|  |  |
| --- | --- |
| A picture of a winding road and trees  The Application of Predictive Analytics on COVID 19 Cases and Deaths Worldwide | Ragunath Gunasekaran |

## 

Contents

[Introduction 1](#_Toc70036698)

[Problem Statement 2](#_Toc70036699)

[Document Overview 3](#_Toc70036700)

[Scope 4](#_Toc70036701)

[Technical Approach 4](#_Toc70036702)

[Analysis 5](#_Toc70036703)

[Model Deployment 6](#_Toc70036704)

[Testing and Evaluation 6](#_Toc70036705)

[Execution and Management of Project Project Plan 6](#_Toc70036706)

[Project Risk 6](#_Toc70036707)

[References 7](#_Toc70036708)

## Introduction

As of April 19, CDC national forecasts predict 249,000–724,000 new COVID19 cases will likely be reported during the week ending May 15 and CDC national forecast predict 2,100–8,800 new COVID19 deaths will be reported during the week ending May 15. That would bring the total coronavirus cases are 33 Million and deaths are 580K in USA. This forecast and prediction approaches are based for my project to explore COVID19 Cases and Deaths.

COVID-19 was identified in Wuhan, China in December 2019. COVID-19 is caused by the virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a new virus in humans causing respiratory illness which can be spread from person-to-person. On February 11, 2020, the World Health Organization [announced](https://twitter.com/DrTedros/status/1227297754499764230) an official name for the disease that is causing the 2019 novel coronavirus outbreak, first identified in Wuhan China.

The new name of this disease is coronavirus disease 2019, abbreviated as COVID-19. In COVID-19, “CO” stands for corona, “VI” for virus, and ”D” for disease. Formerly, this disease was referred to as “2019 novel coronavirus” or “2019-nCoV.”

COVID-19 is a recent new disease, caused by a coronavirus that has not previously been seen in humans. Because it is a new virus, scientists and this world are learning more each day. Although most people who have COVID-19 have mild symptoms, COVID-19 can also cause severe illness and even death. COVID19 spread and death due to coronavirus is increasing day by day.

Predictive analytics is mathematical and statistical analysis that uses data mining, machine learning algorithms based on historical data, behavior, trend to project the future of our business or outcome.

Python as a programming language has numerous uses such as Machine Learning, AI, mobile applications, statistical , etc. As we know, AI, machine learning, and data analysis/analytics is where it has amassed most of its popularity. Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc.

We need to use four main performance metrics used to evaluate the effectiveness of classification models:

* Accuracy: test’s ability to correctly predict both classes
* Precision: test’s ability to correctly detect positive classes from all predicted positive classes
* Recall (Sensitivity): test’s ability to correctly detect positive classes from all actual positive classes.
* F1 Score: harmonic mean of precision and recall

( Source – Ref 1 )

## Problem Statement

In this pandemic situation, we need to predict pandemic’s behavior within an acceptable degree of uncertainty by establishing when and under which conditions countries can expect increases, peaks, and reductions in new Cases and Deaths (mortality). The Prediction of new cases helps to demand for acute medical services and the prediction help the planning for required health technologies (PPE, ventilators, etc.) and ensure adequate end-to-end supply chain and distribution and managing human resources for an appropriate and timely response. The sad part is death prediction helps to plan for the cremation services and arrangements.

The prediction of the pandemic behavior and outcome alert the society towards prevention and fear to stop the spread.

In this Project, the main goal of the project is to build a predictive model to predict and forecast the number of cases and deaths over the period by using the time series and regression algorithms.

As part of COVID Analysis, we are going to find the solution which can predict and answer the below questions.

Research Questions :

1. Forecasting Confirmed Cases, Deaths & Recovered Cases and Computing Mean Absolute Error (MAE)
2. Predict the Case Fatality Rates and Infection Fatality Rates
3. Prediction for the possible end of new cases of COVID-19 epidemics
4. Support Vector Regression model for prediction of COVID19 cases
5. Evolution of Active Cases, Confirmed Cases, Death Cases and Recovered Cases
6. Calculate Recovery and Death Rates, Deaths per 100k
7. Count of patients : Infected by Virus and Deaths, Distribution

## Document Overview

For this project, I will be modelling various predictive process data related to COVID 19 datasets. The Datasets comes from WHO and Kaggle, which contains 13 columns for Global data and 400 columns for date related data.

