International Journal of Engineering Research and Sustainable Technologies
Volume II, No.1, March 2024, P 1- 8
ISSN: 2584-1394 (Online version)

VISUALIZING THE RIPPLE EFFECT: COVID-19'S IMPACT AND PREDICTIVE INSIGHTS

Sreehari Thirumalai Bhuvaraghavan[*1], Usman Matheen Hameed[2]
[*1,2]Student, MS in Computer Science and Engineering, Illinois Institute of Technology, USA
* **Corresponding author email address**: sthirumalaibhuvaragh@hawk.iit.edu

https://doi.org/000000/000000/

Abstract
Throughout history, humanity has encountered numerous crises and pandemics.Yet, our resilience, innovation, and capacity for creative problem-solving have consistently guided us through these challenges. Presently, the world is confronted with the COVID-19 pandemic, stemming from a novel coronavirus originating in Wuhan, China, and swiftly spreading across the globe, infecting millions. However, in times of adversity, rigorous measures and inventive approaches become imperative. This study is dedicated to analyzing COVID-19 data using advanced visualization techniques to chart and juxtapose the outbreak's progression across various continents and territories, including the USA, China, India, Italy, and Taiwan. Employing metrics such as total confirmed cases, fatalities, and recoveries for each region, this analysis harnesses the powerof data visualization to facilitate comparisons. Ultimately, this research endeavors to furnish a comprehensive insight into the global impact of COVID-19 through the integration of Data Visualization and Machine Learning methodologies.

Keywords: Machine learning, Random forest, Datavisualization, Python and MatplotLib.

1. Introduction

The emergence of the COVID-19 pandemic has sparked a worldwide health emergency, inducing unparalleled levels of anxiety and concern. Originating in Wuhan, China, in December 2019, the novel coronavirus was promptly reported to the World Health Organization (WHO). With a variable incubation period among individuals, the virus swiftly disseminated globally, prompting its classification as a pandemic by the WHO in March 2020. By mid-April, active cases worldwide had exceeded 1 million, with the WHO estimating a mortality rate of 3.4% as of early March. Symptoms complications such as respiratory distress and anosmia. Elderly individuals and those with underlying health issues face heightened risks. The pandemic has strained healthcare systems and economies worldwide, leading to widespread lockdowns and travel restrictions. To comprehensively grasp the pandemic's impact, we will analyze diverse data points including total cases, fatalities, recoveries, lockdown timelines, and testing rates across various countries and regions. Additionally, we will draw comparisons with past epidemics like SARS, MERS, and Ebola, assessing their severity, mortality rates, and transmission dynamics. Through this analysis, we aim to glean insights to mitigate the impact of COVID-19 and bolster preparedness against future pandemics.

2. Related Work

Several studies have investigated the impact of the COVID-19 epidemic on financial markets. Salisu and Vo (2020) explored the potential of utilizing healthcare-related Google search data from affected nations and territories with high fatality rates to predict stock returns. Sharif et al. (2020) employed wavelet-based regression analysis and fidelity wavelet tests to analyze the relationships between oil prices, the COVID-19 epidemic, economic uncertainty, political instability, and the U.S. stock market. Liu et al. (2020) and Khan et al. (2020) investigated the performance of stock market indexes in heavily affected countries due to the coronavirus outbreak. Research has also focused on stock market volatility in emerging economies across regions like the Arab World, Latin America, and Central and Eastern Europe (Anser et al., 2021; Salisu and Obiora, 2021). Some studies have aimed at preventing stock market crashes during the pandemic, examining broader economic indicators in Montenegro using a Bayesian VARX approach and the impact of the COVID-19 outbreak on Chinese- listed tourism revenue shares (Wu et al., 2021; Dai et al., 2021). While previous research has utilized data visualization techniques with Python libraries in the Jupyter platform to understand the global prevalence of COVID-19, limitations exist as these methods were primarily used for visualization purposes only, employing attributes such as GDP, stringency measures, and total deaths across 16 datasets.

3. Materials and Methodologies

Our team undertook a research initiative focusing on the COVID-19 pandemic, utilizing a dataset sourced from Kaggle. This comprehensive dataset facilitated our exploration into the varied impacts of the pandemic across different regions worldwide. Our aim was to elucidate the global ramifications of the crisis. Additionally, we endeavored to forecast the implementation of stringent healthcare regulations in various countries based on their individual circumstances. To achieve this, we employed a trio of prediction algorithms: the XG Boost method, support vector machine, and random forest algorithm. Through our efforts, we aimed to provide insights into the multifaceted dimensions of the pandemic's effects and offer predictive analyses to aid in strategic decision-making.

