COVID-19: Visualize worldwide situation

Runaj Khatiwada, Shweta Chalise, Udisha Mahaju

Data Science, Data Science, Business Intelligence
Dalarna University
Borlänge, Sweden
h19runkh@du.se , v20shwch@du.se , v20udimah@du.se

***Abstract* —** A newly emerging infectious virus that causes COVID-19 has been detected in December 2019. An outbreak of pneumonia was detected in Wuhan (Hubei, China) whose cause was not detected initially. The cause was quickly determined to be novel coronavirus, namely severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus has spread worldwide causing death to huge number of people. It has hindered the lifestyle of people all around the world. In this project, we used statistical learning method to analyze the pandemic characteristics of all cases reported in the world as up-to the date and predict how long does it take to recover from the pandemic. Supervised learning method such as linear regression has been implemented in our project. The outcome of our project can play a vital role for the government of any country to take the safety measures in order to prevent the COVID 19.

Keywords – COVID-19, Corona Virus, Pneumonia, Pandemic.

I. INTRODUCTION

In December 2019, an outbreak of pneumonia was detected in Wuhan (Hubei, China) whose cause was not detected initially. The cause was quickly determined to be novel coronavirus, namely severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The virus has spread worldwide causing death to huge number of people. Its pathogenesis and proliferation pathways are still unknown. This is the reason why there is still no vaccine or definitive treatment of this virus. It has been affecting 213 countries around the world. There are 4,894,278 reported cases to the date 19 May 2020 among which 320,198 have lost their life [2]. Its spread has hindered the lifestyle of people all around the world.

In this study, we analyze the pandemic characteristics of all COVID 19 cases and derive the conclusion on how long it takes to flatten the curve of new infection cases of the given country. The user should set the thresholds in order to complete the task. For e.g. if the user define the criteria as 100 new cases in a particular day with 5 new deaths for the particular country, our program will predict the number of days required to meet the criteria from the date of execution.

This project is very helpful for the government as well as for the people to take the safety measures like maintaining the social distance, formulating the necessary health laws, providing the instruction and preventative measures. In order to complete the task, we use the various libraries and modules from python programming language. Pandas is a python library that has numerous data structures and tools for working with structured data sets. These are common to statistics, finance, social sciences, and many other fields. The pandas library provides integrated routines for data manipulations and analysis on the available data sets. It aims to be the foundational layer for the future of statistical computing in Python. It has been a complement to the existing scientific Python stack while implementing and improving over the data manipulation tools found in other statistical programming languages [3].

Furthermore, pandas have many features that can be implemented in our project. Python provides different modules for data visualization and matplotlib is one them. Matplotlib is one of the best modules for the visualization of machine learning techniques such as regression, classification, clustering and so on. This module provides various functions for plotting the data based on the requirements. The plot may be either line, graph, or area [4]. Tkinter is a library that provides a fast and easy way to create Graphical User Interface (GUI) application. It is the standard GUI library for python. In our project, we have used the tkinter so that the user finds it convenient to use the program and choose the record of the desired country.

II. PROJECT MOTIVATION

The motivation of this project is to determine how it has affected the global communities. Some countries have taken the strong measure to control the spread of virus by total lockdown whereas some countries are against the lockdown. Therefore, it is necessary to understand, how the virus can be controlled or how long does it take the world to be in usual state and what will be the aftereffects of the virus.

This project is to develop a prototype for the prediction using Data Mining modelling techniques and pandas in python to discover and extract hidden information (patterns and relationships) of COVID-19.This project will enhance the visualization and case of interpretation to display the result in tabular and graphical forms. The data source for this project is obtained from https://api.statworx.com/covid.

It provides the data of people infected and dead daily. The project consists various phases. Data understanding phase focuses on understanding the objectives and requirements from a health perspective and designing a preliminary plan to

achieve an objective as it uses the raw data to identify its quality and subsets. Data preparation phase constructs the final data set that will be fed to a modelling tool. This includes table, record, attributes, selection, data cleaning and transformation. The modelling phase selects and implies various techniques to calibrate their parameter to optimal values. The evaluation phase evaluates the model to ensure that it achieves its objectives in a productive and active ways.

