iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.34958

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

A Time Series Forecasting with different Visualization Modes of COVID-19 Cases throughout the World

D. Naga Jyothi[1], G. Mamatha[2]
[1, 2]*Assistant Professor, CSE, CBIT, Hyderabad, Telangana, India.*

Abstract: ***In real world applications, one of the prosperous field of science is time series forecasting due to its recognition though having some challenges in the development of methods. In medical field, time series forecasting models have been successfully used in various applications to predict progress of the disease, measure the risk dependent on time and the mortality rate. However due to the availability of many techniques which excel in each of a particular scenario, choosing an appropriate model has become challenging. When a huge dataset is considered it is obvious that machine learning is the best way to perform predictive analysis or pattern recognition tasks on the data. Before machine learning can be used, the time series forecasting problems should be reframed into supervised learning problems. The purpose of machine learning in this field is also to tackle the different challenges like data pre-processing, data modelling, training and any other refinement required with respect to the actual data. This paper deals with the predictive analysis and various visualization applications for time series forecasting of COVID- 19 patients throughout the world.***

Keywords: Analytics, visualization, prediction, Time series forecasting, COVID-19.

I. INTRODUCTION

The diverse industries like finance, supply chain management, production, inventory planning and medical are using the time series forecasting which is the most useful technique. Time series talks about the series of observations that are gathered in constant intervals of time that can be daily, weekly, monthly, or yearly. Time series analysis involves the development of various models which describe and understand the observed time series about the dataset. This involves interpretations and creating ideas about the given data. The complex processing of current and previous data is done and a best fitting model is formed to predict the future observations. Thus, machine learning proved to be the most effective for pattern analysis and prediction of the sequence in structured as well as unstructured data [1]. One important difference we must know is between prediction and forecasting.

A. Prediction and Forecasting

1) The outcomes of the unseen data are assessed using Prediction. A model is fit to the training samples which forms an estimator called f(z) which can estimate for the new samples z. A sub- type of Prediction is forecasting which makes predictions of the future sample data based on time i.e. the time series data. Thus, the only difference between the forecasting and prediction is the historical dimension. The components of Time series forecasting are

2) Trend is to describe the growing or shrinking behaviour of time series data.

3) Seasonality is to highlight the repeated pattern or cycles of behaviour over time.

4) Noise/ irregularity is to describe the non – methodical aspect of time series deviating from common model values.

5) Cyclic property is to identify the repetitive changes in the time series and its positioning in the cycle.

The methods to implement Time series forecasting can be classified as

B. Classical Methods

1) *Naïve Model:* For naïve forecasts, all forecasts are set to be the value of the last observation. In many cases, Naïve models are applied as a random walk (the last observed value is used as a unit for the next period forecast) and seasonal random walk (with a value from the same period of the last observed time span which is used as a unit of the forecast).

2) *ARIMA/ SARIMA:* ARIMA stands for Autoregressive Integrated Moving Average model used to build a composite model of the time series. AR, Autoregression uses the dependency relationship between the observation and few other lagged observations. I, integrated means the use of differentiation between the raw observations to make the time series "stationary". MA, Moving Average uses the dependency between an observation and residual error. SARIMA stands for Seasonal Autoregressive Integrated Moving Average to broaden the application of the ARIMA by including a linear combination of seasonal past values and forecast errors.

3) *Linear Regression Method:* Linear regression is the statistical technique mainly used for predictive modelling by supplying an equation of independent variables, on which our target variable is constructed upon.

C. Machine Learning Methods

1) *Multilayer Perceptron (MLP):* MLP model infers the multi-layered structure of the primary layer of the network which takes in an input, a hidden layer of nodes, and an output layer used to make a prediction and to solve complex problems.
2) *Recurrent Neural Networks (RNN):* RNN's are for sequential data and due to its internal memory, it is very precise in predicting the future data. RNNs are basically neural networks with memory that can be used for predicting time-related targets and can remember the previously captured state of an input to make a choice for the future time-step.
3) *Long Short-Term Memory (LSTM):* LSTM cells (special RNN cells) were developed to find the solution to the issue with gradients by presenting several gates to help the model decide on what information to mark as important and what information to ignore.

D. Time Series Forecasting Process

The time series forecasting process follows the following steps.

1) Problem definition
2) Gathering and exploration of Data
3) Preparation of Data
4) Application of Time series forecasting method.
5) Evaluation/validation

Before discussing the details, primarily the problem should be defined properly identifying the details of domain of forecast operation including important terms, keys and models relating to specific domain. After defining the basics, it becomes clear about the data to be collected, applying techniques of building the graphs and visualization charts. Now, it reaches a level of data exploration and estimates the pivots and trends for further evaluating the variations. Next, the data need to be cleaned for relevant insights and further deducting the variables of importance. The key aspects of data preparation process are targeting the areas of knowledge about the domain that are crucial for designing the new features in the existing dataset. Based on the data preparation and exploratory analysis, the next step is to work with several models and ensure that we choose the best model and consider the variables which are essential for the forecasting process. The evaluation step considers the optimization of the forecasting model parameters and attain high performance.

