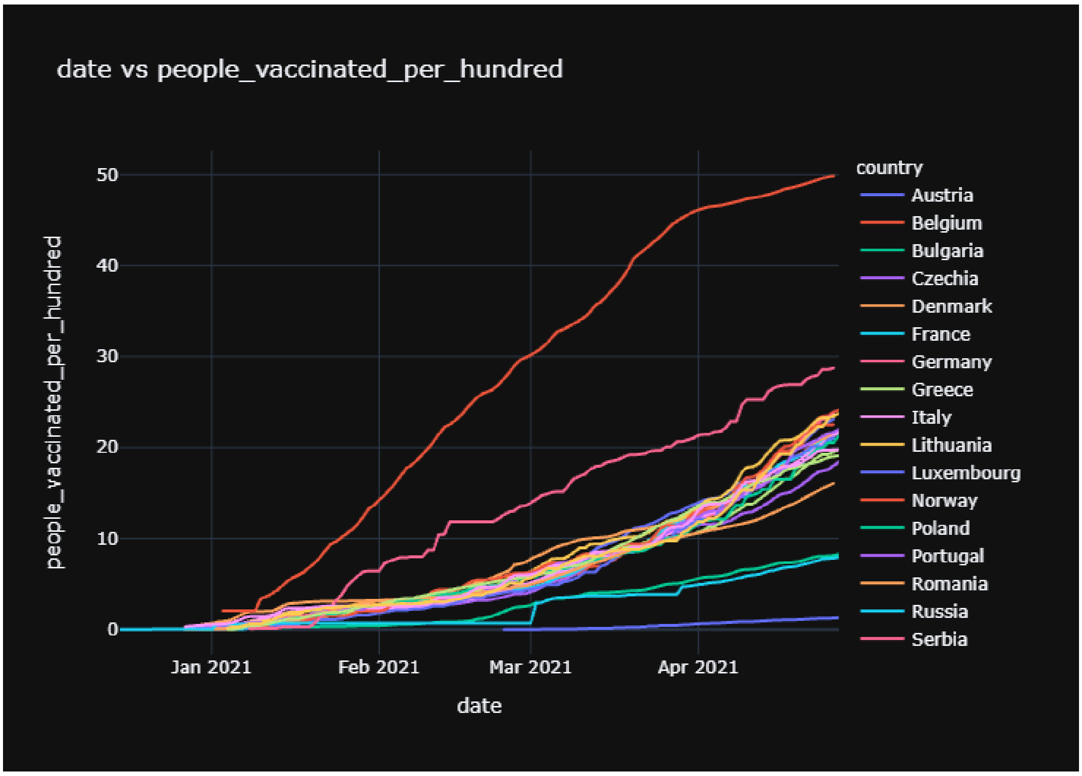
The procedure implemented to carry out the comparative study of models is as follows:

a. **Data Collection** – The data was collected as mentioned in the previous section.

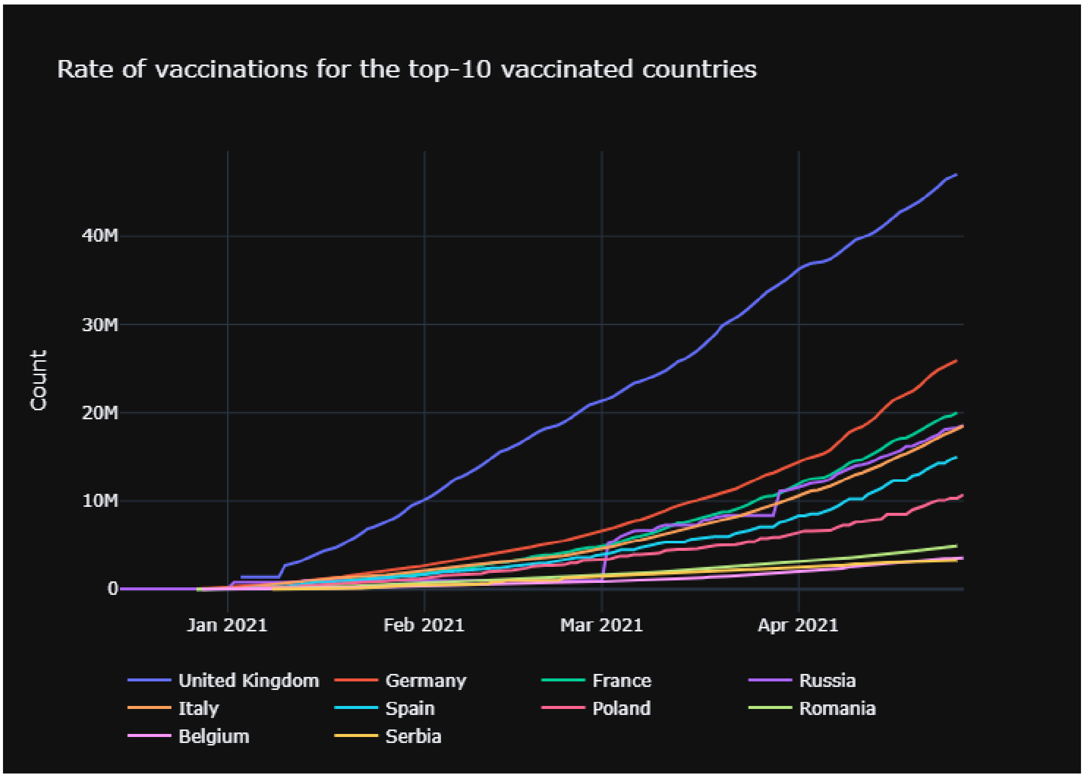
b. **EDA and Visualization** – The dataset chosen is analyzed studied extensively. Then all the unnecessary data was removed, and dates were converted to datetime format. Finally, the datasets of confirmed cases, recovered cases and fatalities was merged with the mobility dataset. Mobility dataset has the column:

- retail_and_recreation_percent_change_from_baseline
- grocery_and_pharmacy_percent_change_from_baseline
- parks_percent_change_from_baseline
- transit_stations_percent_change_from_baseline
- workplaces_percent_change_from_baseline
- residential_percent_change_from_baseline

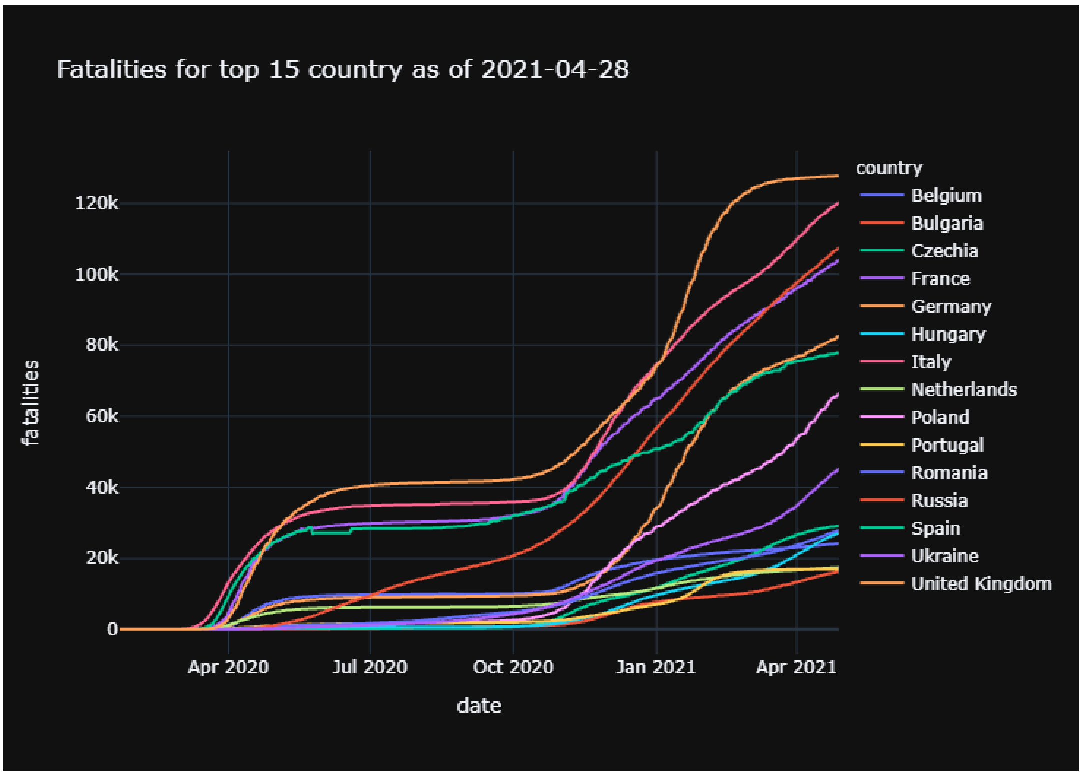
Data visualization was done to a great depth to figure out trends of various parameters with respect to time.



