Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765847

COVID-19 IMPACT ANALYSIS USING DATA VISUALIZATION AND PREDICTION

[1] Mr.S.Mohan, [1]Dr.G.Victo Sudha George, [1]Dr.P.Dinesh Kumar,
[2] Mr.P.Sankar Ganesh, [2] Mr.Surampudi Lokesh Ratna Teja, [2] Mr.T.Sree Hari
[1]Professor, Department of CSE, Dr.M.G.R. Educational and Research Institute.
[2]Student, B.Tech CSE, Department of CSE, Dr.M.G.R. Educational and Research Institute.

*** Corresponding author:**
Mail id – mohan.cse@drmgrdu.ac.in

Abstract: Throughout history, the world has faced various crises and pandemics, but the resilience and innovation of humanity, along with their ability to think creatively and unconventionally, have helped them to overcome these challenges. Currently, the world is grappling with the COVID-19 pandemic, caused by a deadly new coronavirus that originated in Wuhan, China and has since spread globally, infecting millions of people. However, difficult times require stringent measures and innovative solutions. As such, the purpose of this research is to analyze COVID-19 data using visualization techniques to map and compare the outbreak in different continents and territories such as the USA, China, India, Italy, and Taiwan. This analysis utilizes metrics such as the total number of confirmed cases, casualties, and recoveries for each region, and employs data visualization to aid in comparison. Ultimately, this research aims to provide a comprehensive understanding of the impact of COVID-19 across the globe through the use of Data Visualization and Machine Learning.

Key Words – Matplotlib, machine learning, random forest, data visualization, python

I. INTRODUCTION

The COVID-19 pandemic has unleashed a global health crisis that has caused an unprecedented level of distress and fear around the world. The outbreak of the novel coronavirus was first detected in Wuhan, China in December 2019 and reported to the World Health Organization (WHO) shortly thereafter. With an incubation period that varies among patients, the virus has rapidly spread across the globe, leading to its declaration as a global pandemic by the WHO in March 2020. By mid-April, the number of active cases worldwide had surpassed the 1 million mark, and the mortality rate had been estimated by the WHO to be 3.4% as of early March. The symptoms of COVID-19 range from flu-like symptoms such as fever and cough to more severe symptoms like shortness of breath and loss of smell. The elderly and those with pre-existing medical conditions are at a higher risk of severe illness and death due to COVID-19. The pandemic has wreaked havoc on healthcare systems and economies around the world, leading to lockdowns and travel restrictions in many countries. To better understand the impact of COVID-19 across the globe, we will analyze various data points, including the total number of cases, casualties, recoveries, lockdown dates, and tests conducted in different countries and regions. We will also compare COVID-19 to previous epidemics like SARS, MERS, and Ebola, evaluating the severity, mortality rate, and contagiousness of each disease. By analyzing this information, we hope to gain insights that can help us mitigate the impact of COVID-19 and prevent future pandemics.

III. RELATED WORK

Numerous research have examined how the COVID-19 epidemic has affected financial markets. Salisu and Vo (2020) studied whether healthcare information found via Google searches might be leveraged to forecast stock returns employing info from all the nations impacted by the epidemic as well as the territories that had the most fatalities during the outbreak. Employing wavelet-based Regression analysis and fidelity wavelet tests, Sharif et al. (2020) examined the time-frequency relationships between price of oil, the COVID-19 epidemic, economic uncertainty, political turmoil, and the share market in the U.S. As a result of the coronavirus upsurge, Liu et al. (2020) and Khan et al. (2020) examined the brief stock exchange indexes of the most impacted nations. In several research, emerging economies in the Arab World, Latin America, and Central and Eastern Europe have experienced stock market unpredictability (Anser etc. 2021; Salisu and Obiora 2021). Several research has mainly concentrated on preventing stock market crashes as a result of economic volatility during the pandemic, investigating the broader economic criteria of Montenegro using a Bayesian VARX approach, and the effect of the COVID-19 flare - up on Chinese-listed tourist revenue shares (Wu, et al., 2021). (Dai etc. 2021). To date, research on our issue has only revealed data visualisation of data sets using python libraries in the Jupyter platform, and their findings have helped to understand the covid illness and its prevalence throughout the globe. They employed 16 attributes in their datasets, including GDP, stringency, total deaths, etc. for that, but the main limitation of their work was that it could only be used for display.