Fig.1: Time series forecasting

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

II. SIMULATION RESULTS

Time series forecasting can be implemented using python. The various aspects that can be dealt are:

A. What makes Time Series Special?

Time series is a collection of data values which are indeed collected at constant time intervals. To forecast the future, the data values are analysed to find the long-term trend. A Few points by which we can say that time series is different and special from normal regression problems are the **time dependency** and **seasonality trends** (for example if we consider the sales of cotton clothes, it would be invariably high in summer season). Firstly, we started with loading a Time series (TS) object in Python. We have used COVID-19 dataset.

B. Loading and Handling Time Series in Pandas

Pandas has dedicated libraries for handling TS objects, particularly the **datatime64[ns]** class which stores time information and allows us to perform some operations really fast. Considering the arguments :

1) Parse_dates – This column specifies and contains the date – time information(ex: column name is 'Month')
2) Index_col – the main idea of using Pandas for TS data is the index should be the variable depicting the date-time information. This tells pandas to use 'Month' as index.
3) Date_parser: This specifies a fuction to convert an input string into datetime variable. Default format in pandas is 'YYYY-MM-DD HH:MM:SS'. If the data is not in this format it would be converted using date_parse function.

C. Forecasting a Time Series

Starting with reading the data from the dataset available:

```
df = pd.read_csv(path)
df.head()
```

S.no	Province/ State	Country/ Region	Lat	Long	Date	Confirmed	Deaths	Recovered
0	NaN	Afghanistan	33.0000	65.0000	1/22/20	0	0	0
1	NaN	Albania	41.1533	20.1683	1/22/20	0	0	0

This way six rows are displayed.

```
df = pd.read_csv(path,parse_dates=['Date'])
df.head()
```

Here the parse_dates would convert the columns which are in the dataset to date format wherever required.

S.no	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered
0	NaN	Afghanistan	33	65	1/22/20	0	0	0
1	NaN	Albania	41.1533	20.1683	1/22/20	0	0	0
2	NaN	Algeria	28.0339	1.6596	1/22/20	0	0	0
3	NaN	Andorra	42.5063	1.5218	1/22/20	0	0	0
4	NaN	Angola	-11.203	17.8739	1/22/20	0	0	0

df.info() – to display the data types of the columns in the dataset.

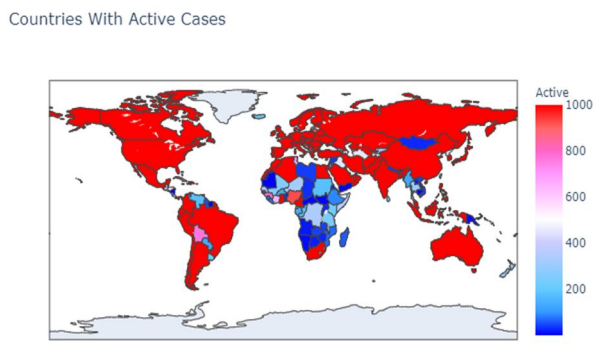
#	Column	Non-Null Count	Dtype
0	Province/State	7600 non-null	object
1	Country/Region	24890 non-null	object
2	Lat	24890 non-null	float64
3	Long	24890 non-null	float64
4	Date	24890 non-null	datetime64[ns]
5	Confirmed	24890 non-null	int64
6	Deaths	24890 non-null	int64
7	Recovered	24890 non-null	int64

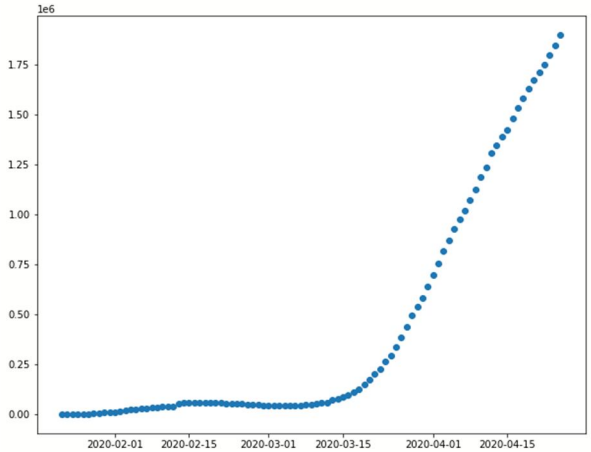
```
active = df['Confirmed'] - df['Recovered'] - df['Deaths']
df['Active'] = active
df.tail(10)
```