III. About Data Source

The data source for Covid19 is obtained from the free API provided by statworx [5],[6]. Statworx is business management consultant company in Frankfurt, Germany. This API uses official data provided by the European Center for Disease Prevention and Control and delivers a clear and concise data structure for further processing and analysis.

This API need a post request from any programming languages, for python we need requests [16] and json [17] library to post the request by passing the request json. After posting the request, we receive the response in json format [6].

If we are using R, we need httr and jsonlite libraries to push the request and receive response. This API useful for our project and our detail analysis regarding COVID-19[6].

IV. Process used

Multiple processes of Business Intelligence are used in this software. All these theories are described as follows:

A. Data mining

Data mining is the process of converting the raw data into useful information [7]. In this case study, we have collected the daily data that represent of the COVID 19 cases for different countries. We used this data for visualizing the status of the virus in particular country. Along with that, we have used the techniques of data mining as regression to forecast some useful outcome that can help the governments of different countries to decide the further action to be taken based on regression.

B. Structured query language, SQL

SQL is a domain specific programming language, and it is used to manage and process the data, which are stored in relational database system (RDBMS) [19]. This language is only used for structured databases, which has a fixed schema for storing the data such as tables in rows and columns representation. [8]

C. Data Warehousing

Data warehousing is the method by which data capacity is built and utilized. Data capacity is made by consolidating data from a few heterogeneous sources that permit expository announcing, organized and/or ad hoc inquiries and decision-making exercises. Data handling includes framework security, framework creation and data disposal. In this project, we have extracted the data from the source, and transferred it into the MS SQL server. [9]

D. Pyplot

Pyplot or matplotlib.pyplot is a plotting library which is used for graphical illustrations in python. We have used this library to plot the graph in different ways like line, area and bar. These graphical representation helps the user to observe the position of the data.[10]

E. Linear Regression

The linear regression [11] is a linear approach to model the relationship between a response variable and predictors. For example,

$$Y = a + bX \quad (1)$$

Where,

X is the explanatory variable
Y is the dependent variable
b is the slope of the line
a is an intercept

In this project, we have the dataset containing the variables like date, daily cases, daily deaths, etc. So, for our linear regression, the output or the dependent variable is number of days, which is a dummy variable added in the data frame which is the day count from the start date when the infection started. The predictors are daily cases and daily deaths. So, the equation for linear regression for this is,

$$Y = a + b1.X1 + b2.X2 \quad (2)$$

Where,

X1 is the daily infected cases
X2 is daily death cases
Y is number of days

V. METHODS

The method we have used for this project is Data Mining. Data mining is the process of reading a huge amount of data and watch the pattern of data to derive some valid conclusion. In other words, Data mining is a method of extracting data from various sources and summarizing it to the appropriate details. The primary aim is to identify similarities or trends between hundreds of fields in vast datasets.

Data mining can be used in marketing process to exploit progressively broad datasets and boost customer segmentation. It can be used in banking process, to minimize the market risk and for the better competition in the market. Data mining can be used in medical sector and one of the best examples is for analyzing the data of COVID 19 in particular country to decide the upcoming action for their people. Our

project includes the same process. The various stages in our project, is shown by the following figure.

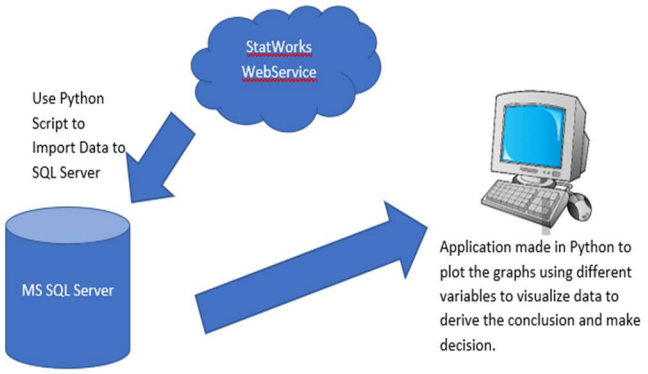

Fig.1: System Prototype

The methodologies comprised in our project are further described as:

A. Choose the external data source

For this project, we need the daily data for total cases, new cases, total deaths and new deaths caused by COVID – 19. We have chosen API provided by StatWorx. It is a statistical consulting firm with expertise in health research. This provides the daily updates for COVID – 19 cases. When we post request on the URL (https://api.statworx.com/covid) for the api we receive the dataset in the form of json.