1. WHO Dashboard: <https://covid19.who.int/table>
2. WHO Situational Reports: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>
3. WHO Time Series CSV: <https://covid19.who.int/WHO-COVID-19-global-data.csv>
4. <https://www.cdc.gov/coronavirus/2019-ncov/faq.html#Basics>

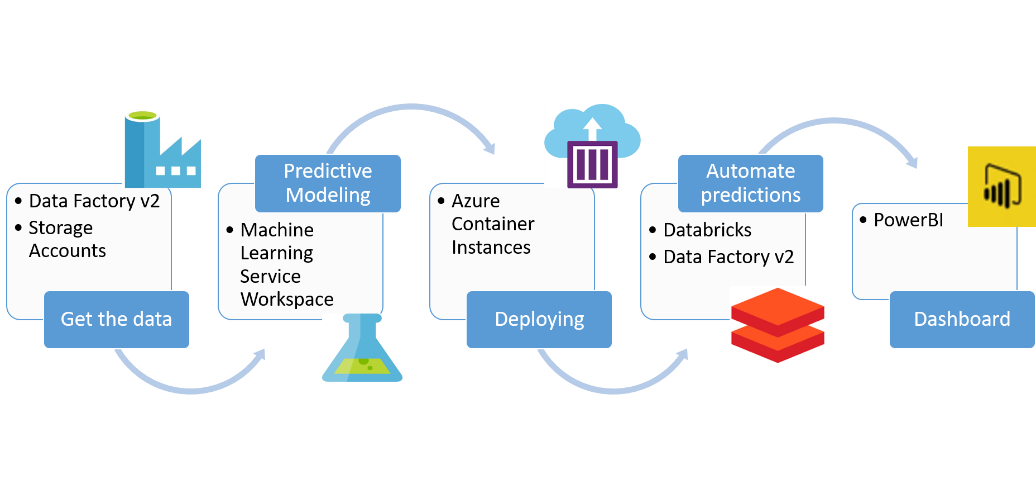
## Scope

The scope of this project will be to analyze the COVID19 data in WHO and Kaggle to identify and predict, forecast the cases and deaths. As part of the COVID 19 analysis, we need to predict cases and deaths by end of August and so on. Also need include vaccination analytics.

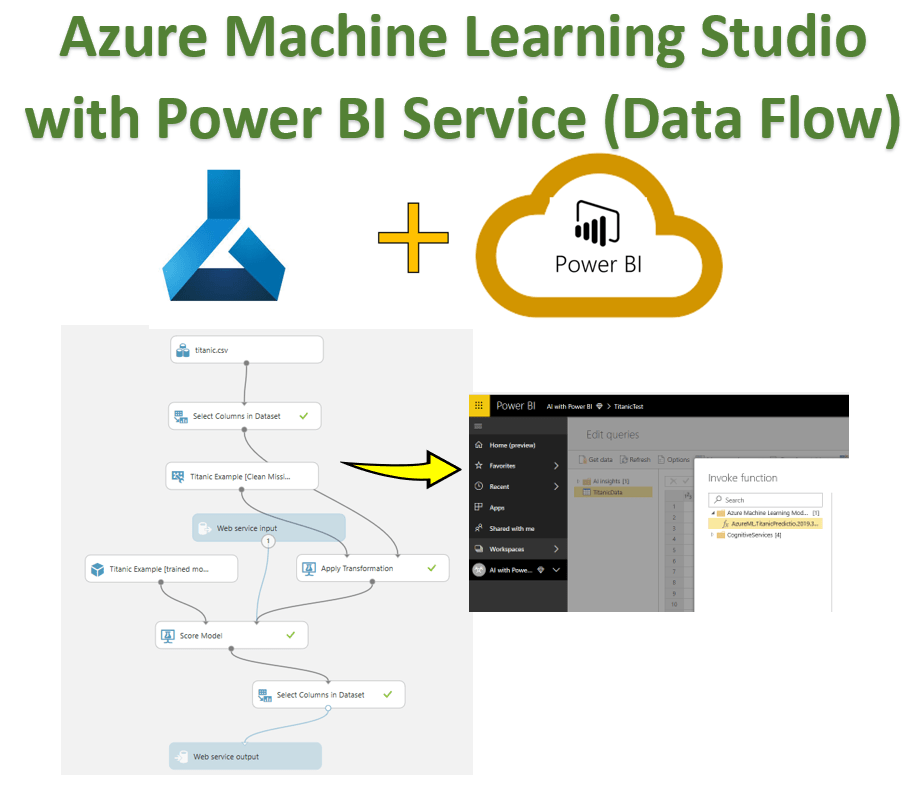
1. Forecasting Confirmed Cases, Deaths & Recovered Cases and Computing Mean Absolute Error (MAE)
2. Predict the Case Fatality Rates and Infection Fatality Rates
3. Prediction for the possible end of new cases of COVID-19 epidemics
4. Support Vector Regression model for prediction of COVID19 cases
5. Evolution of Active Cases, Confirmed Cases, Death Cases and Recovered Cases
6. Calculate Recovery and Death Rates, Deaths per 100k
7. Count of patients : Infected by Virus and Deaths, Distribution