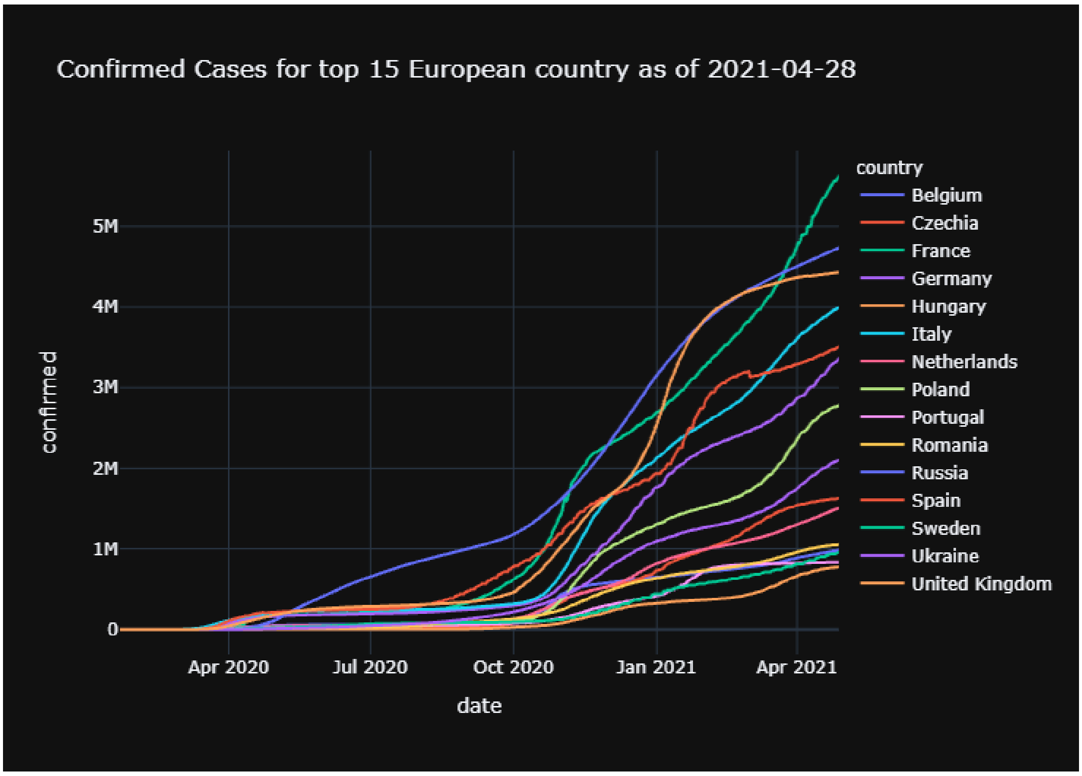
Vaccination per hundred Graph



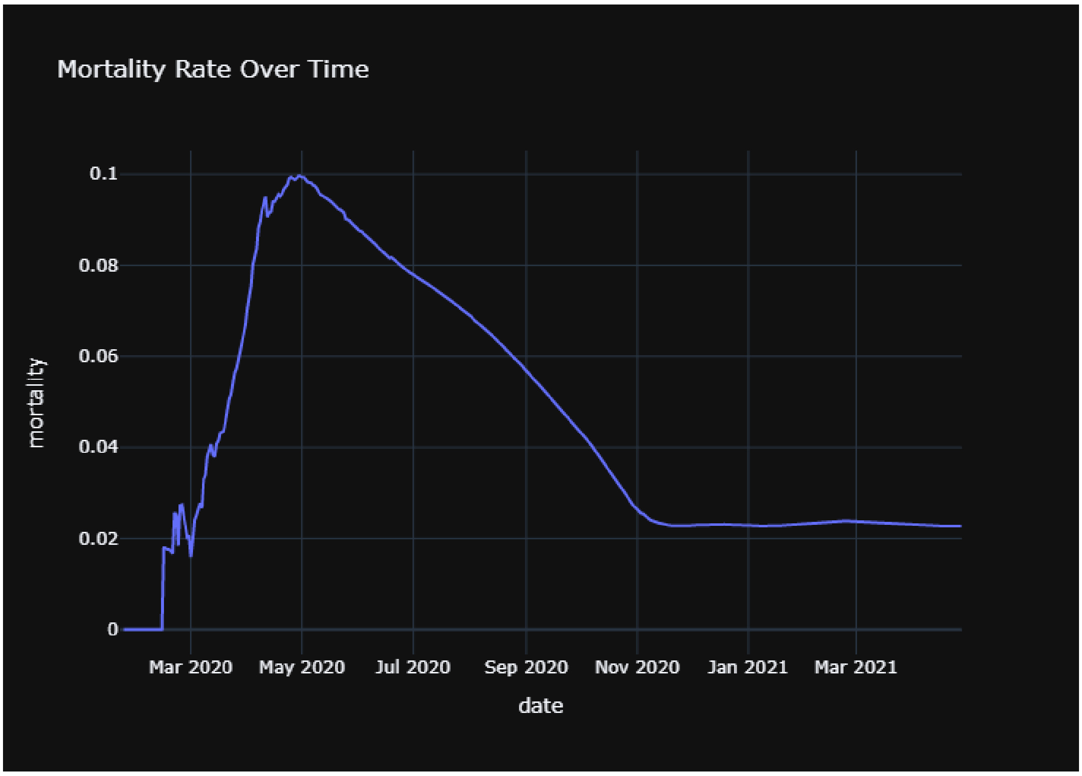
Rate of Vaccination Graph



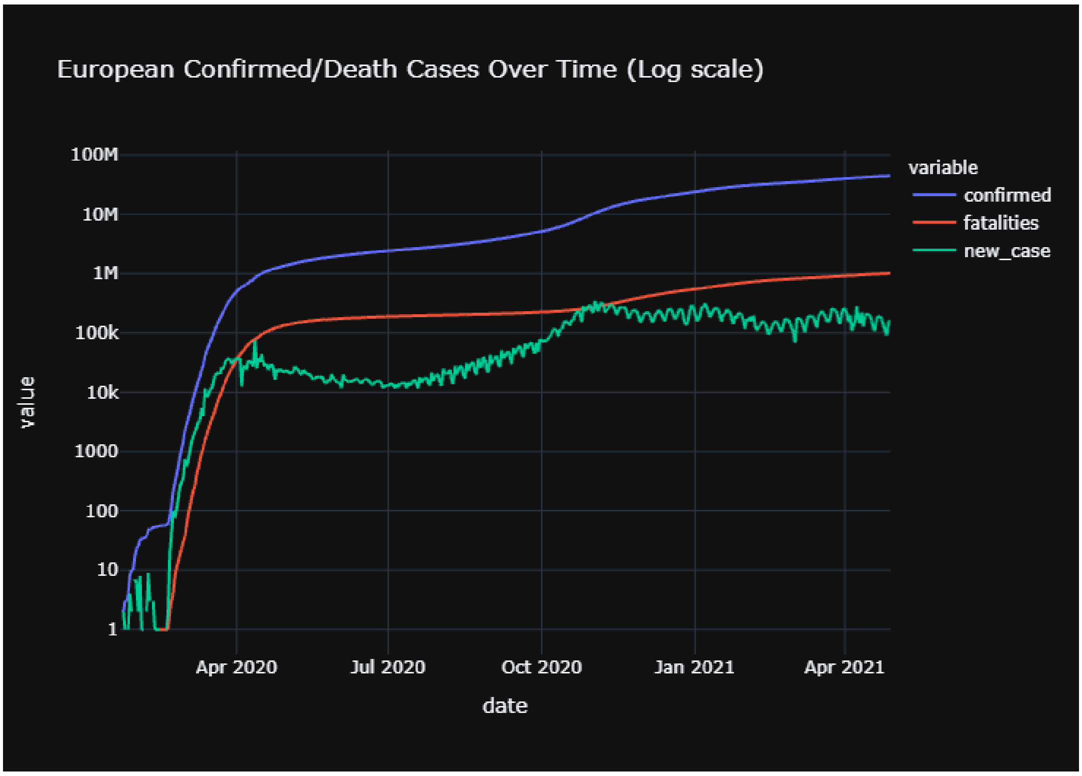
Fatalities Graph



Confirmed Cases Graph



Mortality Chart



Log Graph of cumulative data of Europe

c. **Comparative Study of Machine Learning models** - We have presented predictions based on three regression-based models and two Holt’s time series-based models were used.

i. Linear regression - a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

ii. Polynomial regression - It is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.

iii. Support Vector Regression (SVR) - It uses the same principles as the Support-vector machines (SVM) for classification, with only a few minor differences. First of all, because output is a real number it becomes very. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

iv. Holt's Linear Model - Holt linear attempts to capture the high-level trends in the time series data and fits the data with a straight line. The method can be summarized as follows:

Forecast, level, and trend equations respectively

$$\hat{y}_{t+h} = l_t + h \cdot b_t$$

$$l_t = a \cdot y_t + (1 - \alpha) \cdot (l_{t-1} + b_{t-1})$$

$$b_t = \beta \cdot (l_t - l_{t-1}) + (1 - \beta) \cdot b_{t-1}$$