B. Perform ETL process

ETL [13] process is major part of data mining process. At first, data extraction is done. In this project, we send a request to API provided by statworx. For this, we need to pass the filter in the json format, we have chosen the filter value as ISO codes for the countries. We have chosen those countries where there is very large number of corona infected people as well as deaths. The list of those countries and their ISO codes are kept in MS SQL table under the application database named COVID19. We have run a loop in python to send the request for each country which are kept in data frame fetching it from SQL table. There is no need of authentication for this. After sending the request, we receive the response in json format, which is converted into pandas data frame. These data in data frame is stored in the MS SQL table named *covid_case_details.*

This task is done on daily basis, to run this python script, we have created a executable batch file, on executing the file, the python script is executed and data for that day is inserted in the SQL table. In order to get uninterruptable execution on daily basis, a daily scheduled task is defined in windows system where the software and database server is hosted. The task is scheduled by the default feature of windows 10, Task Scheduler where the schedule is made to run at 11:45 PM every day. Thus, we can have our new data for the current day on daily basis.

C. Create a GUI for the application

The major part for the visualization here is a software made with GUI. We have designed a GUI by the help of python inbuilt library, Tkinter [14]. This allows us to create a window canvas with desired geometry and we can add different fields like text box, combo box, buttons, labels, etc.

In Fig. 2, we can see the GUI of the software. There are 3 combo boxes where we can choose the values. The first combo box is Country, on clicking the combo box, we see the list of countries where there are maximum cases of COVID 19 patients, among those options we can choose one of the desired country to see the condition of COVID 19. The second combo box is the Data Options, that means if we want to see the daily cases/deaths against the date or cumulative daily cases/death. There are two options Daily and Total. Along with that, we can see another combo box from where we can set the graph options, for this we have kept the options as Line plot, area plot and bar plot.

Below those combo boxes we can see three buttons, Plot, Exit and Forecast. After choosing all the parameters we can click on the plot button, where we can see the plot of given inputs.