The analysis of data pertaining to the COVID-19 pandemic from the moment of its onset until January 1st, 2020 is the main objective of this study. The data was gathered from official government websites and balanced periodic panels. In order to obtain insight into the varied effects of the virus on various places throughout the world, the research intends to investigate a number of variables, including the length of lockdown periods, fatality rates, and contagiousness. The data's source and representation are the two main factors that have been taken into account in this research. This study aims to offer a thorough knowledge of the COVID-19 pandemic and its effects on various regions of the world by using credible data sources and suitable data visualisation tools.

The provided figure outlines the research procedure. First, we start by collecting information from several websites. The gathered data next goes through preprocessing, where a variety of methods are used to convert unprocessed data into processed data. Plot diagrams are used to visually evaluate the data after preprocessing in order to improve comprehension. After the data is ready, we use three classifiers to assess the cleaned data: Random Forest, Support Vector Machine (SVM), and XG Boost. Training and testing sets of data are created before the models are analysed. All three approaches use the training dataset to train their models. The trained models are then put to the test and utilised to forecast the algorithms' accuracy levels.

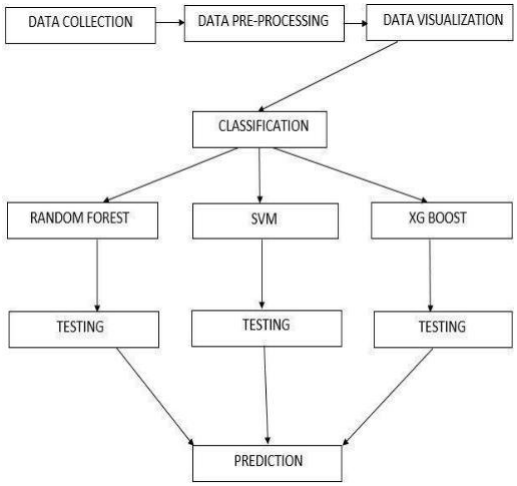

Fig 1: Architecture Diagram of Proposed System

Here, we have divided the whole process into three modules. They are Data Handling, Algorithms used, Analysis.

3.1 Data Handling – Data Collection

Data serves as a crucial component of any Artificial Intelligence (AI) system and is primarily responsible for the current surge in popularity of machine learning. The availability of data has enabled the development of scalable ML algorithms that can now function as independent solutions, adding tangible value to businesses rather than merely complementing their core operations. The dataset utilized for analyzing the impacts of COVID-19 is sourced from Kaggle, encompassing data from 170 countries spanning January 1st, 2020, to October 19th, 2020, totaling 294 days. To ensure a comprehensive and conclusive evaluation, diverse data sources are leveraged, a common practice owing to the abundance of available datasets.

3.2 Data Handling - 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. During these phases, a Replace Missing Values filter was utilized to address missing data, followed by the application of an Interquartile Range (IQR) filter to identify outlier and extreme values. To eliminate noisy data points, we employed an entropy-based binning technique. This method categorizes continuous or numerical variables by assessing the majority class label within each bin or group. Entropy calculations are performed on the target class labels, and splits are determined based on maximizing information gain. While data cleaning often involves these processes, it encompasses more than just one approach. The primary focus lies in identifying and, whenever feasible, rectifying rogue data, which may be incomplete, inaccurate, irrelevant, corrupted, or improperly formatted. Deduplication, commonly referred to as "deduping," is another essential aspect of the process, involving the consolidation or removal of similar data entries. Data integration, which merges data from disparate sources to form a unified perspective, is integral. This process includes cleansing, ETL mapping, and transformation, commencing with data ingestion. Through data integration, analytics technologies can generate actionable business insights. Transformation is critical in various stages such as data integration, migration, warehousing, and wrangling. Constructive transformation involves adding, duplicating, or replicating data, while destructive transformation entails deleting records and fields. Aesthetic transformation standardizes specific values, while structural transformation involves renaming, relocating, and merging columns.

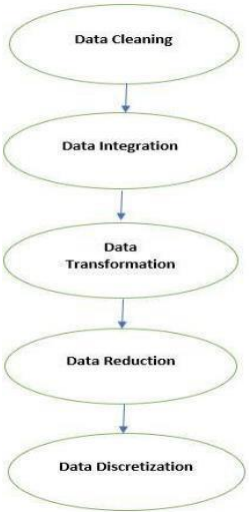

Fig.2 Pre-Processing Stages

3.3 Data Handling - Data Visualization

In this paper, we have employed statistical analysis methods to construct a descriptive framework covering data collection, analysis, interpretation, presentation, and modeling. Our framework categorizes the COVID-19 pandemic into two tiers: country-level and continent-level. Although there are similarities between the two tiers, each possesses distinct attributes, allowing us to derive separate insights from the data through inferential analysis, a component of statistical analysis.