In the above equations, α and β are constants that can be configured. The values lt and bt represent the level and trend values respectively. The trend value is the slope of the linear forecast function and the level value is the y-intercept of the linear forecast function. The slope and y-intercept values are continuously updated using the second and third update equations. Finally, the slope and y-intercept are used to calculate the forecast yt+h (in equation 1), which is h time steps ahead of the current time step.

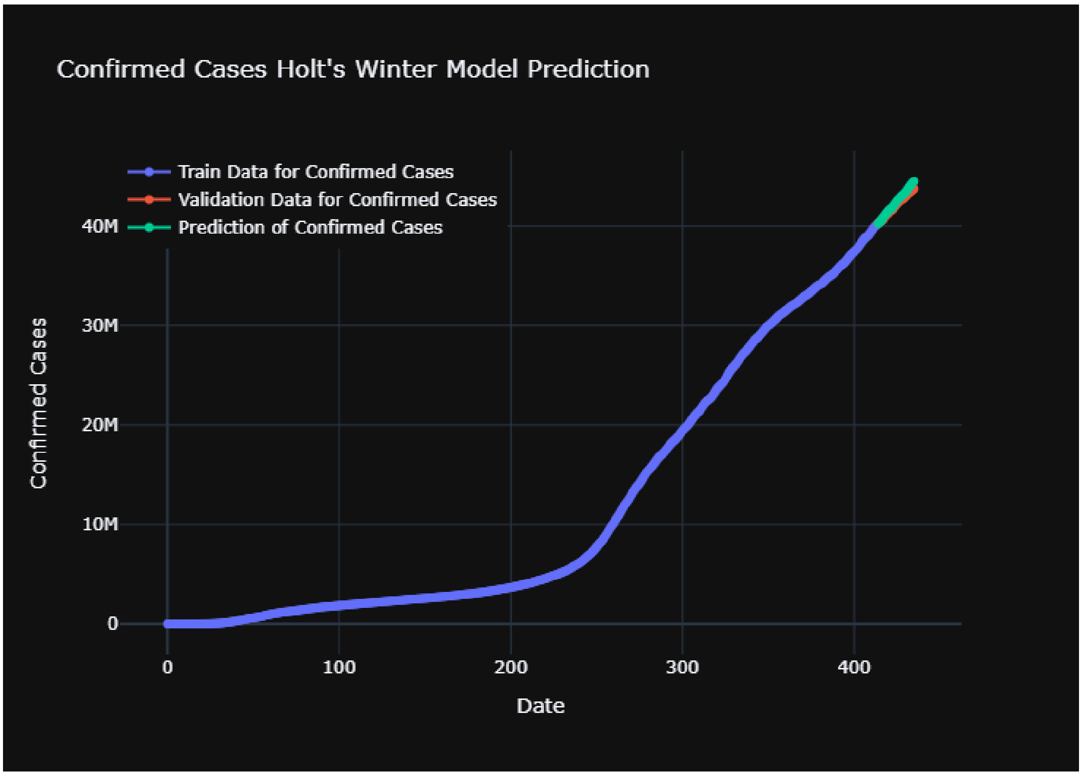
v. Holts's Winter Model -Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict “typical” values for the present and future. Exponential smoothing assigns exponentially decreasing weights and values against historical data to decrease the value of the weight for the older data. In other words, more recent historical data is assigned more weight in forecasting than the older results. Specifically, Holt-Winter's Multiplicative method has been implemented with the following equations:

$$\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\
\ell_t &= \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma\frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}
\end{aligned}$$

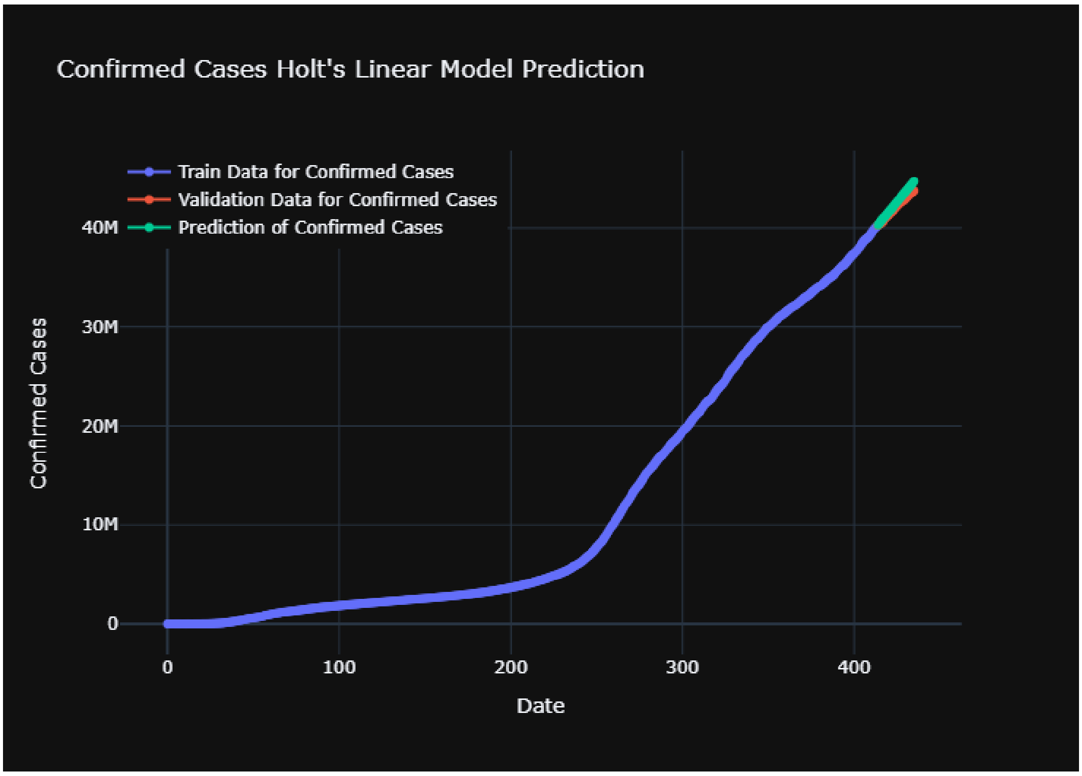
4. Results

This section provides the MAPE (Mean Absolute Prediction Error) score for all the models used to obtain the predictions in the project.

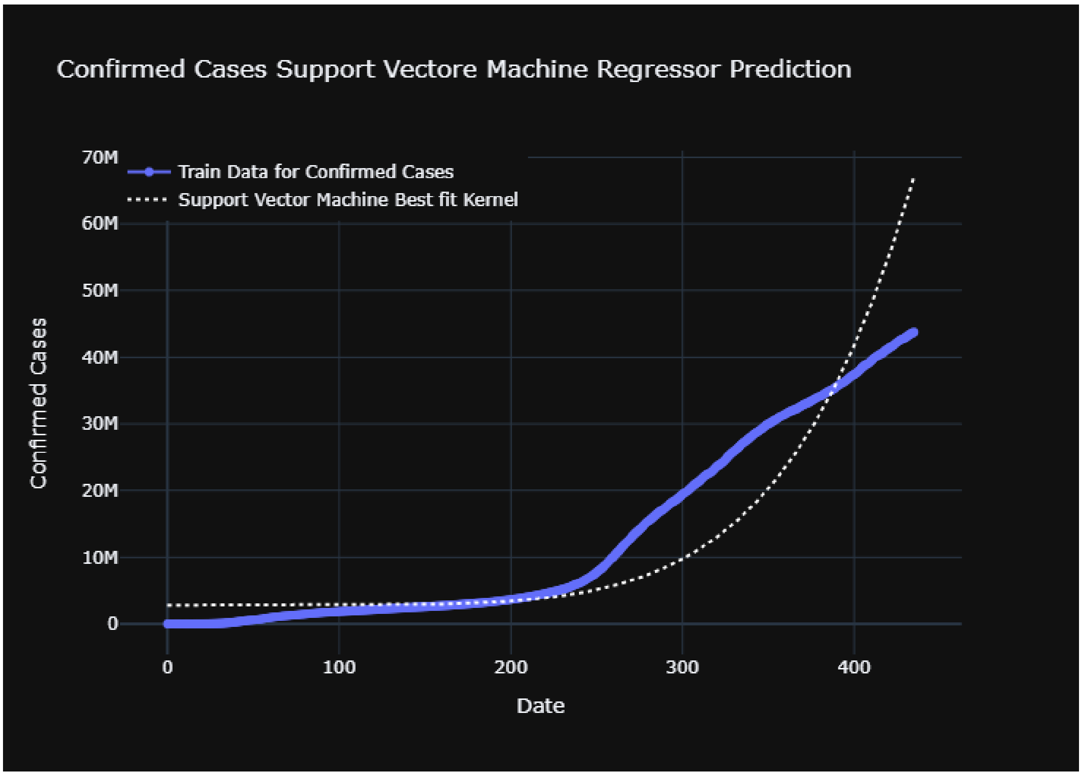
MODELS	MAPE
Holt's Winter	0.7535
Holt's Linear	1.0712
Polynomial Regression	8.8331
Linear Regression	22.8336
SVM	39.1468



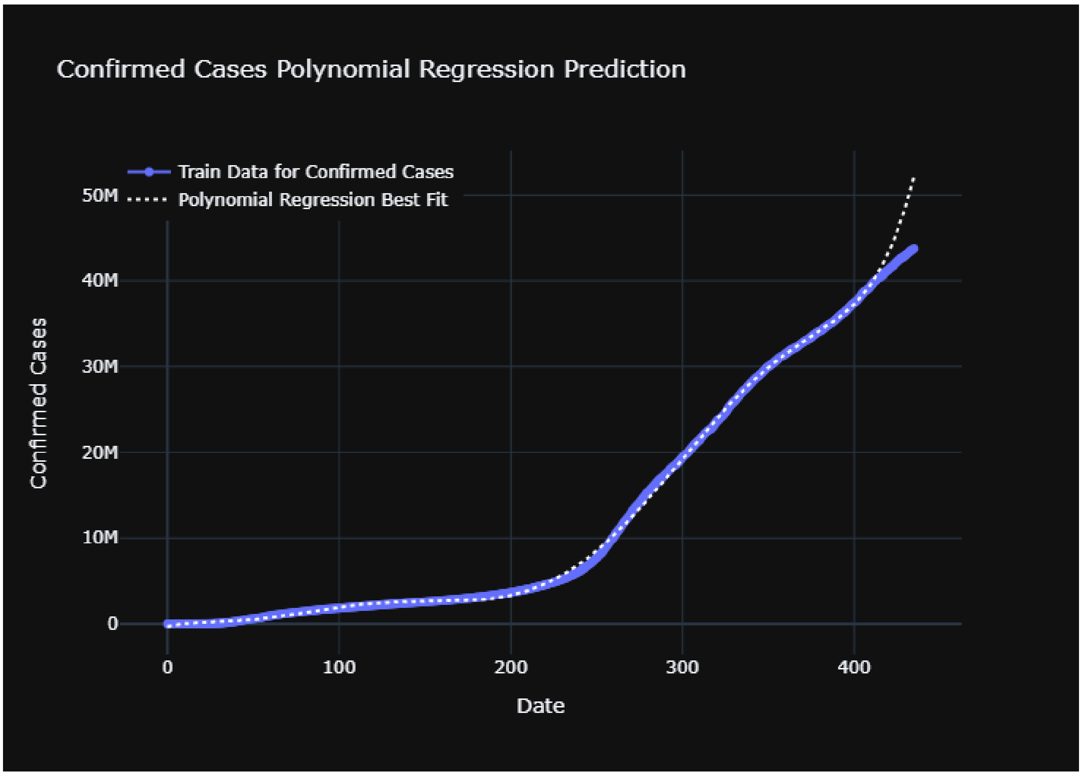
Holt's Winter Model Prediction Graph



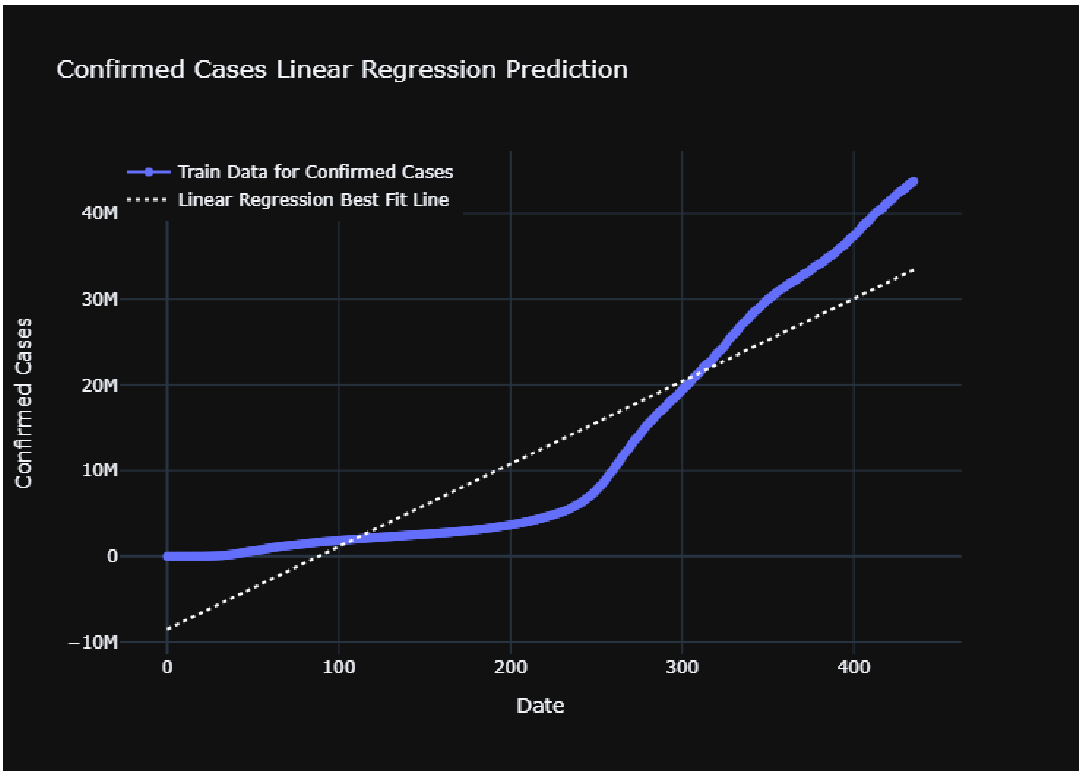
Holt’s Liner Model Prediction Graph



SVM Regressor Prediction Graph



Polynomial Regression Prediction Graph

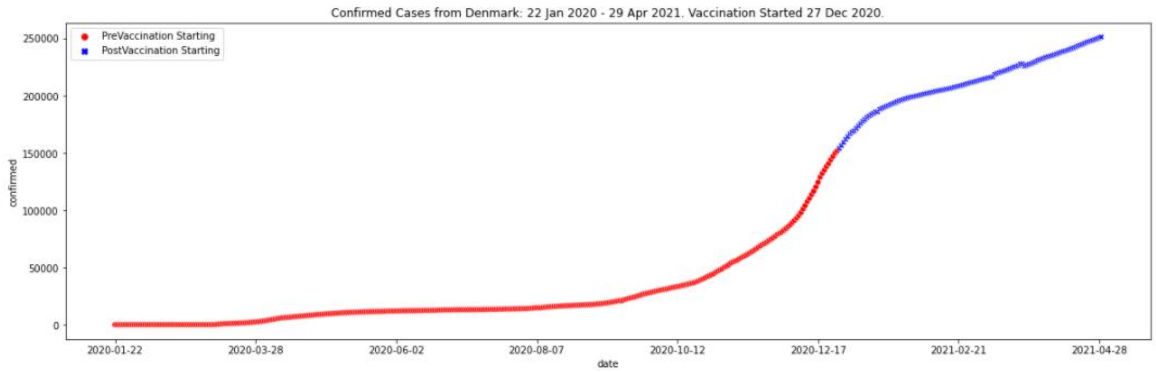


Linear Regression Prediction Graph

From the above MAPE values and graphs for different model it is clear that Holt’s Winter Model is the best model amongst the selected ones.

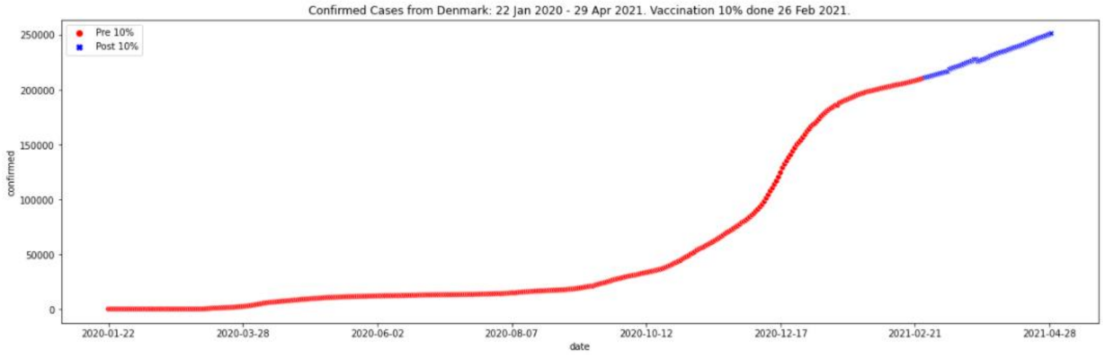
Qualitative Analysis:

