# Covid Case Study and Forecasting

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Team 17**

**Members: Amal V Nair, Aruna Chandran, Devika C D, Vyshakh Krishnan T**

**Name: Vyshakh Krishnan T**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

## List of Figures

Figure 1: Total Cases in United States …………………………………………………………………… 8

Figure 2: Total Cases in India ………………………………………………………………………………… 8

Figure 3: Total Cases in Brazil …………………………………………………………………………………9

Figure 4: Vaccination and Case count .………………………………………………………………….. 9

Figure 5: Vaccination and death ……………………………………………………………………………10

Figure 6: Heat map ……………………………………………………………………………………………….10

Figure 7: Cases recorded in India……………………………………………………………………………11

Figure 8: Seasonality component……………………………………………………………………………11

Figure 9: Trend component…………………………………………………………………………………….12

Figure 10: Residual component………………………………………………………………………………12

Figure 11: Transformation and Differencing……………………………………………………………13

Figure 12: SARIMA test vs prediction………………………………………………………………………14

Figure 13: Prophet Model……………………………………………………………………………………….15

Figure 14: Holt-Winters test vs prediction……………………………………………………………….16

Figure 15: Holt-Winters vs New Data……………………………………………………………………….17

Figure 16: Future cases Prediction……………………………………………………………………………19

## Table of Contents

[**Covid Case Study and Forecasting**](#_heading=h.o5tynh11tjuk) **1**

[List of Figures](#_heading=h.e1nfihu7yre1) 2

[Table of Contents](#_heading=h.t6mf4sn0jj5z) 3

[Abstract](#_heading=h.xgx17fkuh68z) 4

[1. Problem Definition](#_heading=h.3s0k97ihto4t) 5

[2. Introduction](#_heading=h.n71ezn71y97q) 6

[Exploratory Data Analysis](#_heading=h.shpht2fom9cj) 6

[Top countries in terms of Covid Count:](#_heading=h.atgyh7ahcbiz) 7

[Vaccination Drive in India](#_heading=h.62sjepd1p3ef) 8

[Effect of Other diseases and smoking on Covid](#_heading=h.hp9uh3n7cqie) 10

[Time Series Preprocessing](#_heading=h.whsfnbu9trea) 10

[Seasonal Decomposition](#_heading=h.6aqgh5opi6k1) 10

[Stationarity Test](#_heading=h.ku73qy7xljey) 13

[Modeling and Evaluation](#_heading=h.8ok0bpah5qte) 13

[ARIMA Model](#_heading=h.kbob22hd9vm2) 13

[Result:](#_heading=h.wvdffaz0fusi) 14

[SARIMA Model](#_heading=h.ysp0l06ukodz) 14

[Result:](#_heading=h.gbq7jmjb70t) 14

[SARIMA+GARCH](#_heading=h.tzx0qvqsal7k) 15

[FB Prophet](#_heading=h.crtiu1zeh96) 15

[Result:](#_heading=h.i5g3ev4g6jay) 16

[Holt-Winters Triple exponential Smoothing model](#_heading=h.1kxr3y9ohf6o) 16

[Result:](#_heading=h.kq4y0vbe6wor) 16

[Covid Self Test App](#_heading=h.22b8bfw76x49) 17

[3. Literature Survey](#_heading=h.vihqyquuexmo) 18

[4. Result](#_heading=h.l5sn4r98s4k) 19

[5. Conclusion](#_heading=h.444j009ugwjl) 20

[References](#_heading=h.cww1af9yyuza) 21

## 

## 

## 

## 

## 

## 

## Abstract

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. Effective screening of SARS-CoV-2 enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed. In this project, we **(Group 17, Batch 1)** aim to:

1) Analyze the world-wide covid data procured from [our world in data](https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv) which include variables such as confirmed cases, confirmed deaths, tests conducted, hospitalization details, vaccination details and other such features, and to build a regression model that will predict new covid cases reported worldwide.

2) Create a web application based on a classification model that will predict the chances of a person having COVID-19, given certain symptoms, using data originally from Israeli Ministry of Health website available at [nshomron](https://github.com/nshomron/covidpred/blob/master/data/corona_tested_individuals_ver_006.english.csv.zip).

## 

## 

## 

## 

## 1. Problem Definition

**1.1 Overview**

Apply Machine Learning for the analysis and forecasting of Covid-19.

**1.2 Problem Statement**

* The Covid pandemic has changed the world scenario. In this context, we aim to analyze the Covid data world wide using exploratory data analysis to gain insight into various aspects of the pandemic such as influence of socio economic factors, vaccination drive, and policy response.
* Further, we intend to perform time series analysis and forecasting of Covid cases in India using machine learning to gain some idea as to what the case count would be in the coming months.
* Lastly we intend to create a web application based on a ML classification model that would suggest the possibility of being Covid positive based on the symptoms the person has; we hope this may serve as an early warning to take Covid tests.

## 2. Introduction

According to the World Health Organization (WHO), "COVID-19 is a disease caused by a new Coronavirus strain. The abbreviation of COVID stands for: 'CO' stands for corona, 'VI' for the virus, and 'D' for the disease. This disease was formally referred to as '2019 novel coronavirus' or '2019-nCoV.' The COVID-19 virus is a new virus linked to the same family of viruses as Severe Acute Respiratory Syndrome (SARS) and some types of a common cold." It is a highly infectious disease which can be transmitted through person-to-person contact and through direct contact with respiratory droplets generated when an infected person coughs or sneezes. About 435,302,116 people were infected worldwide by February 27, 2022, whereas the number of deaths was 5,965,830, and the number of recoveries was 365,575,268.

The governments across the world implemented strict measures such as social distancing and lockdowns to curb the spread of the pandemic. Even so, the disease saw a hike in numbers across three waves. The global economy was severely impacted, mostly affecting the developing countries.

In this project we attempt to analyze Covid data from 15 countries across all continents, create visualizations, gain insights and predict future cases through time series analysis and machine learning using python.

### Exploratory Data Analysis

On analyzing the data we found that the dataset contains 31 columns and 10545 rows. The countries chosen are: Australia, Brazil, Canada, China , France, Italy, India, Kenya, Mexico, Russia, South Africa, Turkey, UAE, UK, USA. The dataset contains covid details such as case counts, deaths counts, vaccination details, policy response, socio-economic features of the countries etc. Only four of the features are of object type, namely iso code, location, continent, and date. The date column is in YY-MM-DD which can be converted to date type. Below are some of the main topics we explored.

#### Top countries in terms of Covid Count:

U.S, India and Brazil are the top three countries with the most cases. Visually, the new cases and total case count are represented as:

Figure 1

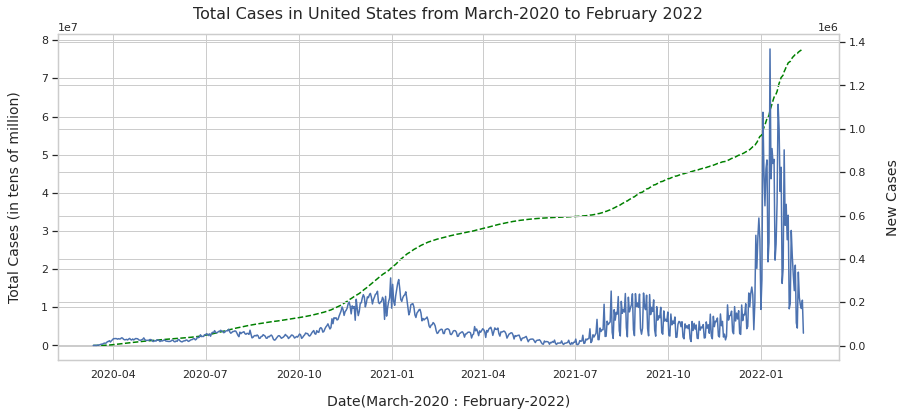


Figure 2

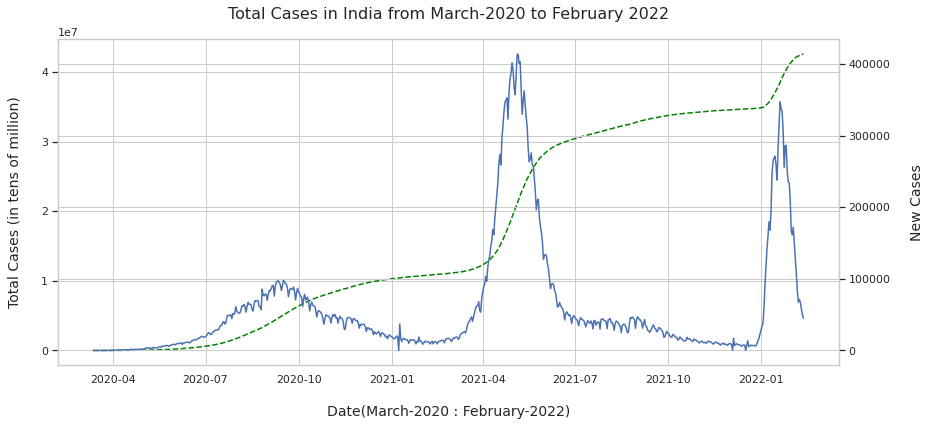
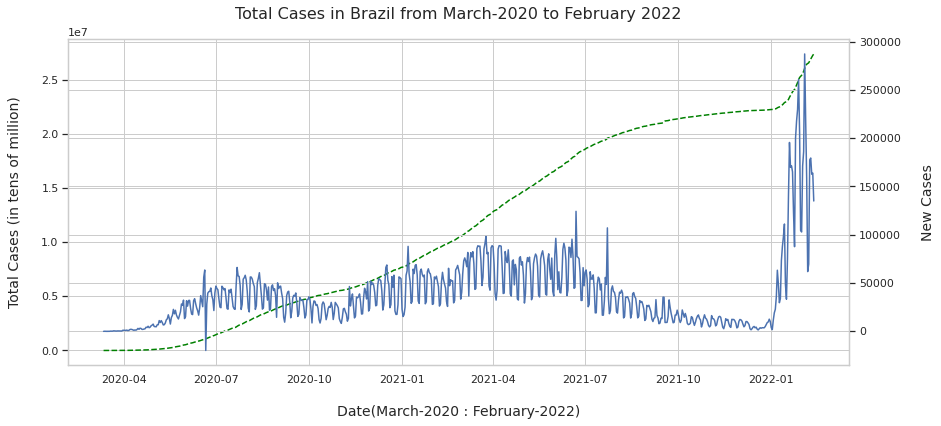


Figure 3



#### Vaccination Drive in India

We further explored the effectiveness of vaccination drive in India, namely the effect of vaccination on cases counts and deaths.

Figure 4: Vaccination and Case count

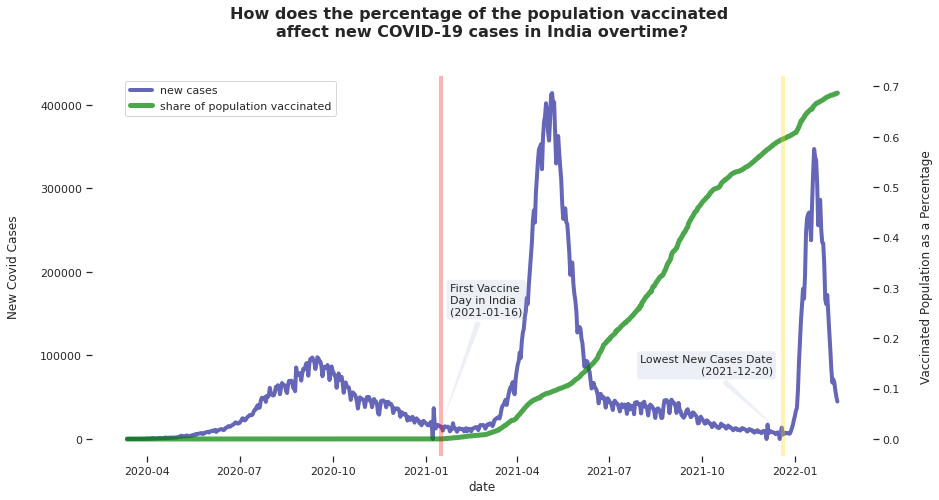
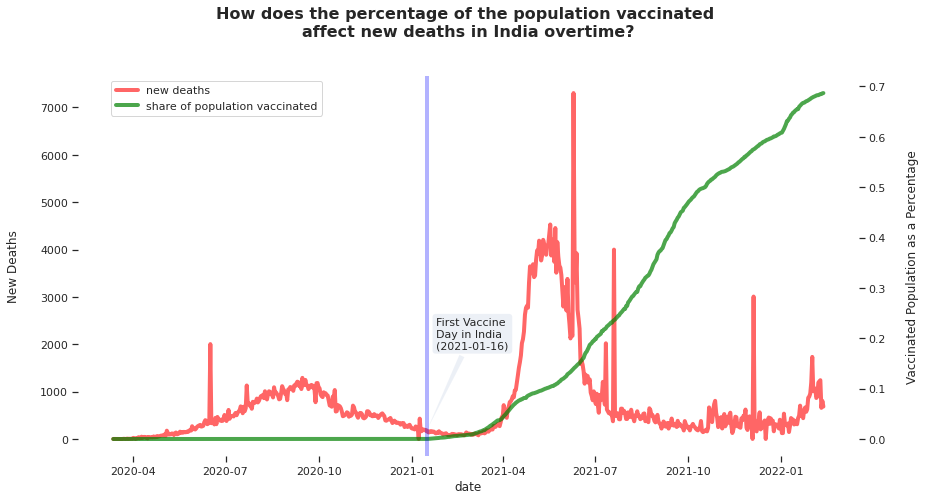


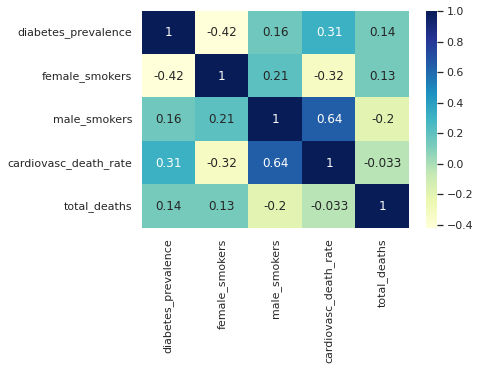
Figure 5: Vaccination and Death counts



#### Effect of Other diseases and smoking on Covid

We explored the effect of other disease on Covid mortality through correlation

Figure 6

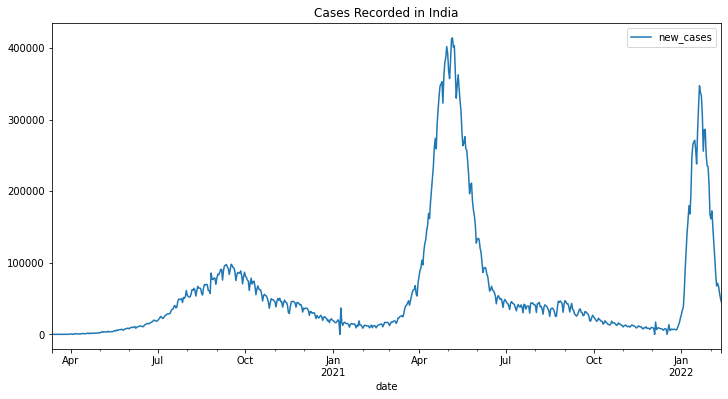
****

### Time Series Preprocessing

#### Seasonal Decomposition

We have attempted to perform time series forecasting on India’s covid data sampled from the main dataset. New cases count is the target column chosen. Plotting the new cases in India, we were able to observe the three Covid waves:

Figure 7:



For the purpose of time series forecasting, we have to analyze various aspects of the series such as seasonality, trend, and residue. Using seasonal decomposition in statsmodels library we observed:

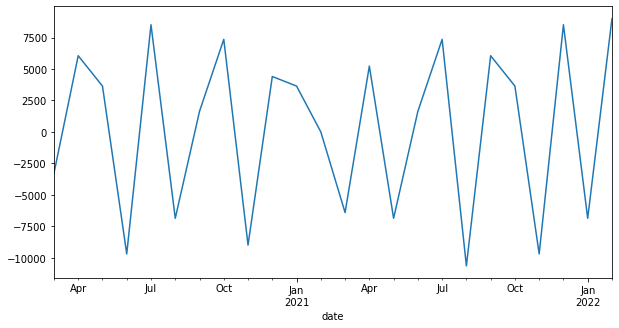
Figure 8: Seasonality Component

Figure 9: Trend Component

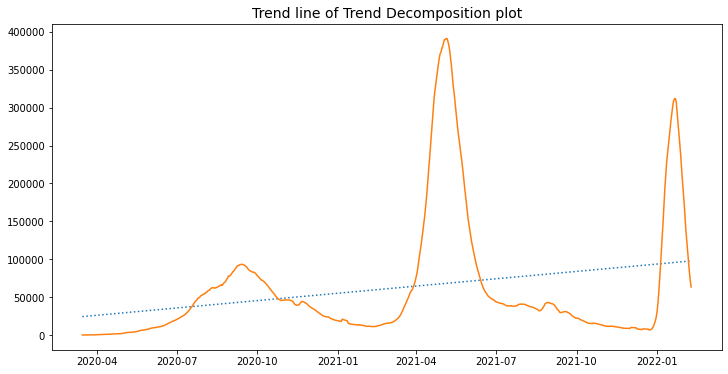
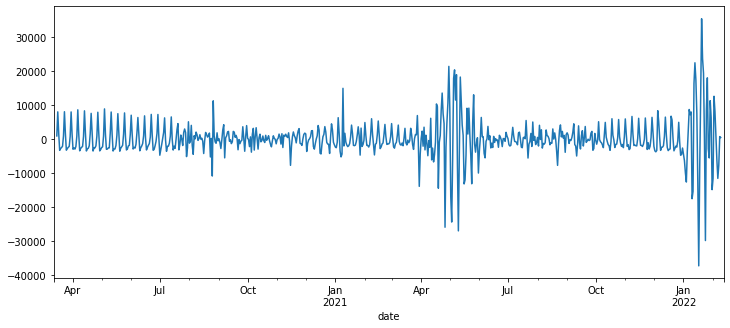


Figure 10: Residual Component

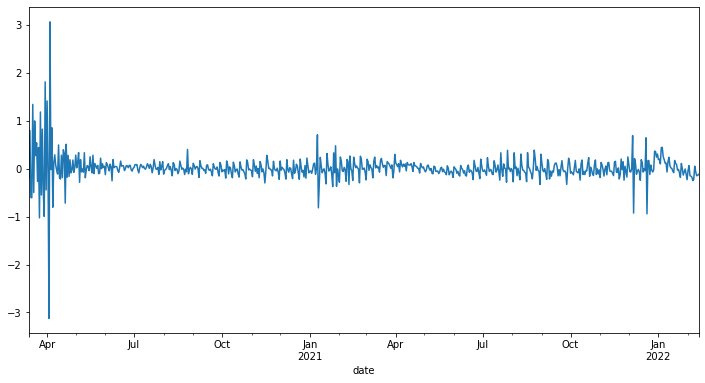
****

We can see that the data contains seasonality and trend. So the data is not just white noise and can be used for forecasting.

#### Stationarity Test

We test the data for stationarity using visualization and Augmented Dickey Fuller test. After applying log transformation and First order differencing, we were able to make the data stationary.