	Country	Lat	Long	Date	Confirmed	Deaths	Recovered	Active
24880	Botswana	-22.329	24.6849	2020-04-25	22	1	0	21
24881	Burundi	-3.3731	29.9189	2020-04-25	11	1	4	6
24882	Sierra Leone	8.46056	-11.78	2020-04-25	82	2	10	70
24883	Malawi	-13.254	34.3015	2020-04-25	33	3	4	26
24884	United Kingdom	-51.796	-59.524	2020-04-25	13	0	11	2

The various modes of visualization can be shown as follows

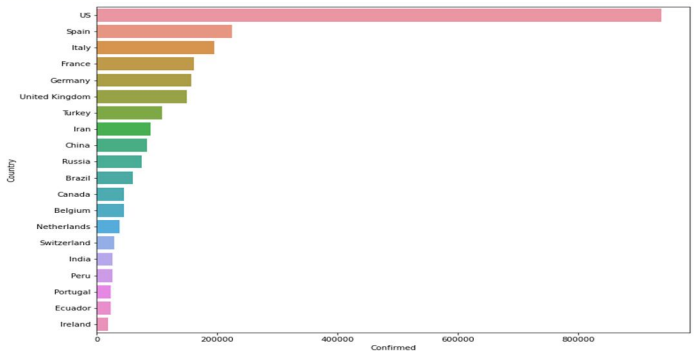
```
#### Plotting On World Map (Active Cases)
figure = px.choropleth(world,locations='Country',locationmode='country names',color='Active',hover_name='Country',range_color
=[1,1000],color_continuous_scale='picnic',title='Countries With Active Cases')
figure.show()
```

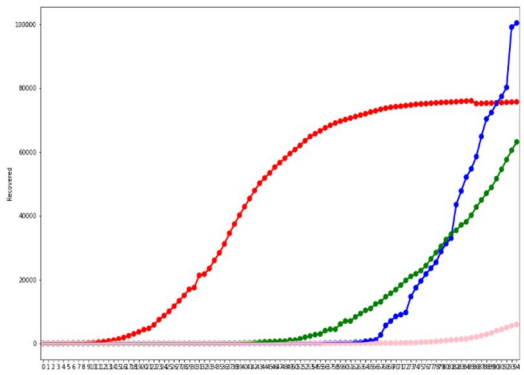




visualization using scatterplot (Active Cases vs date)
plt.figure(figsize=(10,8))
plt.scatter(total_active_cases['Date'],total_active_cases['Active'])
plt.show()
Similarly various other operations can be performed like sorting, checking the active cases for various countries using different visualization plots. Below shown are the bar plot, point plots .

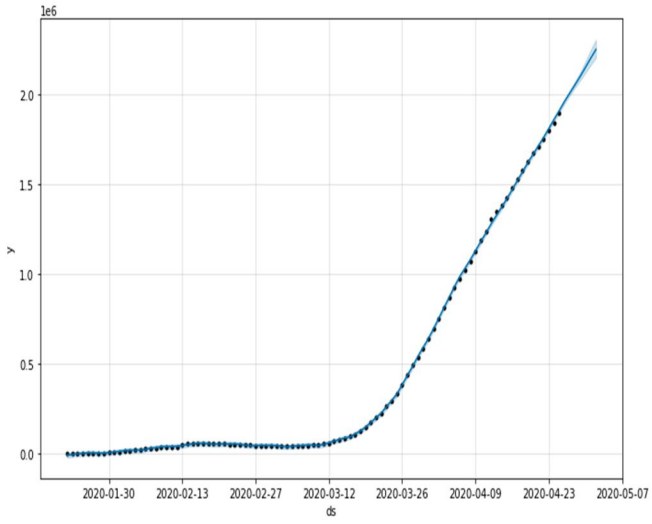
plt.figure(figsize=(12,10))
sns.barplot(top_20['Confirmed'],top_20['Country'])
plt.show()





Finally moving to the time series forecasting in python can be implemented as follows. Here we need to use the library in python .

A library in Python Created by facebook for the time series analysis (forecasting)
We need to build the model, train the model and then find the forecast values.
future = model.make_future_dataframe(periods = 7)
forecast = model.predict(future)
active_plot = model.plot(forecast)