For quanlitative analysis, we plotted some graphs on data of selected countries and tried to develop a hypothesis. Here is the study on one of the countries which is Denmark. Initially we plotted a graph of number of confirmed cases before and after the start of vaccination.



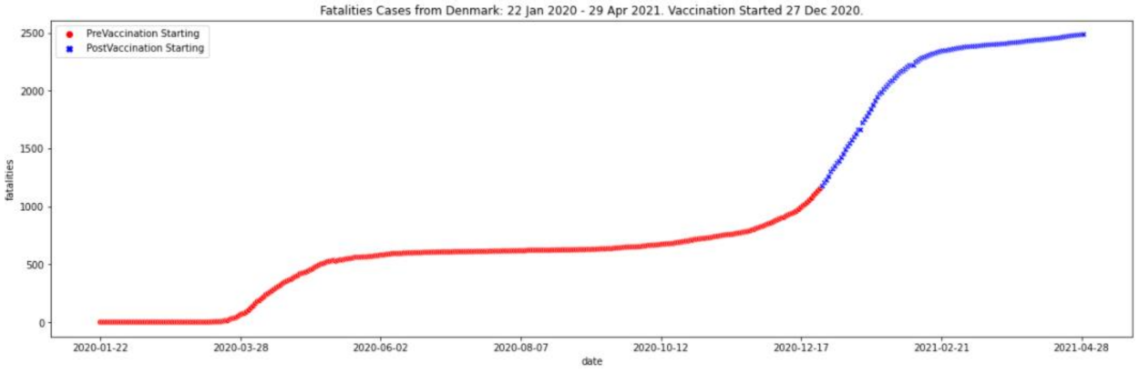
From this graph, we can observe that the slope of graph does not reduces immediately but after some time, the slope starts decreasing. Thus, it can be concluded that the start of vaccination process immediately did not have any major impact on number of cases per day but afterwards when several people got vaccinated, fewer number of confirmed cases were observed.