IV. PROPOSED SYSTEM

Our team conducted research on the COVID pandemic using a dataset from the Kaggle website. We used this information to understand how it is affecting different parts of the world, and we will use our findings to demonstrate its global impact. Furthermore, we predicted which countries would have stricter regulations regarding health care based on their respective situations. We employed three prediction algorithms--the XG boost method, the support vector machine, and the random forest algorithm--to create our results.

V. METHODOLOGY

This study focuses on analyzing data related to the COVID-19 pandemic from the time of its emergence until January 1st, 2020. The information has been obtained from balanced periodic panels and official government websites. The study aims to examine various factors such as the duration of lockdown periods, mortality rates, and contagiousness, to gain insights into the varying impact of the virus on different regions around the world. The two key aspects that have been considered in this analysis are the source of data and the representation of the data. By utilizing reliable data sources and appropriate data visualization techniques, this study seeks to provide a comprehensive understanding of the COVID-19 pandemic and its impact on different parts of the world..

1. Architecture Diagram :

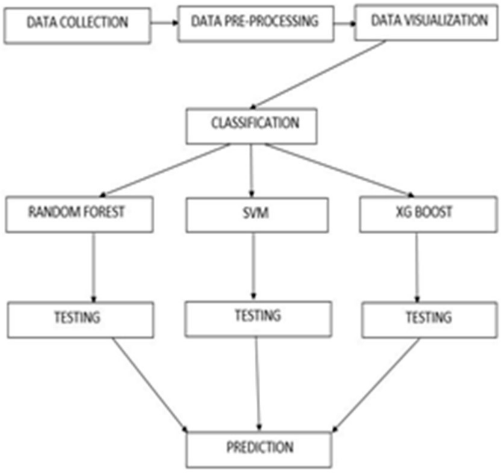


Fig 1: Architecture Diagram of Proposed System

We have described the research work flow in the diagram above. First, we begin by collecting data from websites. Next, the data proceeded into the pre-processing step, where we pre-processed it using a variety of techniques to create transformed data from the raw data. Following that, we checked the data using plot diagrams for data visualisation. After preparing the data, we will use the three classifiers Random forest, SVM, and XG boost to assess the clean data. While we analysing the model, we splitted the data into test and train. we have used the trained dataset for training the model and we added the trained data into all three algorithms and after that, we used the trained data for testing purpose and also for predicting the accuracy level of algorithms.

V. MODULES

Here, we have divided the whole process into three modules. They are

1. Data Handling.
2. Algorithms used.
3. Analysis.

1. DATA HANDLING:

1.1 Dataset Collection :

Data is a critical element of any Artificial intelligence system and, essentially, the only cause of the current boom in machine learning's popularity. Scalable ML algorithms are now feasible as standalone solutions that can add value to a business instead of being a by-product of its core operations because to the availability of data. The dataset we are using to analyse the impacts of covid-19 is downloaded from Kaggle. It has data of 170 countries from January 1st 2020 to October 19th 2020 (294 Days). Various sources are utilized to gather divergent data for a more comprehensive and conclusive evaluation, and this approach is widely employed due to the abundance of available datasets.

1.2 Data Pre-Processing :

Data pre-processing is vital for any machine learning or data mining strategy since the performance of a machine learning methodology relies on how well the dataset is prepared and formatted. To achieve that, we followed several steps.