## Technical Approach

As part of Technical Approach, I wanted to use Microsoft Azure ML and Power BI Technologies to automate end to end flow of Azure ML model deployment. The Predicted outcome, we need to use Power BI Dashboard to showcase the results which may help Business leaders to identify easily.



Azure Machine Learning Studio and Power BI integrated



## Analysis

The Source Data comes from various sources, I am planning to perform the EDA ( exploratory Data Analysis) to find any outliers, missing values, handling n/a values, removing duplicates and patterns. Also, clean and preprocess the data before modelling. Since we are dealing with three different domains in the Predictive application study.

As part of Data Analysis, use some Visualization charts, dashboard to view the data at each layer. Also use the probability and Statistics to summarize the data at each variable level and identify the behavior or characteristic of each variables. Summary Statistics should be generated for each attribute or study to define and show the process outcome with detailed information.

As part of Data Analysis, make sure that we have prepared for the Date field population because the Time series forecasting depends on the Date field data.

## Model Deployment

I am planning to run Forecasting and Time Series model to forecast and predict the various count regarding to COVID 19 Spread and deaths and deploy Support Vector Regression model for prediction of COVID19 cases.

## Testing and Evaluation

We will test and Evaluate the models by using training and Testing datasets ( Considering data across and Random data for Training ). By running those models, we can evaluate and identify the Insurance and Credit Card Fraud customer, and anomalies. For model and probability evaluation, we will use various tests and functions.

## Execution and Management of Project Project Plan

I am planning to follow the below steps to execute this project.

1. Define the Project Goal
2. Data Collection – Data Sources
3. Clean Data ( EDA ) and Enrich Data
4. Find insights and Visualize.
5. Model Deployment
6. Iterate the Model.
7. Testing and verification
8. Present the Results for Decision making / Conclusion.

## Project Risk

It is difficult to exactly quantify the amount of uncertainty in a model, but there are specific types of uncertainties that can affect a model’s performance. COVID 19 Prediction or Forecasting can differ because of some action like Government plan, Vaccination, Awareness etc.

## References

1. <https://www.cdc.gov/coronavirus/2019-ncov/faq.html>
2. WHO Dashboard: <https://covid19.who.int/table>
3. WHO Situational Reports: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>
4. WHO Time Series CSV: <https://covid19.who.int/WHO-COVID-19-global-data.csv>
5. <https://www.cdc.gov/coronavirus/2019-ncov/faq.html#Basics>
6. <https://www.cdc.gov/nchs/covid19/covid-19-mortality-data-files.htm>
7. <https://www.kaggle.com/khoongweihao/covid-19-novel-coronavirus-eda-forecasting-cases#Forecasting-Total-Number-of-Cases-Worldwide>