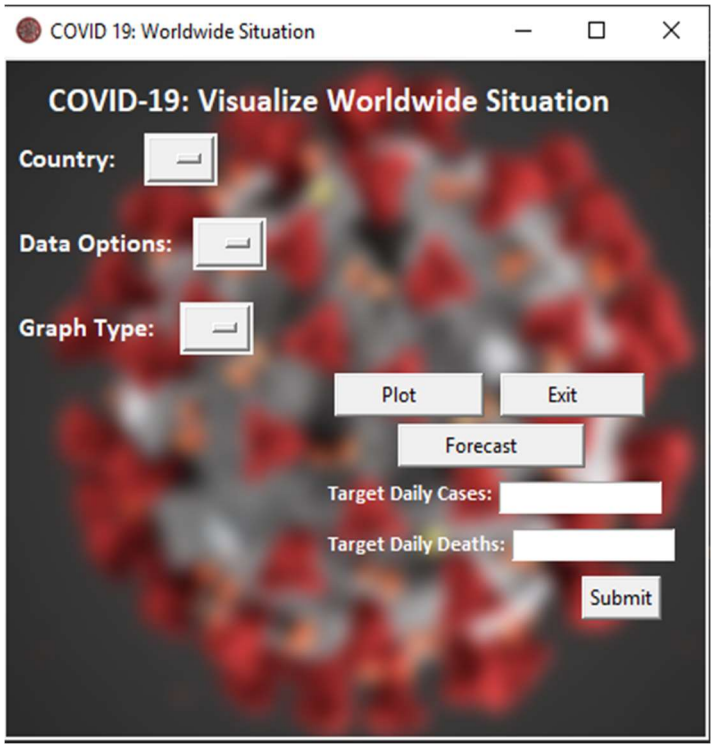

Fig. 2: Screen shot for forecasting software

D. *Fetching the data from SQL and plot in graph*

A major part of our software is to plot the data and forecast the data. To forecast the data the first thing is to get the data source. We already have extracted, transformed, and loaded our data into the MS SQL server that is our data warehouse. For this, we need the python library named pyodbc [15], which should be imported while coding. By the help of this library we can connect to the MS SQL database. In this software we have used context connection, for this we do not need the username and password rather it gets connected by windows authentication. After connecting the database, we hit a query for selecting the table named covid_case_details and created a data frame for all the data in the table, which is the sorted by the date. After getting the data frame, we provide the input through the combo boxes. We choose the plot button; thus, we can see the plots.

E. *Forecasting the date to meet the target*

This is the key feature of our software, where we will receive the predicted date for the given target cases and deaths for any particular country. The concept behind this is the linear regression method. The data frame obtained from SQL table, consists of different columns date, cases, deaths, country, cases_cum (cumulative cases), deaths_cum (cumulative deaths). To forecast the date for target cases, we have used the existing data as training data. For the training data we have chosen daily cases, and daily deaths as predictors and we have added one dummy variable called number of days, which is a counter integer from the beginning of the day when corona cases started to appear i.e. 2019-12-31 in our dataset. This dummy variable is set as output or response variable. To create a linear regression model in python [12] we need sklearn library [18], from where we have to import linear_model. After creating the model, we use this model to predict the output variable. There are two text boxes to set target values, after receiving those values from user, those are converted into data frame as testing data, and use predict function from the model to predict the output variable. The output is compared with the value of number of days from the current day. If the output is less than the number of days from the current day, then we set the days remaining to reach the target is the difference between number of days from the current date. And if number day on current day is less than output then, then we set the days remaining to reach the target is the difference between output value and number day on current day. After calculation, we display the plot and forecasting message as an output. (The examples and result are discussed on results section.)

VI. SOFTWARE DEMONSTRATION

Our project is targeted to the general people not only the technical expert. That is why we have tried our best to make it user-friendly. We have designed a Graphical User Interface (GUI) so that all the users find it very convenient to use. When the user runs the program, they will get an interface with various options. The user can select desired country whose data they want to visualize through a drop box. After selecting the required country, they can choose the date. We have also provided the options whether they want to visualize the data in the form of line graph or area graph or the bar graph.

On the other side of the GUI, user can set the thresholds as "Target daily cases" and "Target daily deaths". When they have entered all the required information, they have to choose the option "Plot". The program will forecast the result based on the previous records. It will perform the linear regression against the total new cases and the new death cases on the current date and plot the relationship. The final output of the program indicates the time period required to reach the threshold defined by the user. The user has to select the "Exit" option in order to terminate the program.

(The demo video is attached along with the report, which makes easier to use the software)

VII. RESULTS

We have used the software to forecast target cases as 100 cases per day and 10 deaths per day in Sweden as an example. To visualize this, we have plotted a line graph to see the condition of corona infected people and deaths.

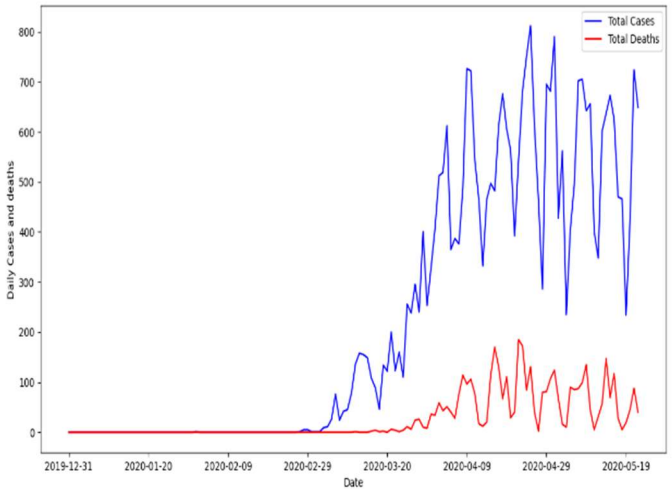

Fig. 3: Sweden Daily Cases and Daily Deaths

From Fig. 3, we can see the daily cases and daily deaths, in Sweden where blue line represents the daily cases and red lines represents daily deaths. On analyzing graph, we can see the daily cases are fluctuating, someday cases are high around 700 and someday it reduces to approximately 300 cases. Likewise, we can see similar type of pattern in death case as