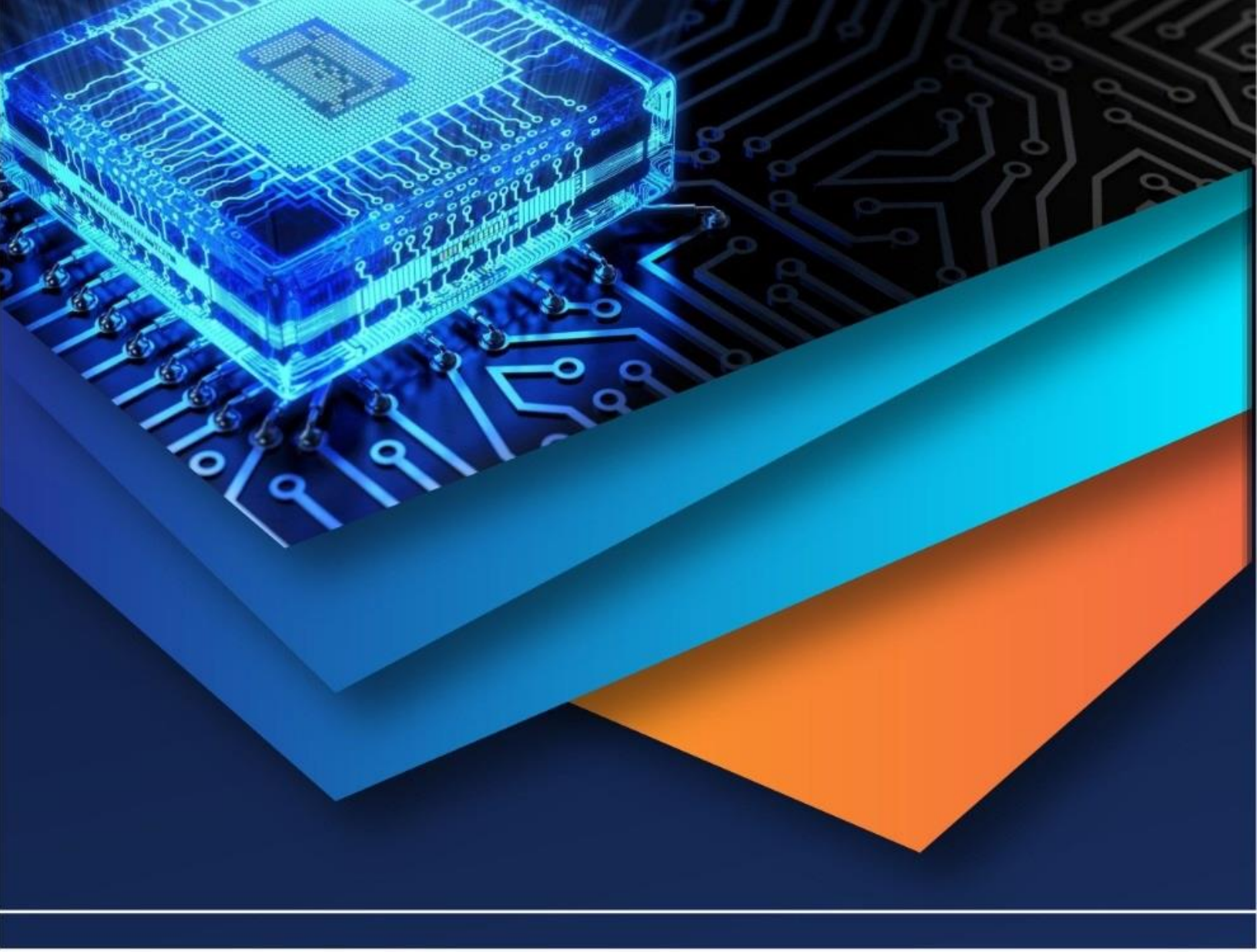
Here if we examine the data set is available only till end of April 2020, but using the time series forecasting we could predict and visualize till may 2020 for which we don't have the values. Based on the trend and seasonality we can model and predict based on time.

III. CONCLUSIONS

As the dataset increases in size there would be higher accuracy of predictions. In case of lack of historical data mainly there would be limitations of using Machine learning associated with the target variable. Thus, lack of data may result in overall decrease of forecasting precision. In general, domain knowledge can also increase the quality of models which will indeed increase the forecasting accuracy. Time series forecasting is similar to any other prediction except that time is involved. Since time is mainly involved the outputs obtained at different times may be different even if the input values are same. In this the training of the model cannot be based on random shuffle. It should follow a sequence. Thus, with the advances in computing techniques and modelling like Deep learning large amounts of data can be handled with time series analysis with greater speed and with higher accuracy.

REFERENCES

[1] https://codeit.us/blog/machine-learning-time-series-forecasting
[2] https://www.datascienceblog.net/post/machine-learning/forecasting_vs_prediction/
[3] Guide to time series forecasting ...musgraveanalytics.com
[4] https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/
[5] https://colab.research.google.com/drive/1iwy-69LunpEmeyLQYAE1TYr9mbdgCuyj
[6] https://blogs.oracle.com/datascience/7-ways-time-series-forecasting-differs-from-machine-learning
[7] https://www.imperial.ac.uk/media/imperial-college/medicine/sph/timeseries.pdf
[8] https://www.frontiersin.org/articles/10.3389/fdata.2020.00004/full
[9] Implemented with the help of workshop attended at IIT Roorkee in collaboration with finland labs.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY