

**University Mohamed VI Polytechnique**

**EMINES School of Industrial Management**

Analytics edge report

**Topic**

Predictions of COVID-19

Forecasting COVID-19 daily confirmed cases

**Accomplished by: Under the guidance of:**

**Ms. GUESSOUS Maryam**

**BAJDIR Abdelmonim**

**CHRAIBI Youssef**

**ELBARMAQUI Adil**

**SAADI Hatim**

University year 2020-2021

**Thanks**

At the end of this work, we would first like to thank the entire EMINES School of Industrial Management educational team and the professional speakers responsible for the training we had during this year.

Of course, words will not be able to show our gratitude and our warm thanks to all those who contributed directly or indirectly to the realization of this project.

Our deep and lasting gratitude and heartfelt thanks go especially to our supervisor who did not hesitate.

**Abstract**

This project presents a dedicated machine learning model to predict the number of cases infected by the Coronavirus; the case of Morocco and Worldwide was chosen to validate this study. It is Completely realized with the 'Python' language and tested for a certain number of algorithms generated on datasets coming from dedicated sources to gather Covid19 data in the world. The results show the possibility of achieving better scores prediction after using the proposed method. The proposed Machine Learning model can be applied to data from any country in the world. We have applied it in this paper to the case of Morocco and Worldwide. We are sending this work to the world to help them fight COVID-19.

Keywords:

Machine Learning – Model – Predictions – Data preprocessing – Forecast – Fit - Split

Abbreviations list

|  |  |
| --- | --- |
| ts | Time series |
| RMSE | Root Mean Square Error |
| ML | Machine Learning |
| DF | Data frame |
| ds | Date series |
| MAPE | Mean absolute percentage error |

Shapes table

[Figure 1: Import df 11](#_Toc62504799)

[Figure 2 : Rename variables 11](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504800)

[Figure 3 : Extract worldwide data and create new df. 12](#_Toc62504801)

[Figure 4 : Create new time series variable 12](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504802)

[Figure 5: Split data into train and test 13](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504803)

[Figure 6: Max and min of date 14](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504804)

[Figure 7: Train and test visualization 14](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504805)

[Figure 8: Import packages 15](#_Toc62504806)

[Figure 9: Import dataset 15](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504807)

[Figure 10: Head of df 15](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504808)

[Figure 11: Extract data Morocco and create new df 16](#_Toc62504809)

[Figure 12: Daily cases confirmed df tail 16](#_Toc62504810)

[Figure 13: Plot df 17](#_Toc62504811)

[Figure 14: Normalization of data 17](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504812)

[Figure 15: Outliers 18](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504813)

[Figure 16: Stationarity 18](#_Toc62504814)

[Figure 17: Splitting data for linear regression 20](#_Toc62504815)

[Figure 18: fitting linear regression model 20](#_Toc62504816)

[Figure 19: Comparing predicted and actual values 21](#_Toc62504817)

[Figure 20: Preparing data for linear regression (Morocco) 21](#_Toc62504818)

[Figure 21: Comparing predictions and actual values 22](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504819)

[Figure 22: Training the model 22](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504820)

[Figure 23: Define our dependent variable and the target 22](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504821)

[Figure 24: define our model 23](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504822)

[Figure 25: define prophet model for Morocco 23](#_Toc62504823)

[Figure 26: Split data into train and test 23](#_Toc62504824)

[Figure 27: Training the model 23](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504825)

[Figure 28: Results 24](#_Toc62504826)

[Figure 29: Hyperparameters optimization for ARIMA 24](#_Toc62504827)

[Figure 30: Best parameters 25](#_Toc62504828)

[Figure 31: ARIMA results 25](#_Toc62504829)

[Figure 32: Model building with ARIMA for Morocco 26](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504830)

[Figure 33: Predict last months 26](#_Toc62504831)

[Figure 34: Plot actual and predicted values 26](#_Toc62504832)

[Figure 35: Hyperparameters optimization for ARIMA 27](#_Toc62504833)

[Figure 36: Best parameters 27](#_Toc62504834)

[Figure 37: Plot actual and predicted values after hyperparameters optimization 27](#_Toc62504835)

[Figure 38: Calculate RMSE 28](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504836)

[Figure 39: Comparing actual and predicted values 28](https://um6p-my.sharepoint.com/personal/hatim_saadi_emines_um6p_ma/Documents/Rapport.docx#_Toc62504837)

[Figure 40: Forecast next month 29](#_Toc62504838)

Table of Contents

[Abbreviations list 5](#_Toc62504773)

[Shapes table 6](#_Toc62504774)

[Table of Contents 7](#_Toc62504775)

[Introduction 9](#_Toc62504776)

[Chapter 1 10](#_Toc62504777)

[Data preprocessing 10](#_Toc62504778)

[Data preprocessing 11](#_Toc62504779)

[**1.1.** **Worldwide study:** 11](#_Toc62504780)

[1.1.1. Importing dataset 11](#_Toc62504781)

[1.1.2. Splitting data 13](#_Toc62504782)

[1.1.3. Visualization of the train and test 14](#_Toc62504783)

[1.2. Morocco: 14](#_Toc62504784)

[Chapter 2 19](#_Toc62504785)

[Machine learning Models 19](#_Toc62504786)

[Machine learning Models: 20](#_Toc62504787)

[2.1. Linear regression Model 20](#_Toc62504788)

[2.1.1. Worldwide 20](#_Toc62504789)

[2.1.2. Morocco 21](#_Toc62504790)

[2.2. Prophet model 22](#_Toc62504791)

[2.2.1. Worldwide 22](#_Toc62504792)

[2.2.2. Morocco 23](#_Toc62504793)

[2.3. Arima Model 24](#_Toc62504794)

[2.3.1. Worldwide 24](#_Toc62504795)

[2.3.2. Morocco 26](#_Toc62504796)

[Conclusion 30](#_Toc62504797)

[Bibliography 31](#_Toc62504798)

# Introduction

The analytics Edge is the science of using data to build models, improve decisions and ultimately add institutions and individuals. It provides a unified, insightful, modern, and entertaining treatment of analytics.

At EMINES-School of industrial management, as part of learning by doing, we are working on a project which aims to develop predictions about a specific phenomenon. As a group of four students, we choose to work on an interesting topic which is the current health crisis of COVID-19.