To look further upon this hypothesis, we tried to observe relation between number of cases before and after completion of vaccination of 10% of total population.



The vaccination of 10 percent population of Denmark got completed on 26th February 2021. From the graph we can say that in the starting when less people were vaccinated, the number of cases were increasing gradually day by day but afterwards the number of cases starts decreasing which is depicted by the blue line as more and more people are vaccinated and we got this result for most of the countries.

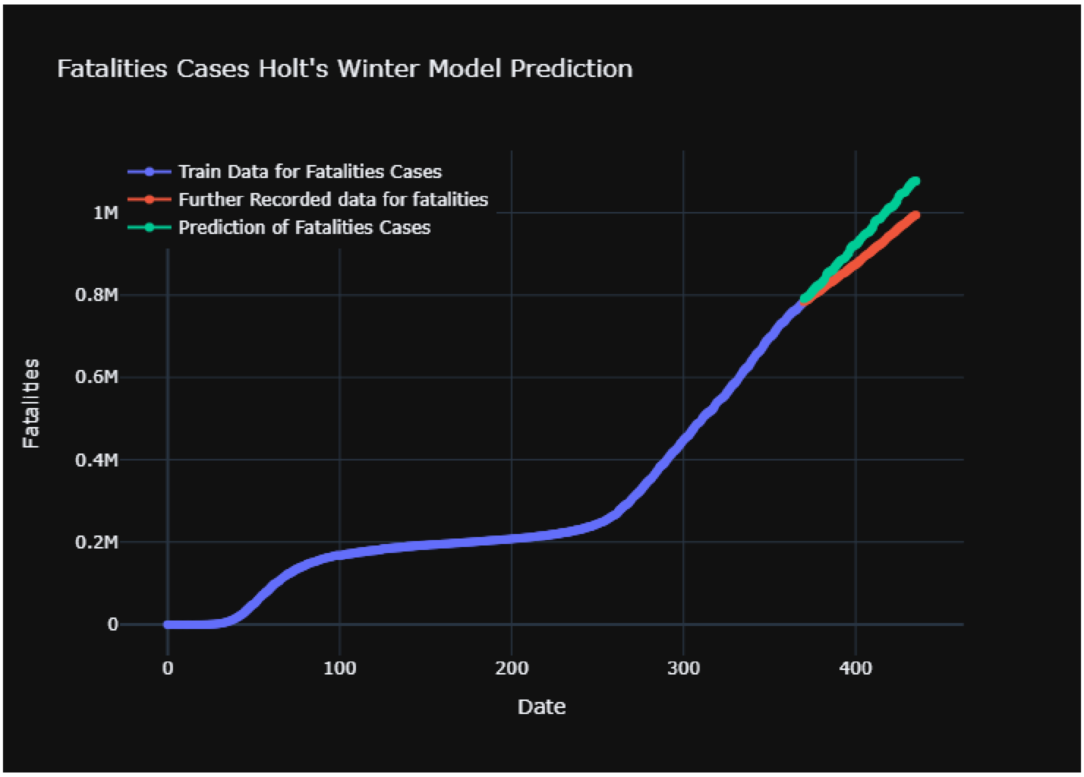
Now we look upon the relation between vaccination and number of deaths recorded.



The time versus deaths plot is more telling of how important vaccination has been in the fight against Covid. After the start of vaccination process, the cases don't immediately starts decreasing as only few people are vaccinated in the starting stage of vaccination but slowly after some time, number of deaths and thus mortality rate has reduced in a regular manner.

Quantitative Analysis:

The graph below compares predicted fatalities, according to Holt's Winter Model which gave an accuracy of 0.7535% during forecasting, and actual fatalities. We can observe that actual number of deaths is lower than the predicted counterpart and keeps on decreasing with the passage of time. Hence, we can clearly say that vaccination results in fewer deaths due to COVID-19. On an average cumulative fatality have decreased by 4.87% and number of fatalities have reduced by 47,243 when compared with the predicted values.



5. Conclusion

The comparison of the MAPE values of the models used we can conclude that the best model that can be implemented for forecasting the spread of COVID-19 is the Holt's Winter method, with a mean absolute percentage error of just **0.7535%.** When analyzed qualitatively and quantitatively, we can clearly see that vaccination has a very positive effect on the COVID-19 pandemic, that is, it has significantly reduced the fatalities caused. We can see that fatalities have been decreased by **4.87%** from the daily predicted values of Holt's Winter Model, which sums upto **47,243** less deaths daily than the predicted numbers.

6. Acknowledgement

We are very thankful of Ashwin Sir and Tirthraj Sir for providing us a opportunity to work on such an open-ended project. It helped us to explore more in field of data collection and preprocessing, machine learning algorithms and learn more about current COVID situation.

7. References

[1] Scikit-learn documentation on Linear Regression implementation: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[2] Scikit-learn documentation on Support Vector Machine implementation: https://scikit-learn.org/stable/modules/svm.html

[3] Scikit-learn documentation on Polynomial Regression implementation: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

[4] Stats Model documentation on Holt's Winter Model : https://www.statsmodels.org/stable/generated/statsmodels.tsa.holtwinters.Holt.html

[5] Plotly documentation : https://plotly.com/python/