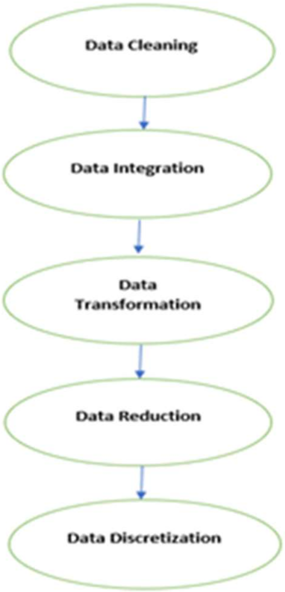


Fig 2: Pre-Processing Stages

During these stages, a Replace Missing Values filter was used to manage missing data, and after that, an Interquartile Range (IQR) filter was used to find outlier and extreme values. To remove noisy values we have used entropy-based binning method. The continuous or numerical variable is classified using the entropy-based binning algorithm when the majority of values in a bin or group correspond to the same class label. Entropy for the target class labels is calculated, and the split is categorised according to maximal information gain. Although this is frequently a component of data cleaning, it is not the only method. The bulk of the effort is put into finding rogue data and (when possible) fixing it. "Rogue data" refers to information that is insufficient, erroneous, irrelevant, corrupt, or improperly formatted. Deduplicating, commonly known as "deduping," is a part of the process. In essence, this involves combining or eliminating similar data pieces. The process of merging data from various sources into a single, cohesive perspective is known as data integration. Integration involves procedures like cleansing, ETL mapping, and transformation and starts with the ingestion process. Analytics technologies can finally create useful, actionable business knowledge thanks to data integration. In several processes, including data integration, migration, warehousing, and wrangling, transformation is a crucial stage. Constructive data transformation involves adding, copying, or replicating data. Destructive data transformation involves deleting records and fields. Aesthetic data transformation involves standardising particular values. Structural data transformation involves renaming, moving, and combining columns.

1.3 Data Visualization :

In this paper, we have utilized statistical analysis techniques to create a descriptive model that encompasses data collection, analysis, interpretation, presentation, and modelling. Our model classifies the COVID-19 pandemic into two categories: country-level and continent-level. While both categories share some common features, each has unique characteristics, and we have drawn distinct conclusions from the data we collected by utilizing inferential analysis, which is a part of statistical analysis.

2. AGORITHMS USED :

2.1 Classification :

2.1.1 Random Forest :

The Random Forest algorithm is employed for both classification and regression. It constructs a decision tree based on the data and makes predictions based on it. It can be applied to extensive datasets and maintains consistency in the output, even if there are null values in the data set. The samples obtained from the decision tree can be preserved and used for other data. The algorithm has two stages: creating a model and predicting using the random forest classifier generated in the first stage. Gini index is used in this classifier to determine the purity of the data. The Gini Index, also known as Gini impurity, quantifies the likelihood that a specific variable would be incorrectly classified when it is selected randomly. It can be referred to be pure if every element is a member of a single class. The Gini Index ranges from 0 to 1, where a value of 0 indicates that all elements stand with a single class or that there is only one class (pure), and a value of 1 indicates that the elements are distributed randomly across the classes (impure). An equal distribution of components within some classes is indicated by a Gini Index value of 0.5.

$$\boxed{\begin{aligned}\text{Gini Index} &= 1\text{-}\ \textstyle\sum_{i=1}^{n}(p_i)^2\\ &= 1\text{-}\ [(p_+)^2 + (p_-)^2\end{aligned}}$$

2.1.2 Support Vector Machine :

In ML, a supervised machine learning approach called the Support Vector Machine (SVM) may be applied to classification or regression problems. However, categorization errors are where it is constantly utilised. Each data point is defined by a points in n-dimensional space (n refers to the number of attributes you have) when using the SVM method, with the value of each feature becoming the value of a unique position. Next, we do classification by point out the hyper-plane that validly delineate the two classes. The projection of two vectors multiplied by the product of two additional vectors is known as the "dot product." SVM's ability to operate on non-linear datasets is one of its most intriguing features, and for this, we employ the "Kernel Trick" to simplify the classification of the points.