4. Algorithms

4.1 Random Forest

The Random Forest algorithm serves dual purposes in classification and regression tasks. It builds a decision tree based on provided data and utilizes it to make predictions. It demonstrates efficacy with large datasets and maintains output consistency even in the presence of null values. Samples derived from the decision tree can be retained for further data utilization. The algorithm comprises two phases: model creation and prediction utilizing the random forest classifier established in the initial phase. Within this classifier, the Gini index assesses data purity. Known as Gini impurity, it gauges the probability of a specific variable being misclassified when chosen randomly. If each element belongs to a single class, the Gini index denotes purity. Ranging from 0 to 1, a value of 0 signifies complete uniformity within classes or a single pure class, while 1 indicates random distribution across classes or impurity. An even distribution among classes corresponds to a Gini index value of 0.5.

$$\text{Gini Index} = 1 - \frac{\sum_{i=1}^{n}(p_i)^2}{p_+)^2 + p_-)^2}$$

4.2 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.

4.3 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:

- An initial model predicting the target variable y is how P0 is defined. There will be a residual associated with this model (y – P0)
- The residuals from the previous phase are fitted to a new model, q1.
- 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.

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

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

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

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

5. ANALYSIS

5.1 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 another 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.

5.2 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.

5.3 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.

5.4 Data Visualization

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, descriptive model through the analysis of collected data and Statistical analysis application, which involves the comprehensive process of data collection, analysis, interpretation, presentation, 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 about the spread of covid all over the world and not only that GDP, human index, covid cases, death rates, economy etc.

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

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

From Figure 2 and Figure 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. 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.

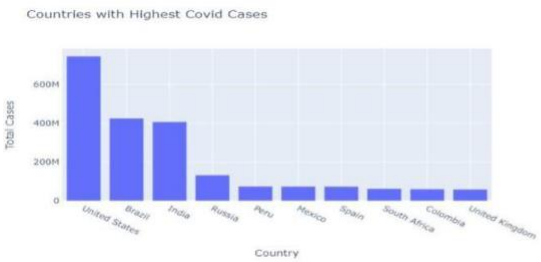

Fig. 3 Covid cases

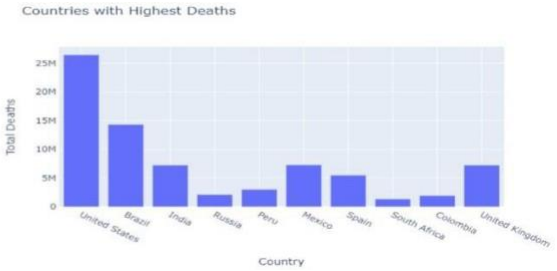

Fig. 4 Covid Deaths

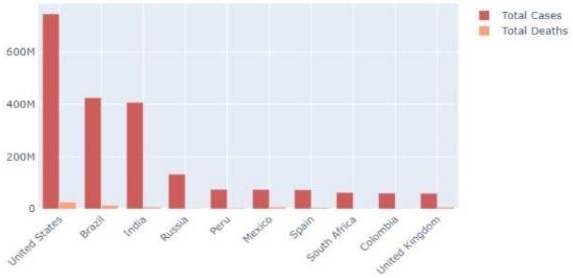

Fig. 5 Total cases and deaths

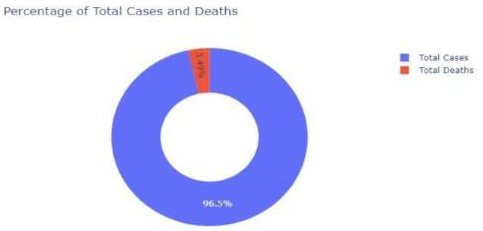

Fig. 6 Percentages of cases and deaths

Table 1. Algorithms with Accuracy Level

S.NO.	Algorithms	Accuracy Level
1.	Random Forest	91%
2.	Support Vector Machine	83%
3.	XG Boost	81%

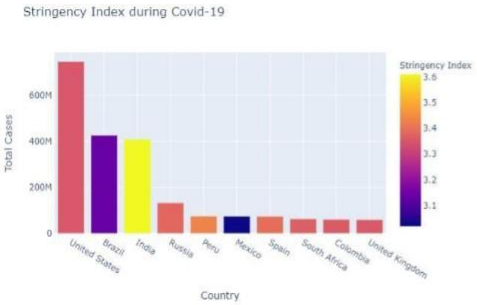

Fig. 7 Stringency index

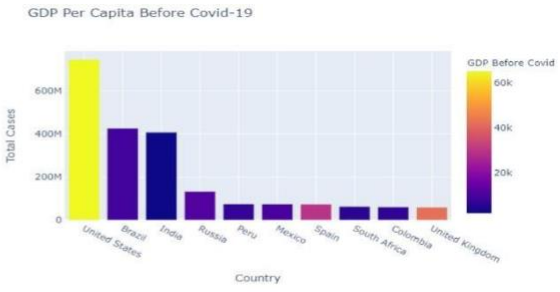

Fig. 8 GDP before Covid

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. 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.