Predicting the evolution of COVID-19 is a challenging topic to work on. Not only because of lack of structured data, but also the complexity of model training and making future predictions.

The choice of our project is due to the lack of works treating the forecast of the evolution of the number of COVID-19 confirmed cases , the importance of this predictions for governments and societies to be well prepared for the different variations in addition to the fact that it’s the news subject and the big issue that doesn’t require yet a solution and the fact that we were also victims of the pandemic and we suffered hardly from the confinement and the life changes so we want to contribute in helping to better predict this international and national issue’s variation.

**Chapter 1**

**Data preprocessing**

This chapter aims to prepare the data used in our predictions

# Data preprocessing

* 1. **Worldwide study:**
     1. Importing dataset

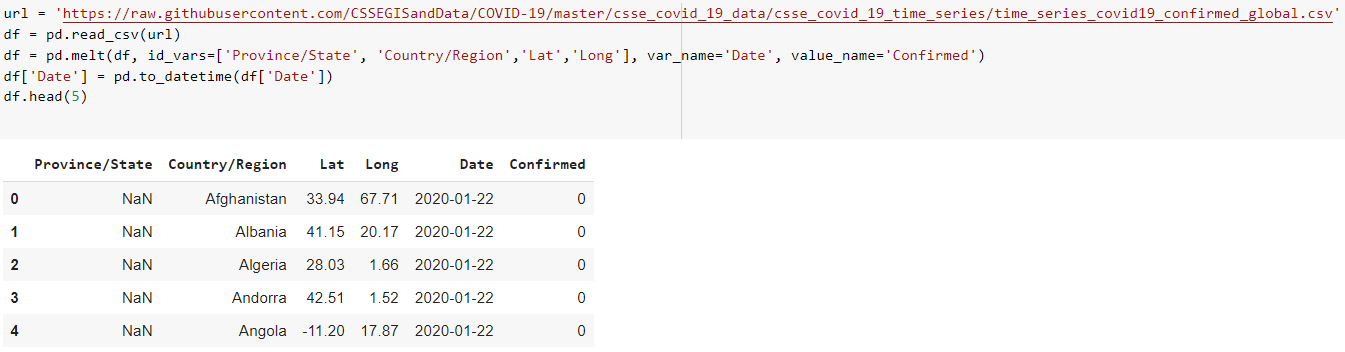


Figure 1: Import df

\*Province/state column has not much importance since it has lot of missed values.

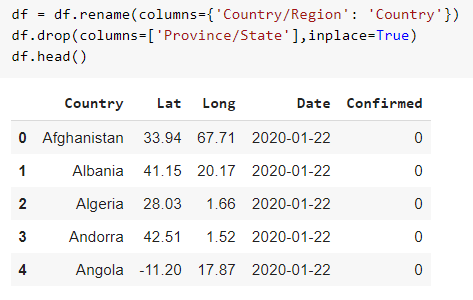


Figure 2 : Rename variables.

First, we group the date by the worldwide cases what is the sum of all values of countries around the whole world. Then we will create new columns to show at the end the evolution of a daily increased COVID-19 worldwide cases, it is about adding: Confirmed\_lag\_1 column, it allows to know the daily increase confirmed cases: which is the Target.



Figure 3 : Extract worldwide data and create new df.

Here we are supposed to create the main variable ts “time series”.

The variable ts will allow us to know how much daily worldwide cases in each date from the start of the propagation of the pandemic.

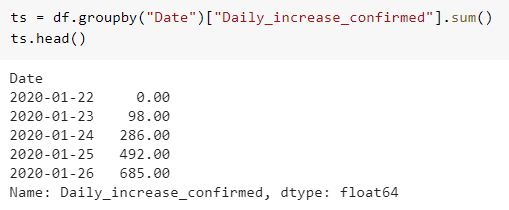


Figure 4 : Create new time series variable.

* + 1. Splitting data

Normally here we are interested in splitting the dates (ts) into training and testing to train the model and test it on 20% of data.

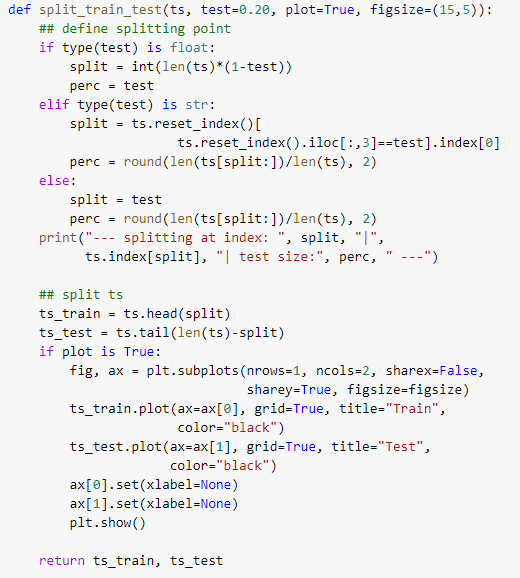


Figure 5: Split data into train and test

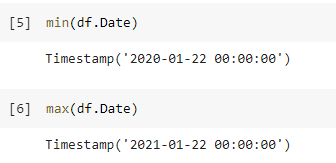
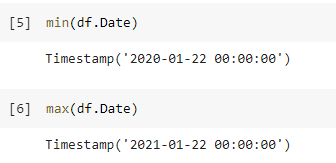
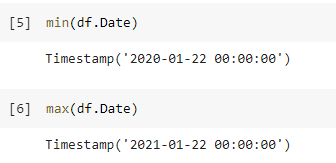
Here is our interval of time of dataset:

Figure 6: Max and min of date



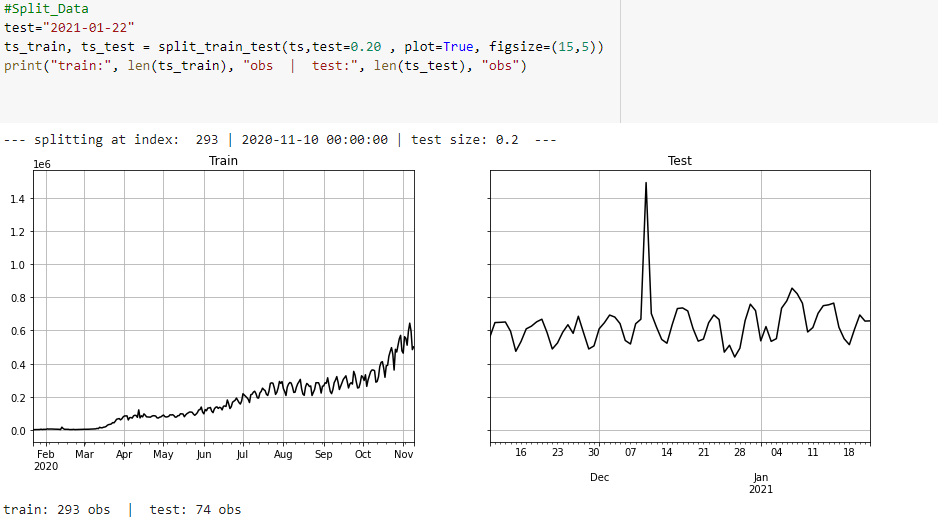
* + 1. Visualization of the train and test

Figure 7: Train and test visualization

* 1. **Morocco:**

First, we will start with importing all necessary packages and libraries for Python:

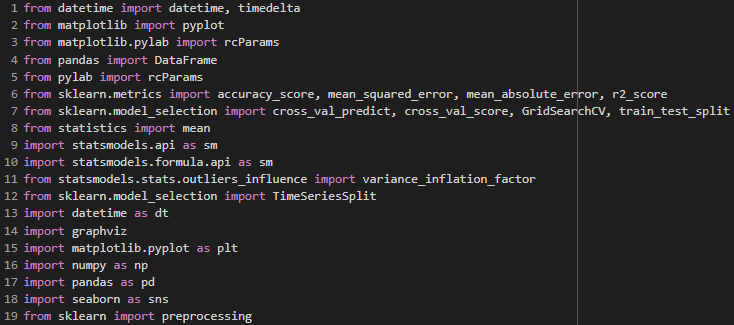


Figure 8: Import packages

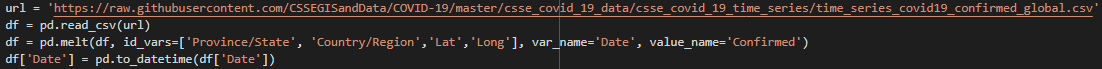
After that, we will use GitHub repository to import our dataset, it is a dataset that contains worldwide data about the evolution of COVID-19. It provides us with daily accumulated cases of each country with the date and information about the countries.

Figure 9: Import dataset

Here is the head of df:



Figure 10: Head of df

At this section, we choose to work on Morocco, which means that the preprocessing of data will be different from the last one (worldwide cases). We will extract the data of Morocco from the general dataset.

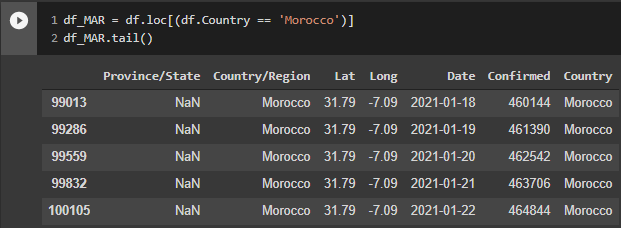


Figure 11: Extract data Morocco and create new df

After that, we delete unnecessary columns:



The dataset provides accumulated cases from the start of COVID-19, but this does not help predicting the evolution of the pandemic and will not give us interesting results. That is why we choose to work with daily positive cases of COVID-19.

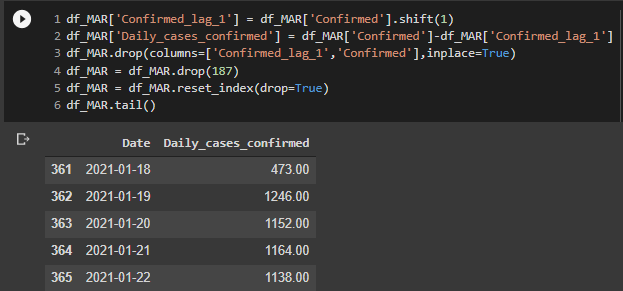


Figure 12: Daily cases confirmed df tail

We will create a time-series variable to analyze our data, to study its outliers, seasonality, and stationarity.

The following graph shows the evolution of COVID-19 for the last 12 months:



Figure 13: Plot df

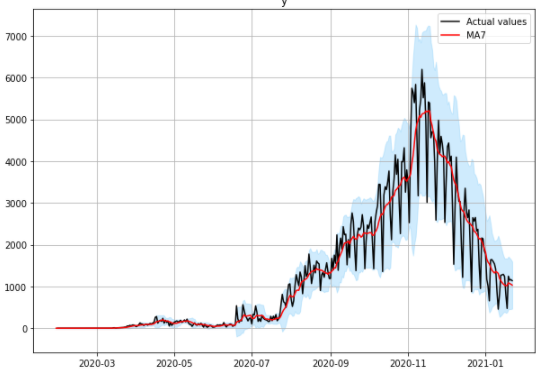
The normalization of the graph helps to identify the regions of high and low increases of daily cases:

Figure 14: Normalization of data

It is normal to have most outliers at the start of COVID-19 because there were approximatively 0 cases/day for the first month, so we will not remove these values because it is logical to have them.