Figure 11: Series after transformation and differencing



We continue onward to forecasting and aim to use models such as ARIMA, SARIMA, FB PROPHET etc.

### Modeling and Evaluation

We have attempted to predict the new cases in India using the data collected through time series forecasting.

#### ARIMA Model

First model evaluated was ARIMA which stands for Auto Regressive Integrated Moving Average. It incorporates changing variable that regresses on its own lagged values(AR component, indicated by parameter p), dependency between an observation and a residual error from a moving average model applied to lagged observations(MA component, indicated by parameter q), and finally integrated component or d which takes non-stationarity into account.

##### Result:

The Arima model failed evaluation on test data, failing to capture seasonality and volatility caused by the third wave.

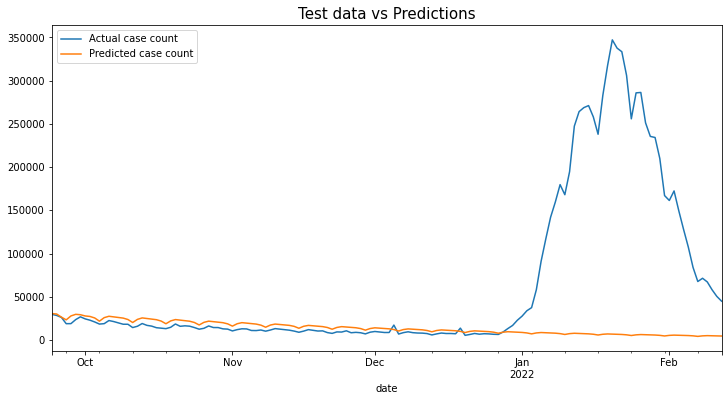
#### SARIMA Model

A seasonal autoregressive integrated moving average (SARIMA) model is one step different from an ARIMA model based on the concept of seasonal trends. In many time series data, frequent seasonal effects come into play. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

##### Result:

As we can see, the seasonality was captured but the third wave was not.

Figure 12: Predictions vs Original cases

****

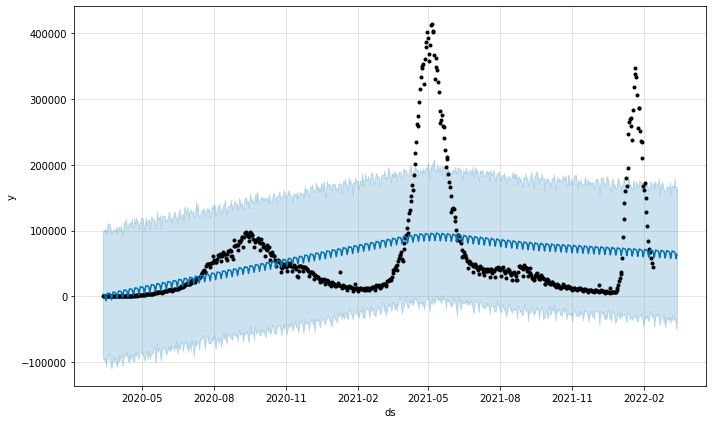
#### SARIMA+GARCH

Generalized Autoregressive Conditional Heteroskedasticity, or GARCH, is an extension of the ARCH model that incorporates a moving average component together with the autoregressive component.GARCH models are used when the variance of the error term is not constant. The GARCH model is actually the ARMA model for the variance, meaning we enable variance to be a function of time and we accept volatility clustering. While ARMA can model the mean of a process, GARCH enables us to model volatility of the process. Applying both together as an ARMA-GARCH model we can model both mean and volatility at the same time, relaxing the condition of homoscedasticity.

We can fit the residuals of the SARIMA model on the GARCH model and calculate the final prediction based on the result. However, in python, the implementation is difficult unlike the R language which has the rugarch package. We can continue after statsmodels implement GARCH.

#### FB Prophet

The Prophet library is an open-source library designed for making forecasts for univariate time series datasets. It is easy to use and designed to automatically find a good set of hyperparameters for the model in an effort to make skillful forecasts for data with trends and seasonal structure by default.