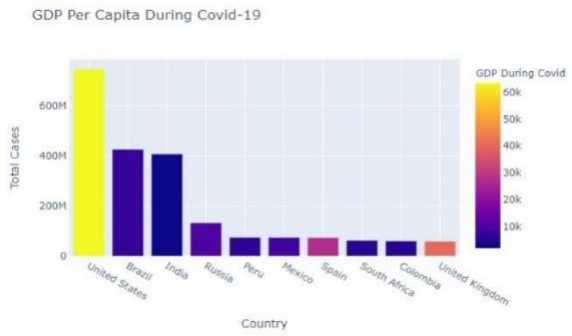

Fig. 9 GDP during Covid

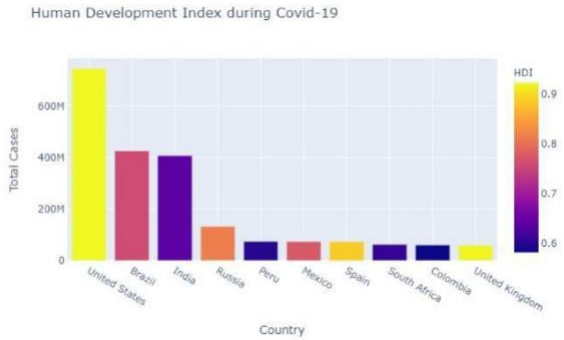

Fig.10 Human Development Index

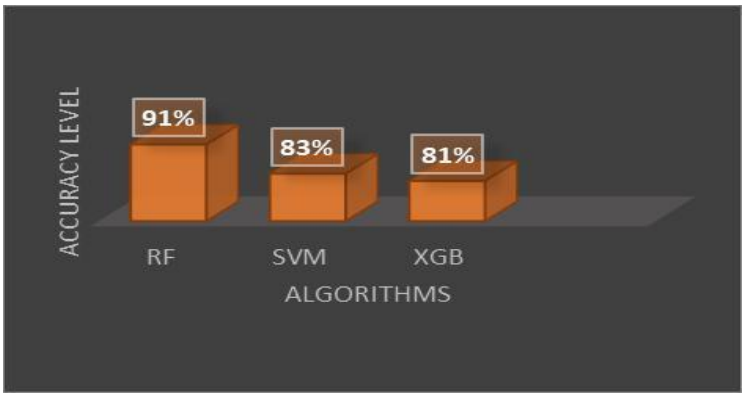

Fig.11 Comparison between accuracies of algorithms

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.

6. 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.

7. Conclusion

In this study, three machine learning techniques, namely Random Forest, XG Boost, and Support Vector Machine (SVM), were employed to predict the stringency level. These algorithms underwent testing on the same dataset to ascertain their respective accuracies. Random Forest achieved a high accuracy score of 91%, followed by XG Boost with 81%, and Support Vector Machine with 83%. Consequently, it can be deduced that the Random Forest algorithm excels in predicting COVID-19 disease. Future iterations of the study could explore additional machine learning algorithms for automating COVID-19 impact analysis. Additionally, powerful Python libraries were utilized to visualize the pandemic scenario using the same dataset employed for prediction, yielding successful outcomes. The final accuracy score graph was generated post-comparison, evaluating each experiment's performance and accuracy through metrics such as the true positive rate.

References

1. 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.
2. Testimony from Internet Search Data: Information Seeking Reactions To Reports Of Regional COVID-19 Cases Felipe Lozano, YongYeol, Coady Wing, Ana I. Bento, Thuy Nguyen, and Kosali Simon
3. "Using generalised logistic regression to anticipate COVID-19 infection among the population," Andy Villalobos, Mario Alberto.
4. “Reinforcement learning to optimise lockdown protocols for epidemic control” Tanuja Ganu, Harshad Khadilkar, and Dev P.
5. Analysis of COVID-19 Impact using Data Visualization Ritik Dixit1 , Rishika Kushwah2 , Samay Pashine3.
6. WHO, “Coronavirus disease 2019 Situation Report – 84”. Wikipedia, “Western African Ebola virus epidemic”.
7. Worldometer, “COVID-19 Coronavirus pandemic”. https:/www.worldometers.info/ coronavirus/
8. Ramifications of the COVID-19 upsurge on Chinese-listed tourist revenue shares (Wu etc. 2021).
9. COVID-19, 20 April 2020. https://github.com/Flame-Atlas/COVID-19 graphs