2.1.3 XG Boost :

This algorithm uses the gradient boosting decision tree algorithm. The gradient boosting method creates new models that do the task of predicting the errors and the residuals of all the prior models, which then, in turn, are added together and then the final prediction is made. The boosting ensemble technique consists of three simple steps:

1. An initial model predicting the target variable y is how P0 is defined. There will be a residual associated with this model (y – P0)

2. The residuals from the previous phase are fitted to a new model, q1.

3. P1, which is a boosted version of P0, is created by combining P0 with q1. There will be a decrease in the mean squared error from P1 compared to P0.

$$p_1(a) < -\ p(a) + q_1(a) \qquad (1)$$

We may model just after residuals of P1 and generate a new model P2 to enhance the performance of P1.

$$p_2(a) < -p_1(a) + q_2(a) \qquad (2)$$

This might be executed for 'n' repetitions, or till the residuals are as little as feasible.

$$p_n(a) < -p_{n-1}(a) + q_n(a) \qquad (3)$$

3. ANALYSIS:

3.1 Splitting dataset into train and test data :

The act of breaking usable data into two halves is known as "data splitting," and it is most frequently done for cross-validator purposes. A predictive model is developed using one set of data, while the effectiveness of the model is assessed using an other set of data. The most of the information is divvied up for training while very little is used for testing when a data set is split into a two sets.

We must first acknowledge a few parameters before we can utilize the class.

3.1.1 Test size

This option specifies the amount of data that must be separated into the validation set. This is displayed as a percentage. If you specify 0.5 as the value, the dataset will be divided in half and used as the test dataset. If you specify this one, you can neglect the next one.

3.1.2 Train size

If the test size is not selected, this option is necessary. This is analogous to test size, however you tell the class what portion of the dataset to split as the training set instead of how much of the dataset to split as the training set.

VI. DATA VISUALISATION

The current system involves the creation of a descriptive model through the analysis of collected data and Statistical analysis application, which involves the comprehensive process of data collection, analysis, interpretation, presentation, and modelling. For graphical representation purposes, the python3 matplotlib and NumPy modules have been employed. These modules have been utilized to create high-quality graphical representations that effectively portray the data's characteristics and trends in a clear and concise manner. We have given the clear detail view about the spread of covid all over the world and not only that GDP, human index, covid cases, death rates, economy etc.

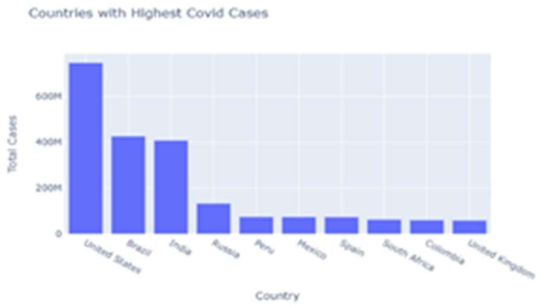


Fig 3: Covid cases

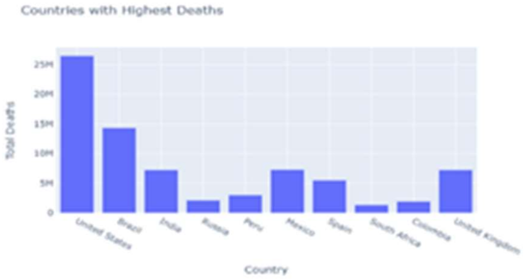


Fig 4: Covid Deaths

From Fig 2&3, we can come to know that compared to all other countries united states got more covid cases and covid deaths and its notifying almost 25M deaths.