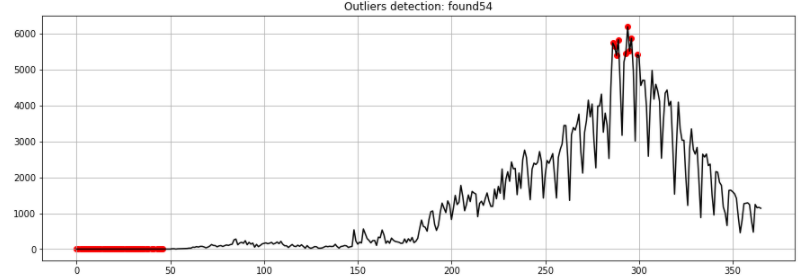


Figure 15: Outliers

The p-value in this case is greater than 0.10, so we have non-stationary data, which means that our distribution changes when shifted in time:

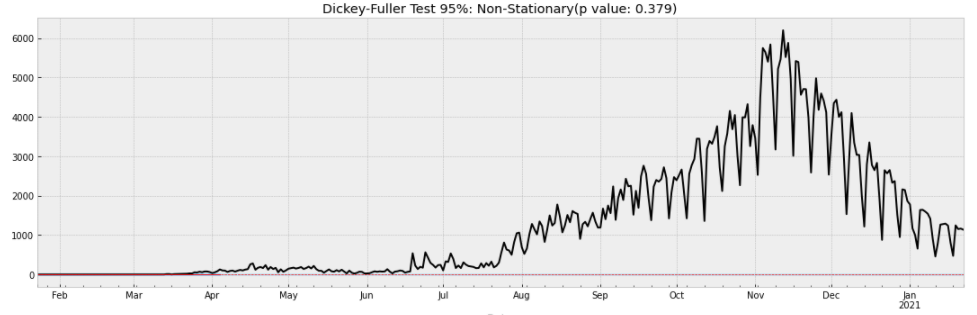


Figure 16: Stationarity

**Chapter 2**

**Machine learning Models.**

This chapter aims to develop ML models to predict the evolution of COVID-19 and forecast for future

# Machine learning Models:

# Linear regression Model

## Worldwide

Preparing the training and testing sets requires splitting the data based on the date. We choose the split\_date variable to have a testing set that is 8 days long.

Une image contenant texte

Description générée automatiquement

Figure 17: Splitting data for linear regression

Defining the model and fitting it to predict the output of the testset:

Une image contenant texte

Description générée automatiquement

Figure 18: fitting linear regression model.

Here is the graph that gives the model and Its RMSE.

The RMSE is high but it is normal because the number of confirmed COVID19 cases in the world is wildly bigger than the number for each country.

Une image contenant texte

Description générée automatiquement

Figure 19: Comparing predicted and actual values.

## Morocco

To predict the number of daily cases using linear regression we start by preparing our dataset. We create new columns that will be helpful for the linear regression model.

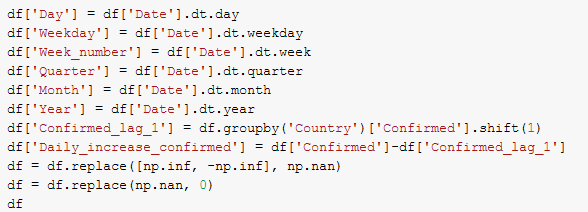


Figure 20: Preparing data for linear regression (Morocco)

df\_Mar is the dataset related to Morocco; it contains columns of df but only rows that have Morocco as Country.

In general, relevant columns for this model are: ['Days\_since\_outbreak\_global','Day','Weekday','Week\_number','Quarter','Month','Year', 'Confirmed\_lag\_7','Days\_since\_outbreak\_country']

Preparing the training and testing sets requires splitting the data based on the date. We choose the split\_date to have a testing set that is 15 days long.

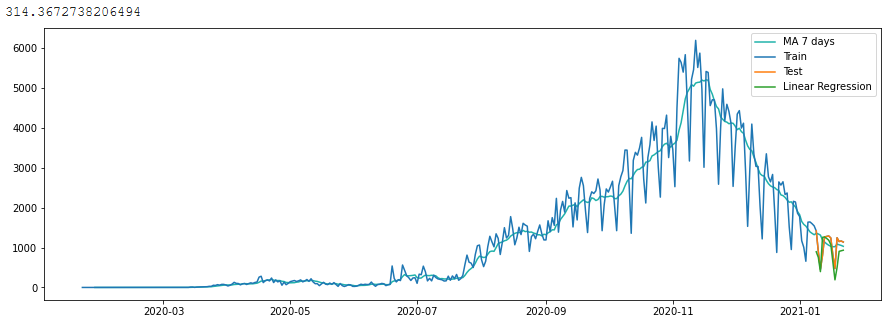
Fitting the model and predicting the cases of the testing set give us this graph.

Figure 21: Comparing predictions and actual values.

The predicted values seem to be close to the real values. The RMSE calculated is 314.36 which means that the model is acceptable, but we can have better results using another model.

# Prophet model

## Worldwide

Let us build our model with ARIMA method:

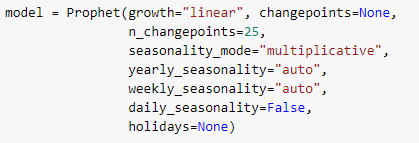


Figure 22: Training the model.

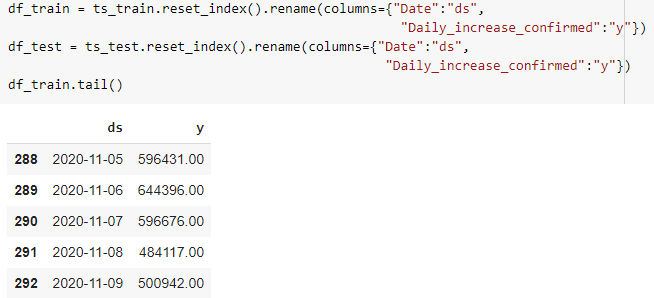


Figure 23: Defining the dependent variable and the target

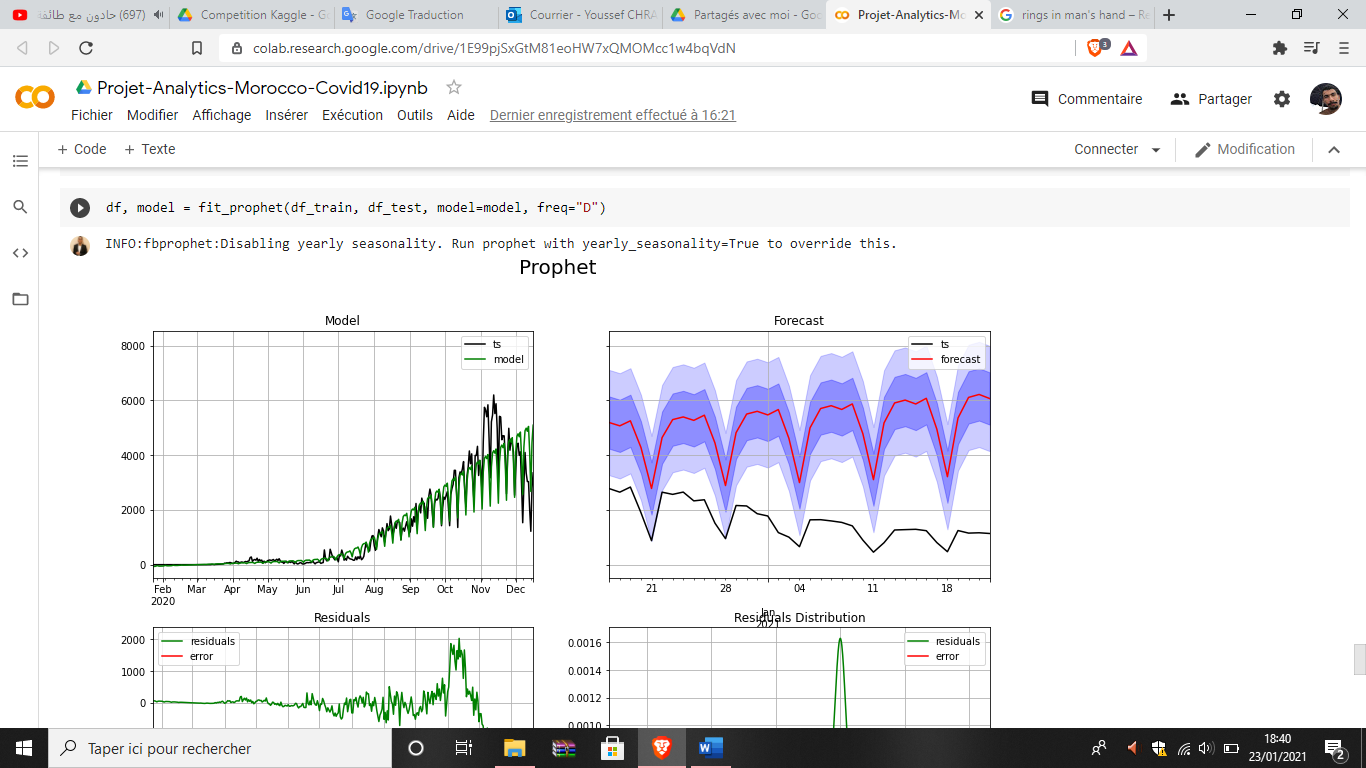


Figure 24: define our model.

## Morocco

First, we define our model based on some defaut parameters such as linear growth, automatical seasonality and a multiplicative seasonality mode, in addition to several potential changepoints equal to 25.

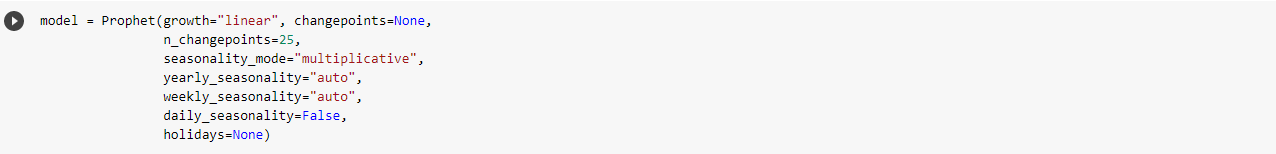


Figure 25: define prophet model for Morocco.

Secondly, we define our train and test sets, and we rename our input and output. Also, we show the tail of our train set for more clarity.

Une image contenant texte

Description générée automatiquement

Figure 26: Split data into train and test

Here is the tail of our training set, it is contained 328 rows and two columns. Our input is ds, and our output is y.

Fitting our prophet model on the training set and testing it to evaluate its forecast.