Figure 13: Prophet Model

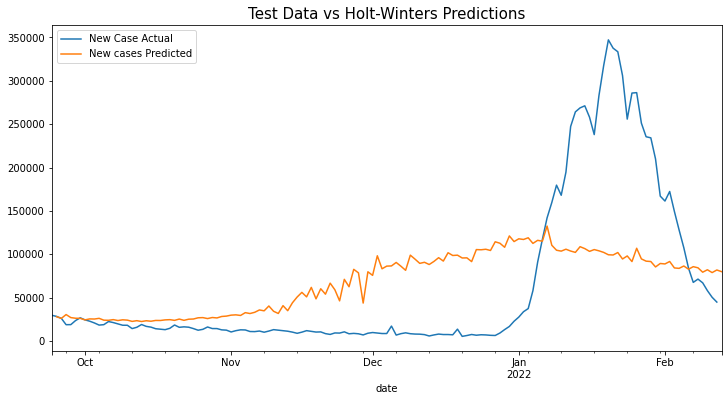
##### Result:

The prophet model also failed to give reasonable predictions.

#### Holt-Winters Triple exponential Smoothing model

Holt-Winters is one of the most popular forecasting techniques for time series. 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).It allows us to add double and triple exponential smoothing to take into account trend and seasonality.

Figure 14: Comparing Holt-Winters prediction with test data

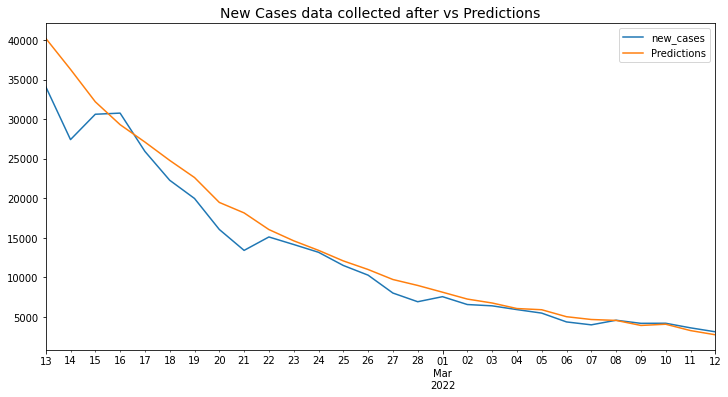


##### Result:

Out of all the models, Holt-Winters actually managed to capture the sudden upward trend, albeit not accurately.

After training the model on full data and evaluating on recently collected unseen data, the model performed reasonably well with an RMSE score of 2543.42 for a series with mean 12841.79; which though not perfect, seems reasonable given the erratic nature of the data.

Figure 15: Evaluating Holt-Winters model using newly collected data



### 

### Covid Self Test App

We also created a self test app using Flask framework and data based on symptoms, travel history, gender and age details. The model on which the app is basing its predictions is a Random Forest Classification model which was selected through extensive evaluation, hyperparameter tuning and also most importantly on the basis of low number of False Negative classifications given(Classified as Covid Negative when in fact the person is positive). The app was hosted using Heroku and can be found at: [Covid Self Test App](http://covidselftest17.herokuapp.com/)

## 

## 3. Result

For the time series analysis part, we tried several models like ARIMA, SARIMA, SARIMA+GARCH(incomplete), FB Prophet, and Holt-Winters. All models except Holt-Winters failed to capture the volatility of the data. Holt-Winters was reasonably accurate though not completely. Comparing the AIC Scores of the three models, Holt-Winters gave the lowest AIC Score of -1915.65. Also the model gave a RMSE score of 2543.42 for new and unseen data having mean of 12841.79 giving us a reasonable forecast.

For the classification model used to build the self test app, several models and approaches were used including balancing techniques, hyperparameter optimization, threshold adjustment and using Voting Classifier to combine three models, to reduce the number of false negatives. However, we were unsuccessful at improving the model beyond a point and resorted to the RandomForest Classifier model that gave an accuracy of 91% and recall score of 80% after applying the balancing technique, SMOTE+TOMEK.