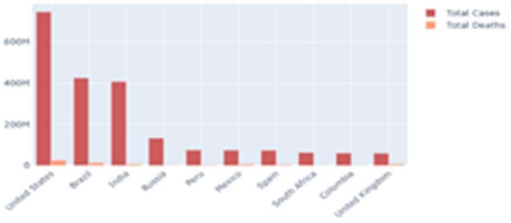


Fig 5: Total cases and deaths

Just like the total number of covid-19 cases, the USA is leading in the deaths, with Brazil and India in the second and third positions. One thing to notice here is that the death rate in India, Russia, and South Africa is comparatively low according to the total number of cases.

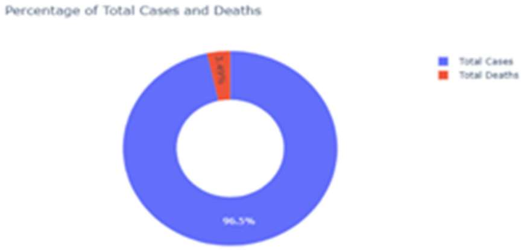
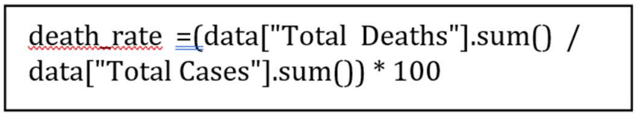


Fig 6: Percentage of cases and deaths

To know about the death rate, we used the below formula



```
death_rate =(data["Total  Deaths"].sum() /
data["Total Cases"].sum()) * 100
```

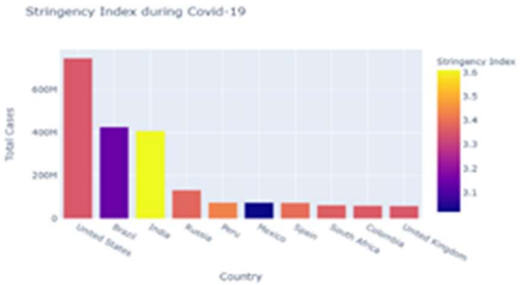


Fig 7: Stringency index

In Fig 6, another important column in this dataset is the stringency index. It is a compound indication of reaction that takes into account things like travel restrictions, job closures, and school closings. It shows how strictly countries are following these evaluates to control the outbreak of covid-19.

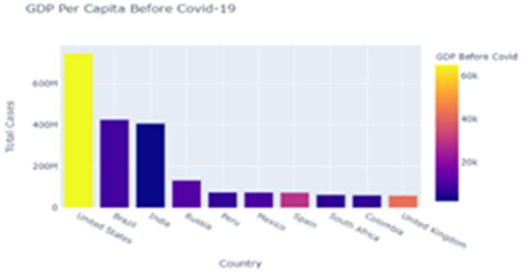


Fig 8: GDP before covid

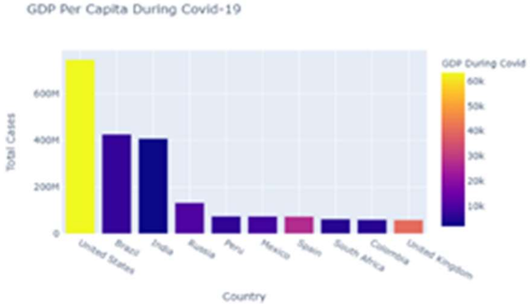


Fig 9: GDP during covid

In Fig 7&8, we can see the huge difference between before covid and during covid times in GDP rate and the pandemic affected the economy completely.

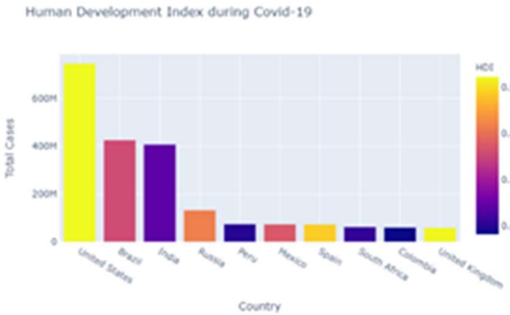


Fig 10: Human development index

Fig 9, One other important economic factor is Human Development Index. It is a statistic composite index of life expectancy, education, and per capita indicators. In that, US holding the first position and followed by brazil and India.