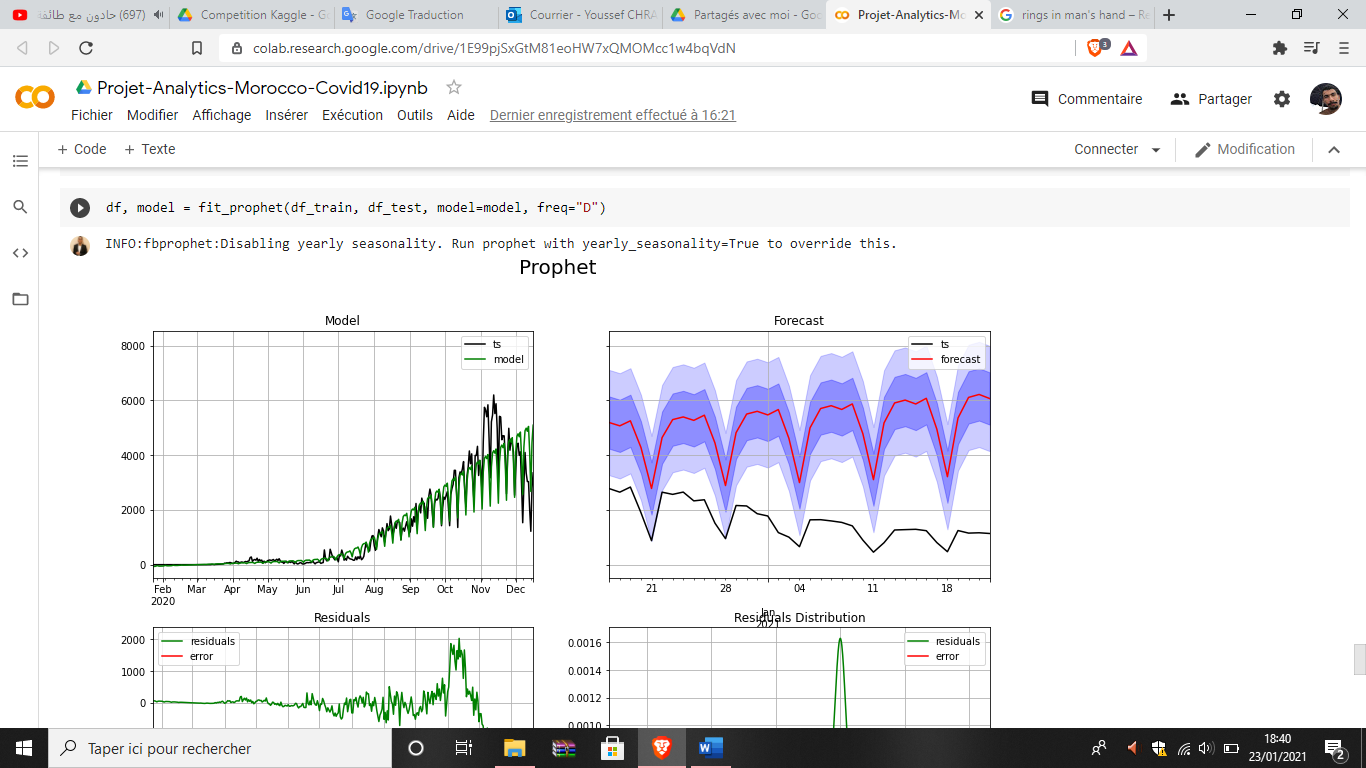
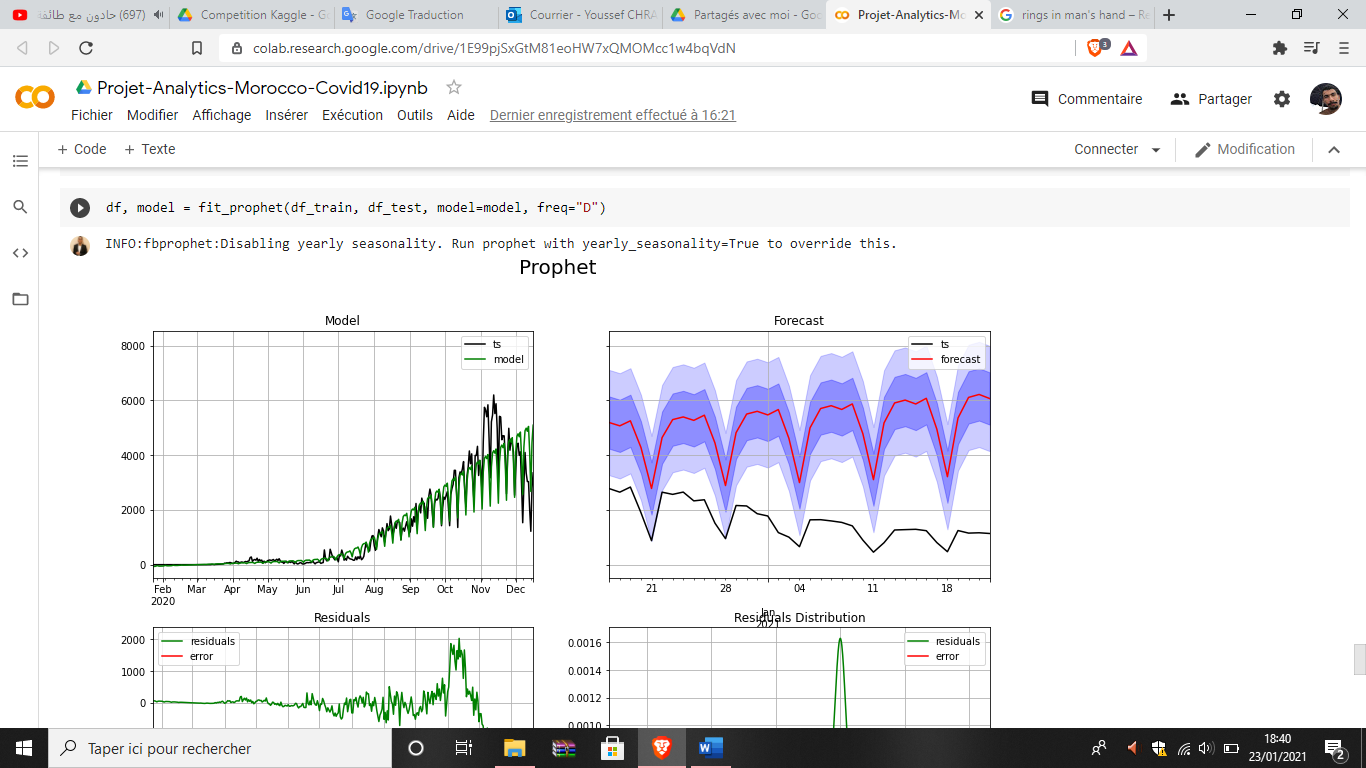


Figure 27: Training the model.

Here’s the graphs of our model’s forecast and its error distribution, for the evaluation of the model we got an RMSE=3632 and MAPE=281% which shows us that the model’s prediction error is higher, so the Prophet model is not that good for the case of Morocco.