The final model was deployed using the Flask framework and Heroku.

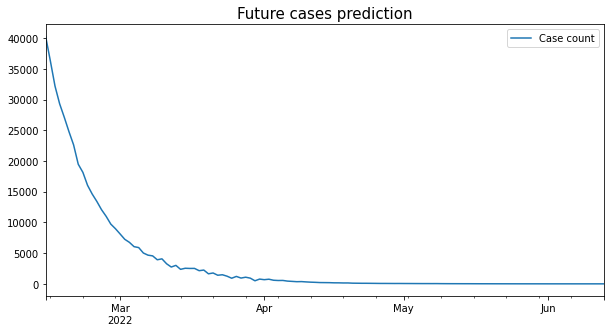
## 

## 4. Conclusion

After visualizing, analyzing and modeling the covid data collected, we were able to:

* Confirm the positive effect of vaccination drive in India by visualizing the decline of new cases and new deaths as the percentage of population who received vaccination increased.
* Create a Holt-Winters time series Model that predicted new cases in India after the decline of the third wave with reasonable accuracy.
* Furthermore, we used the model to forecast three to four months into the future to predict the possibility of a fourth wave.

Figure 16: Future Forecasting with Holt-winters model



Forecasting four months into the future, we see no evidence of a fourth wave in that period. This however doesn't mean a fourth wave is unlikely. We simply lack the means to take into account external factors, vaccination drives and newer Covid strains. Hence new situations could bring about new waves. So it is advisable to maintain safety protocols. Factoring in external influences mentioned above could produce a better model. Implementing GARCH might also improve the model. These could be future endeavors.

* The self test app predicting the possibility of being Covid positive has a recall score of 80% only. Even though not extremely bad, this shouldn’t be used outside an academic purpose for predicting Covid status.

## 5. References

* Covid Time series data: [Ourworldindata](https://github.com/owid/covid-19-data/tree/master/public/data)
* Covid symptoms data: [nshomron-github](https://github.com/nshomron/covidpred/tree/master/data)
* [Analyticsvidhya-times series](https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/)
* [kdnuggets-stationarity in time series](https://www.kdnuggets.com/2019/08/stationarity-time-series-data.html#:~:text=Stationarity%20means%20that%20the%20statistical,and%20models%20rely%20on%20it.)
* [medium-open machine learning course: part 1](https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-9-time-series-analysis-in-python-a270cb05e0b3)
* [medium-open machine learning course: part 2](https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-9-part-3-predicting-the-future-with-facebook-prophet-3f3af145cdc)
* [Eberly College of science: Applied Time Series Analysis](https://online.stat.psu.edu/stat510/lesson/1)
* Udemy - Jose Portilla time series analysis with python course
* [ritvikmath youtube: time series playlist](https://www.youtube.com/playlist?list=PLvcbYUQ5t0UHOLnBzl46_Q6QKtFgfMGc3)
* [Statquest youtube: XGBoost](https://www.youtube.com/playlist?list=PLblh5JKOoLULU0irPgs1SnKO6wqVjKUsQ)
* [Analyticsvidhya: hyperopt technique](https://www.analyticsvidhya.com/blog/2020/09/alternative-hyperparameter-optimization-technique-you-need-to-know-hyperopt/)
* [sklearn - Voting Classifier documentation](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html)
* [geeksforgeeks - Voting Classifier](https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/)
* [machinelearningmastery - threshold moving](https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/)
* [AnalyticsVidhya - Techniques to handle imbalanced data](https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/)
* [AnalyticsVidhya - SMOTE + TOMEK](https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/)
* [kaggle - Jamie Chen: COVID-19 data analysis](https://www.kaggle.com/code/jaimechen00/covid-19-data-analysis-and-visualization)
* [Kaggle - Prashant Banerjee: A guide on xgboost hyperparameter tuning](https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning)
* [Kaggle - BEXGBoost time series notebooks](https://www.kaggle.com/bextuychiev/code)