VII. RESULTS

In the result, we can concluded the accuracy level of prediction with the help of algorithms which we have used for model creation. We explained the level of accuracy in bar graph and in that we can see that in three algorithms random forest got highest level of accuracy compared to other algorithms.

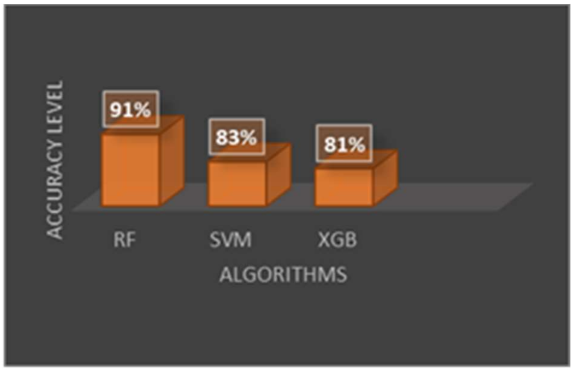


Fig 11: Comparison between accuracies of   algorithms

	S.NO.	ALGORITHMS	ACCURACY LEVEL
	1.	RANDOM FOREST	91%
	2.	SUPPORT VECTOR MACHINE	83%
	3.	XG BOOST	81%

Fig 12: Results

VIII. CONCLUSION

Three machine learning techniques, Random Forest, XG boost, and support vector machine (SVM), were used to predict the stringency level in this study. These three algorithms were tested on the same dataset in order to determine which one was the most accurate. Random Forest has a 91 percent accuracy score, whereas XG boost has an accuracy of 81 percent accuracy score. Support Vector Machine, on the other hand, has a 83 percent success rate. As a result, it may be inferred that the Random Forest algorithm is superior for predicting covid disease. In the future, the study might be expanded or modified to include other machine algorithms in the automation of Covid 19 impaction analysis. More than we have used seveal powerful libraries in python to visualize the covid scenario in pandemic and for that we used the same dataset which we used for prediction and got the output successfully.

We obtain the final score graph for accuracy after all of the comparisons. Each experiment's performance and accuracy are assessed using performance metrics such as the true positive rate.

IX. REFERENCES

[1] Singh KP, Malik YS, Tiwari R Dharma K,Sharun K, Dadar M, et al. Advances and future possibilities in creating vaccines, immunotherapies, and treatments for COVID-19, a new coronavirus illness. Human immunotherapies and vaccinations.

[2] Tateyama Y, Urasaki M , Yamamoto K, Nagayasu Y, Takahashi T, Shimamoto T, et al. Proof of concept and usage study for a health observation application for COVID-19 symptom surveillance combined with personal health information. mHealth and uHealth JMIR.

[3] Testimony from Internet Search Data: Information Seeking Reactions To Reports Of Regional COVID-19 Cases Felipe Lozano, Yong-Yeol, Coady Wing, Ana I. Bento, Thuy Nguyen, and Kosali Simon

[4] "Using generalised logistic regression to anticipate COVID-19 infection among the population," Andy Villalobos, Mario Alberto.

[5] “Reinforcement learning to optimise lockdown protocols for epidemic control” Tanuja Ganu, Harshad Khadilkar, and Dev P.

[6] Analysis of COVID-19 Impact using Data Visualization Ritik Dixit1 , Rishika Kushwah2 , Samay Pashine3.

[7] WHO, “Coronavirus disease 2019 Situation Report – 84”. Wikipedia, “Western African Ebola virus epidemic”.

[8] Worldometer, “COVID-19 Coronavirus pandemic”. https:/www.worldometers.info/coronavirus/

[9] Ramifications of the COVID-19 upsurge on Chinese-listed tourist revenue shares (Wu etc. 2021).

[10] COVID-19, 20 april 2020. https://github.com/Flame-Atlas/COVID-19 graphs .