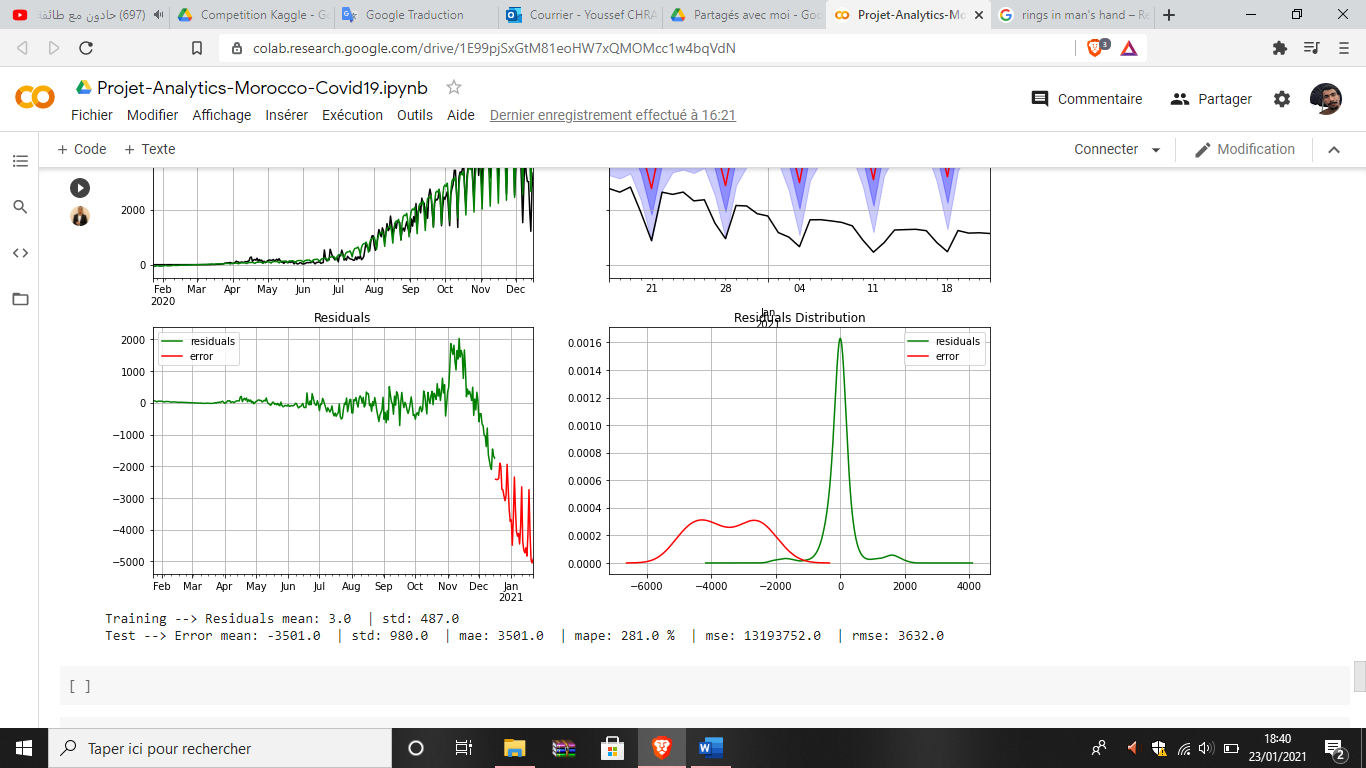


Figure 28: Results

# Arima Model

## Worldwide

To implement the model of Arima we need 7 parameters: (p, d, q), (P, D, Q, s). Therefore, we need to find the best parameters using the function auto\_arima from the library pyramid\_arima ‘pmdarima’.

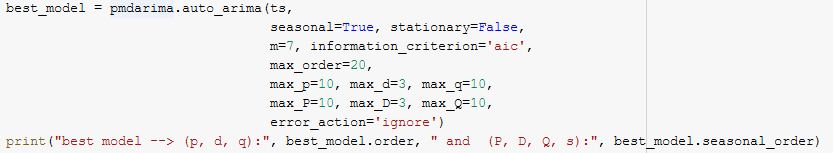


Figure 29: Hyperparameters optimization for ARIMA

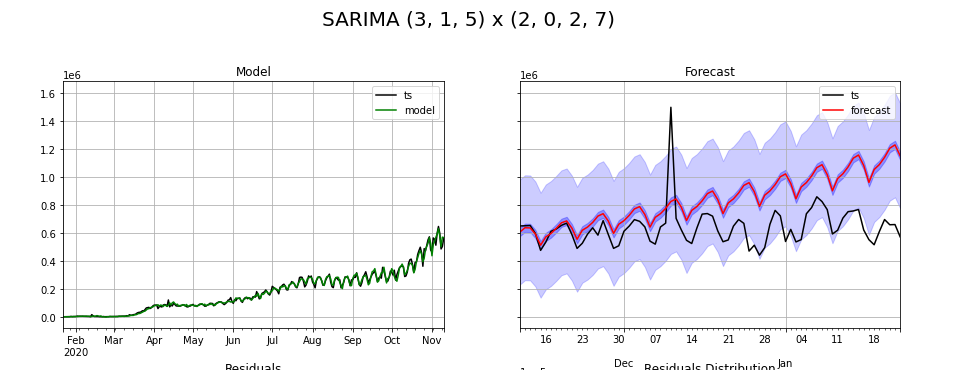
The results were as follows:



Figure 30: Best parameters

A function fit\_sarimax is defined, it takes for parameters the training and testing sets, and the parameters of Arima model (order, seasonality order).

The results were as follows:



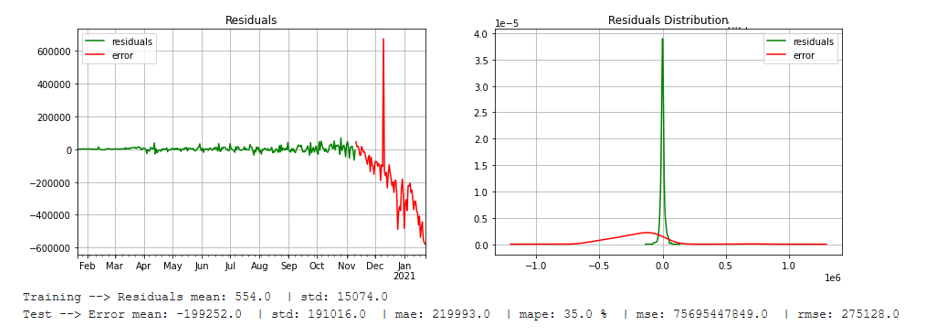


Figure 31: ARIMA results

The graph ‘Forecast’ shows that the predicted values are close to the real ones in the beginning, but after December 2020 the predicted values start to deviate.

## Morocco

Let us build our model with ARIMA method:

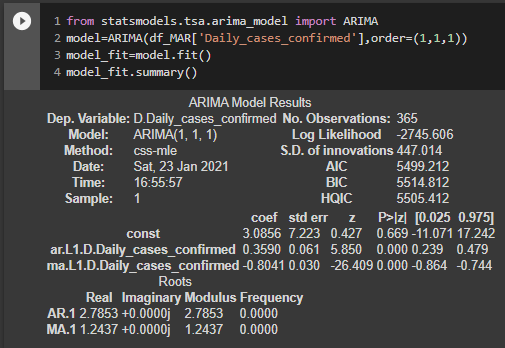


Figure 32: Model building with ARIMA for Morocco.