well. Someday death cases rise to 180 to 200 and someday it reduces below 10.

On visualizing, the total cases till the current date from the beginning; we can see the following visualization:

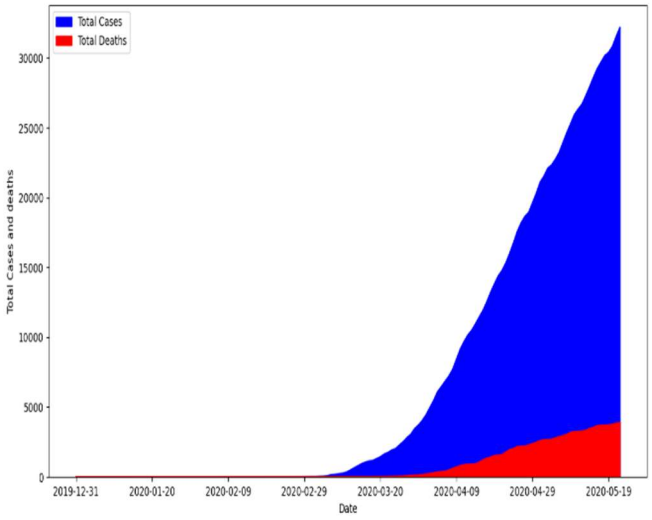

Fig. 4: Sweden total cumulative cases till 2020-05-23

Providing the input on the forecasting parameters, target daily cases and target death cases, we get the following outcome:

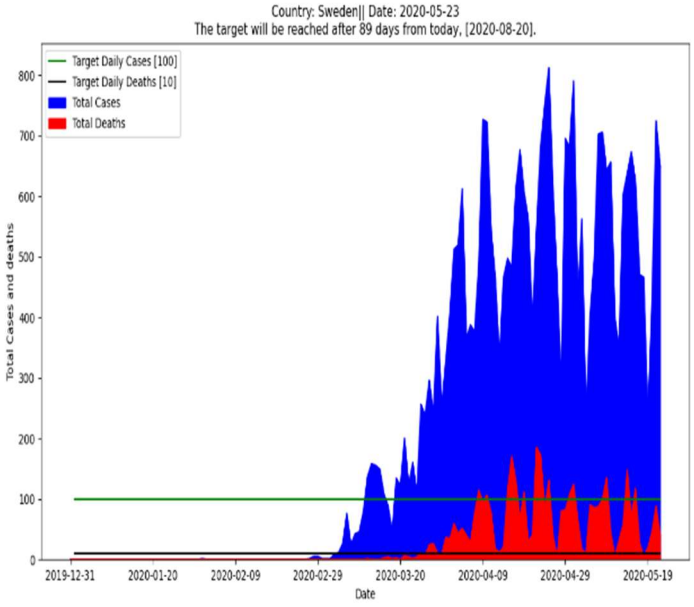

Fig. 5: Forecast Date for target cases and deaths

In Fig. 5, we can see the area plot for Sweden daily cases and daily deaths due to corona virus, where blue area represents the daily cases and red area represents the daily death cases. Along with that, we can see two straight lines, representing the targets for daily deaths and daily cases of corona infected people. At the top of the plot, a forecasting message is displayed, which states that, the target of getting daily 100 cases and 10 deaths will be achieved after 89 days from 2020-05-23, that means the target will be reached on August 20, 2020 in case of Sweden.

VIII. CONCLUSION

COVID-19 has clearly become an international public health problem. Due to its rapid transmission, all the countries should be attentive towards the disease surveillance system. They should scale up the readiness and the response operations.