Now we will split our dataset into train and test, for the test we will work on the last months (26 last days from today) and default hyperparameters for SARIMAX:

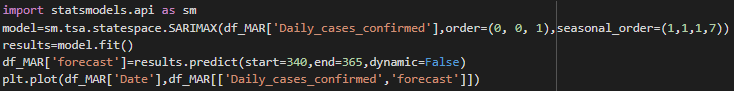


Figure 33: Predict last months

It generates the next graph:

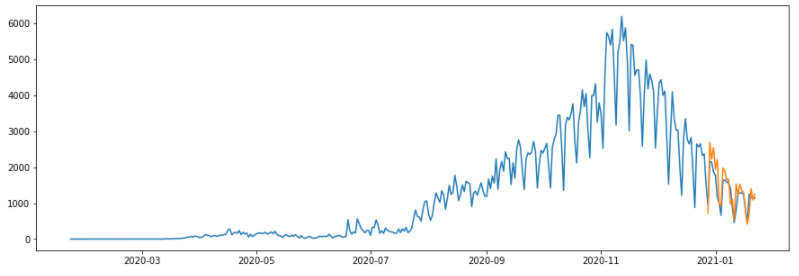


Figure 34: Plot actual and predicted values

Here predictions do not match exactly actual values, it missed some values. Which leads us to search and optimize the hyperparameters as following:

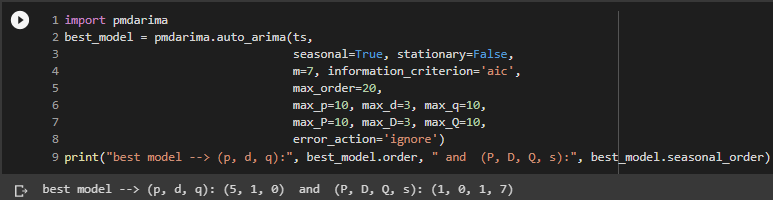


Figure 35: Hyperparameters optimization for ARIMA

Here predictions do not match exactly actual values, it missed some values. Which leads us to search and optimize the hyperparameters as following:

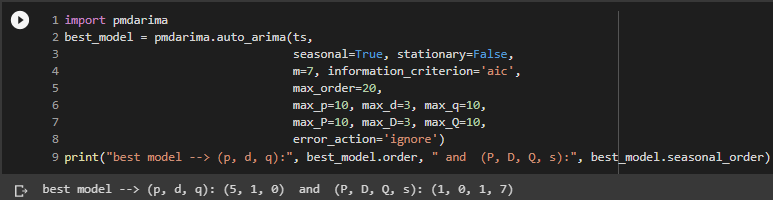


Figure 36: Best parameters

After applying the best hyperparameters for our model, we get the following graph:

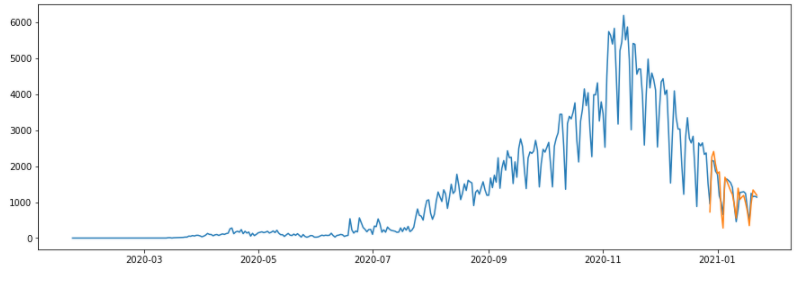


Figure 37: Plot actual and predicted values after hyperparameters optimization

After comparing it with the last one, we get better results, which means that predicted values are very close to actual ones.

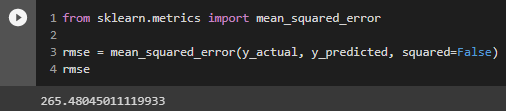
To evaluate our model, we calculate the RMSE:

Figure 38: Calculate RMSE

RMSE = 264

The following table contains a short comparison between actual and predicted values to have an idea about the accuracy of our model and its error:

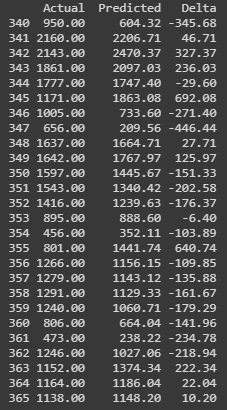


Figure 39: Comparing actual and predicted values.

By comparing RMSE between the three models applied to Morocco, we conclude that ARIMA model gives the closest predictions to actual values, that’s why we will use it to forecast future.

After training the model on the full dataset, we will forecast daily cases in Morocco for the next 21 days.

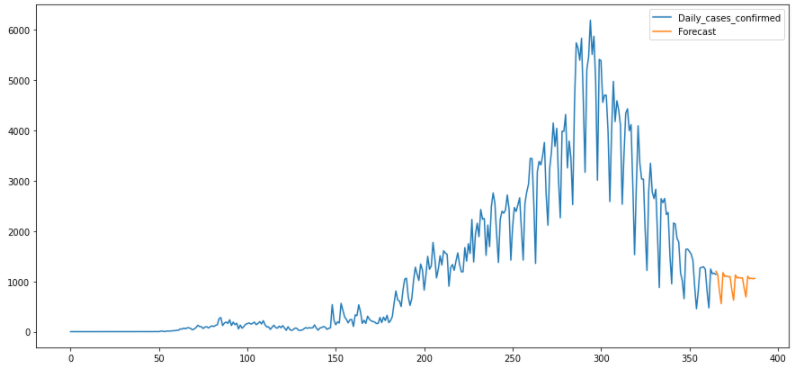


Figure 40: Forecast Next month

The non-stationarity of the data makes predictions hard to make and we will not be sure of the accuracy of predicted values.

# Conclusion

The incertitude of the predictions and the big number of the RMSE metric in some models is explained by the fact that we faced a strong lack of Datasets of COVID-19, the lack of sociological and pandemical studies about relations between confirmed COVID-19 cases and some variables such as number of tests and the portion of respecting some Health measures such as wearing masks, in addition to the non-transparency of some governments and the big number of outliers in the existing datasets.

# Bibliography

* Webography:

<https://towardsdatascience.com/time-series-analysis-for-machine-learning-with-python-626bee0d0205>

<https://github.com/ieee8023/covid-chestxray-dataset>

<https://www.worldometers.info/coronavirus/#countries>

<https://link.springer.com/article/10.1007/s42600-020-00105-4>