" Prevention is better than cure." For individuals to prevent themselves from COVID-19, washing hands every time they touch anything outside, using facemask to minimize the transmission through air, sanitizing the product they buy from stores, not touching their face without proper hand wash are some safety measures.
On the other hand, if the large-scale community if at risk, mitigating social gathering, educational institute closure for temporary period, strictly supervising the symptomatic individual providing essential life supports for them, personal hand hygiene, wearing facemask can be enforced by the government body. If the case is critical, the government can even lock down their country to slow down the spread of COVID-19.

Our analysis was mainly focused on Sweden, the victims are rising rapidly day by day. The number of deaths is also up surging. The situation can go worse when proper care is not done. Not only the government, but also everyone is responsible for the prevention of this pandemic. Our prediction shows that it will take a long duration in order to get back to the normal state. Our prediction shows, Sweden is expected target cases and target death to be on August 20, 2020. Once the disease is gone, it is not guaranteed that it will not have the second wave. However, the spread rate can be decelerated by various measures. Thus, we suggest the people and the government body to become more cautious about the situation until there is any medicine or vaccine against the disease.

IX. ABBREVIATIONS

- COVID-19: Corona Virus Disease 2019.
- SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
- SQL: Structured Query Language
- RDBMS: Relational Database Management System
- GUI: Graphical User Interface

- WHO: World Health Organization
- MS SQL server: Microsoft Structured Query Language
- API: Application Program Interface
- ETL process: Extract Transform Load process.
- ISO: The International Organization for Standardization (in this case ISO code for country)

X. ACKNOWLEDGMENT

The success and outcome of this project required a lot of guidance and assistance of many peoples. We are privileged to receive all these help throughout the entire project.

We would like to thank to our course instructor Hasan Fleyeh for his guidance and support. He gave the idea of different concept used in this project. Also, thanks to our lab instructor Serena who guided us technically to use data frames, pyplots and use of different modellings like regression, clustering, classification, etc.

XI. REFERENCES

[1] WHO. WHO statement regarding cluster of pneumonia cases in Wuhan, China. Jan 9, 2020. https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-Wuhan-china (accessed April 16, 2020)

[2] "Coronavirus Cases:" *Worldometer*. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 19-May-2020].

[3] Mckinney, Wes. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Performance Science Computer.

[4] Srinivasa K.G., G. M. S., H. S. (2018) Getting Started with Visualization in Python. In: Network Data Analytics. Computer Communications and Networks. Springer, Cham

[5] Statworx: https://www.statworx.com/de/academy/

[6] API Description: https://www.statworx.com/ch/blog/making-of-a-free-api-for-covid-19-data/

[7] Data Mining: https://www.geeksforgeeks.org/data-mining/

[8] Structured Query Language, SQL: https://docs.microsoft.com/en-us/sql/odbc/reference/structured-query-language-sql?view=sql-server-ver15

[9] Data warehousing: https://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm

[10] Pyplot (Matplot lib): https://www.edureka.co/blog/python-matplotlib-tutorial/

[11] Linear Regression: Sharda, R., Deel, D., & Turban, E. (2014). Business intelligence and analytics: decision support systems (Tenth edition, Global edition.). Harlow, Essex: Pearson. ISBN: 1292009209

[12] Linear Regression in Python: https://www.geeksforgeeks.org/linear-regression-python-implementation/

[13] ETL Process: https://www.guru99.com/etl-extract-load-process.html

[14] Design GUI by Tkinter: https://www.geeksforgeeks.org/python-gui-tkinter/

[15] PyODBC (Connect MS SQL database in python): https://github.com/mkleehammer/pyodbc/wiki

[16] Requests (get and post) Library in Python: https://www.geeksforgeeks.org/get-post-requests-using-python/

[17] JSON library in python: https://www.w3schools.com/python/python_json.asp

[18] Scikit-learn (sklearn) in python: https://scikit-learn.org/stable/tutorial/basic/tutorial.html

[19] RDBMS: https://en.wikipedia.org/wiki/Relational_database