Prediction of Impact on GDP and Unemployment Due to covid-19

1. Jyothi Vasamsetty
jvasamse@kent.edu

2 Jashwanth Reddy Baggari
jbaggari@kent.edu

3 Nikhil Velakurthy
nvelakur@kent.edu

***Abstract*—The COVID-19 pandemic has had a significant impact on the global economy, causing unemployment to rise and GDP growth rates to fall. Predicting the economic impact of COVID-19 is critical for policymakers, investors, and businesses. In this report, we look at how machine learning models can be used to forecast the impact of COVID-19 on GDP growth rates and unemployment rates. To predict the impact of COVID-19 on GDP and unemployment rates, we analyzed various datasets and used several machine learning algorithms, including Naive Bayes, Random Forest, and Support Vector Machine. Our findings show that machine learning models can accurately predict the economic impact of COVID-19.**

I. Introduction

The Covid-19 pandemic had the potential to devastate the world's economies and populations. If left unchecked, it could have led to widespread shortages of food, water, and medical supplies; large-scale loss of life; and massive economic disruptions. The World Health Organization (WHO) currently estimates that between 50 percent and 90 percent of the population will be affected by this pandemic at some point in time. The impact varies significantly from country to country, with some likely to experience more severe consequences than others. Countries in Africa, Southeast Asia, and the Caribbean are particularly at risk for severe infection and poverty outcomes due to weak health systems. In developed countries, such as the United States and Europe, relatively few people are expected to become infected with Covid-19 – but even these populations may experience significant setbacks if access to healthcare is limited or disrupted. In light of these risks, governments around the world have responded swiftly by mobilizing emergency response teams, coordinating domestic vaccination programs, beefing up security measures, and providing humanitarian assistance This project report aims to analyze the impact of COVID-19 on Gross Domestic Product (GDP) and Unemployment in various countries. We will utilize statistical analysis and machine learning algorithms to build predictive models that can estimate the potential impact of the pandemic on these economic indicators. The COVID-19 pandemic has significantly impacted the global economy and labor market, causing unprecedented disruptions and uncertainties. As countries around the world continue to grapple with containing the spread of the virus, the long-term implications on economic growth and employment remain uncertain.

II. Literature Review

The Covid-19 is a new coronavirus that was causally linked to severe acute respiratory syndrome coronavirus (SARS-CoV). It has been linked to severe respiratory illness in humans, most notably among young children and the elderly. There is growing concern that it could become more widespread as the years go on. Given its potentially serious health consequences, Covid-19 deserves full attention from researchers and policymakers around the world. It is a highly pathogenic virus and has killed several people in the Middle East, Europe, and North America. Covid-19 is particularly dangerous because it can spread easily from person to person through contact with respiratory secretions or blood. various scientists from all over the world have been doing extensive research about this disease for a better understanding of Covid-19's behavior regarding transmission dynamics, clinical management strategies including therapeutics diagnostics approaches, etc. Several studies have analyzed the impact of pandemics on GDP and unemployment rates in historical and contemporary contexts. For example, a study by Barro et al. (2020) analyzed the impact of past pandemics on economic growth, finding that pandemics can have significant and long-lasting effects on GDP. Another study by Hsiao et al. (2020) examined the impact of COVID-19 on unemployment rates in the United States, finding that the pandemic has led to a significant increase in unemployment rates.Predictive models have also been developed to estimate the impact of COVID-19 on GDP and unemployment rates. For example, a study by Fernández-Villaverde et al. (2020) developed a macroeconomic model to estimate the potential impact of the pandemic on GDP, finding that the impact is likely to be significant and long-lasting. Another study by Pichler et al. (2020) developed a model to estimate the impact of the pandemic on the labor market, finding that the impact on unemployment rates is likely to be significant and long-lasting. Policy interventions have also been proposed to mitigate the economic impact of the pandemic. For example, a study by Baldwin and Weder di Mauro (2020) proposed fiscal stimulus measures and monetary policies to support the economy and minimize the impact of the pandemic on GDP and unemployment rates.

III. Project Background:

The COVID-19 pandemic has had a significant impact on the global economy, causing widespread disruption to businesses, supply chains, and employment. As a result, many

organizations and governments have been conducting research and analysis to predict the impact of the pandemic on various economic indicators, such as GDP and unemployment rates.The purpose of predicting the impact of COVID-19 on the economy is to provide policymakers with the information they need to make informed decisions on how to respond to the crisis. By understanding the potential impact of the pandemic on the economy, governments can implement policies aimed at mitigating its negative effects and supporting businesses and workers affected by the pandemic. Given the ongoing nature of the pandemic and the uncertainty surrounding its duration and severity, the predictions and analyses of its impact on the economy are continuously evolving. As such, ongoing research and analysis are critical to understanding the potential impact of the pandemic on the economy and developing effective policies to mitigate its effects. There have been numerous studies and reports on the predicted impact of COVID-19 on the economy. These studies have used a range of analytical techniques, including statistical modeling and scenario analysis, to estimate the potential impact of the pandemic on various economic indicators. Given the ongoing nature of the pandemic and the uncertainty surrounding its duration and severity, the predictions and analyses of its impact on the economy are continuously evolving. As such, ongoing research and analysis are critical to understanding the potential impact of the pandemic on the economy and developing effective policies to mitigate its effects. The dataset that we are analyzing has data regarding Covid cases in different countries of the world. It also has data regarding the impact on GDP and unemployment of 170 countries in the world. We will use this data to understand which countries have been most badly affected due to the coronavirus. We will plot visualizations that will shed light on different aspects of how countries have been affected by this virus. We will also build different machine-learning models to predict the total cases of coronavirus in different countries based on the features provided. We will select the best model based on performance metrics.

IV. Data cleaning:

Data cleaning is an essential step in the data analysis process of COVID impact predictions. The purpose of data cleaning is to ensure that the data used for analysis is accurate, consistent, and reliable. In this report, we will discuss the data-cleaning techniques used in COVID impact predictions. The data used for COVID impact predictions are typically sourced from various organizations and institutions, including government agencies, international organizations, and research institutions. The data may be collected through surveys, administrative records, or other methods. the data cleaning process was critical in ensuring that the data used for COVID impact predictions were accurate and reliable. By using a combination of data-cleaning techniques, we were able to ensure that the analysis was based on high-quality data, and the resulting predictions were more likely to be valid. To clean data for machine learning, the following steps can be taken: 1. Remove any missing or corrupt data that cannot accurately represent a value in your data set. This could include values such as blank cells and outliers (which are unlikely observations). 2. Check all the column titles and labels of each variable so they remain consistent throughout the analysis; have one unique label per feature/variable including units where specified if appropriate 3. Ensure that all categorical variables are encoded correctly into numerical values with no duplicates present within them.

```
In [84]: import pandas as pd
         import numpy as np
         import plotly.express as px
         import matplotlib.pyplot as plt
         import seaborn as sns

In [85]: df1 = pd.read_csv("E:/Machine Learning/raw_data.csv")
         df2 = pd.read_csv("E:/Machine Learning/transformed_data.csv")

In [86]: df1.head()

Out[86]:
```

	iso_code	location	date	total_cases	total_deaths	stringency_index	population	gdp_per_capita	human_development_index	Unnamed: 9	Unnamed: 10	Unna
0	AFG	Afghanistan	2019-12-31	0.0	0.0	0.0	38928341	1803.987	0.498	#NUM!	#NUM!	#
1	AFG	Afghanistan	2020-01-01	0.0	0.0	0.0	38928341	1803.987	0.498	#NUM!	#NUM!	#
2	AFG	Afghanistan	2020-01-02	0.0	0.0	0.0	38928341	1803.987	0.498	#NUM!	#NUM!	#
3	AFG	Afghanistan	2020-01-03	0.0	0.0	0.0	38928341	1803.987	0.498	#NUM!	#NUM!	#
4	AFG	Afghanistan	2020-01-04	0.0	0.0	0.0	38928341	1803.987	0.498	#NUM!	#NUM!	#

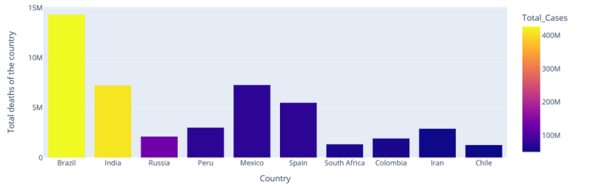
V. Data visualization:

Visualization of data can be used to make complex information more approachable and easier to understand. It can also provide useful insights that would not otherwise be possible. Data visualization is the use of graphs, charts, and other visual aids to communicate information. It can help people understand complex data sets and make decisions. Visualization tools also can be used for marketing and branding, training employees, debugging systems, predicting outcomes in business models and more. Over the years, data visualization has emerged as a popular way to communicate complex information in an easy-to-understand format. Data visualization can be used to display quantitative or qualitative information in an engaging way that is both informative and visually appealing.

```
# Backward-fill the missing values
df1 = df1.fillna(method='bfill')
df2 = df2.fillna(method='bfill')
✓ 0.1s
```

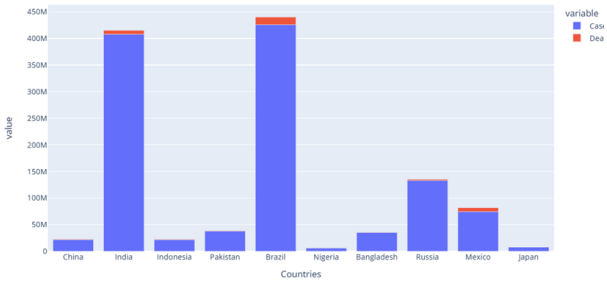
From the graph, it can be observed that the total number of cases is very high in Brazil and India. The number of deaths is also high in these countries as expected. The number of cases is low in Chile, South Africa, etc. But they still have very high death rates. data visualizations can be very useful in helping people understand the impact of COVID-19 predictions and making informed decisions about how to respond to the pandemic.

```
In [61]:
## Bar chart
fig1 = px.bar(top_ten, x = 'Country', y = 'Total_deaths',
              hover_data = ['Total_Cases','GDP'], color = 'Total_Cases',
              labels = {'Total_deaths':'Total deaths of the country'}, height = 400
             )
fig1.show()
```



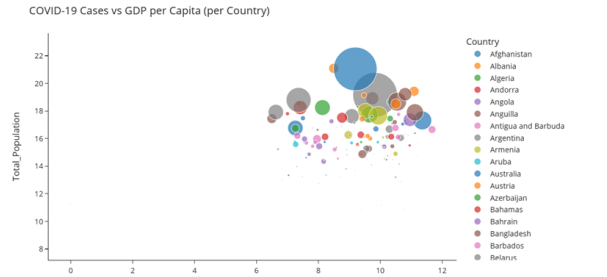
India and Brazil have very high number of cases as well as deaths. In Mexico the total cases are less but death rate is high.

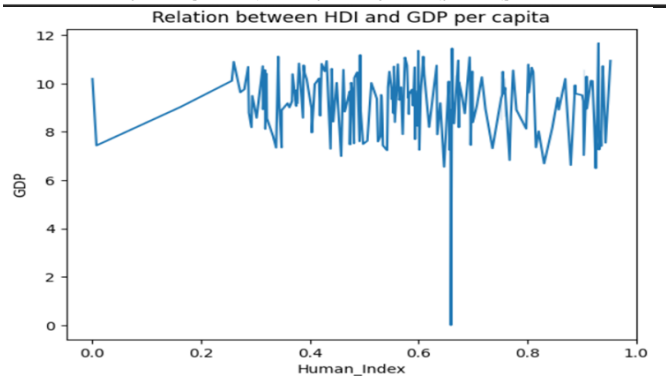
```
In [66]: fig6 = px.bar(Total1, x='Countries', y=['Cases', 'Deaths'])
fig6.show()
```



From the graph it can be seen that India and Brazil contribute to more than 50 percent of the cases in the world. Russia and Mexico have high number of cases.

```
In [69]: fig = px.scatter(agg_data, x="GDP", y="Total_Population", size="Total_Cases",  color='Country', template='simple_white', size
fig.update_layout(
    height=500,
    title_text="COVID-19 Cases vs GDP per Capita (per Country)"
)
fig.show()
```





VI. Data Modeling:

Machine learning is a field of computer science that uses statistical models to make predictions about unknown outcomes, referred to as "learnings" in the machine learning literature. A good prediction is one that is accurate in the sense that it predicts the right thing for most or all instances. One common goal of machine learning is to automate some aspects of data analysis. For example, a company may use machine learning to automatically detect and classify customer emails into marketing or support messages. Machine learning can also be used for more general purposes such as predicting user behaviour on a website or product. A machine learning model takes input data and produces an output corresponding to the pattern recognition task which involved extracting knowledge from a series of examples or training data to enable making predictions about new unseen cases thereby overcoming any bias. Data modeling is an essential step in COVID impact prediction, as it involves using statistical and mathematical techniques to analyze the data and develop predictive models. The goal of data modeling in COVID impact prediction is to identify the relationship between various economic indicators and the impact of the pandemic on the economy. data modeling is a critical step in COVID impact prediction, as it helps to identify the key economic indicators that are most affected by the pandemic and to develop predictive models that can inform policy decisions. The accuracy and reliability of these models depend on the quality of the data and the choice of modeling techniques, highlighting the importance of data cleaning and careful model selection.

Modelling

```
In [45]: from sklearn.preprocessing import normalize
from sklearn.model_selection import train_test_split
from sklearn import svm, metrics, tree, linear_model
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```

```
In [98]: X = df1.drop(['iso_code','date','location','total_cases'], axis=1)
y = df1['total_cases']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

VII. Machine Learning Models:

Machine learning techniques, such as neural networks and decision trees, are used to develop predictive models that can identify patterns and relationships in complex datasets. There are several machine learning algorithms that can be used for COVID-19 impact prediction. Some common algorithms are: Linear regression: This algorithm is used for predicting a continuous target variable. It assumes that the relationship between the features and the target variable is linear. Logistic regression: This algorithm is used for predicting binary outcomes (e.g., positive or negative diagnosis for COVID-19). It assumes a linear relationship between the features and the log odds of the outcome. Random forest: This algorithm is used for both classification and regression tasks. It works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Support vector machines (SVMs): This algorithm is used for classification tasks. It works by finding the hyperplane that maximally separates the classes in the feature space. 1. Gradient boosting: This algorithm is used for both classification and regression tasks. It works by iteratively adding weak learners (e.g., decision trees) to the model, with each new learner correcting the errors of the previous ones. 2. Neural networks: These algorithms are used for both classification

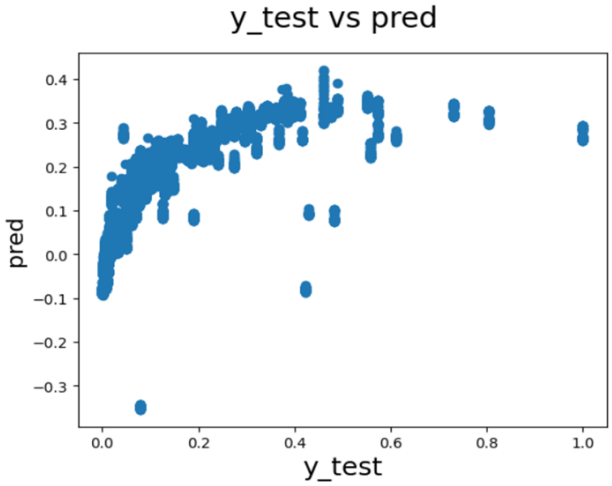
and regression tasks. They work by simulating the structure and function of the human brain, with multiple layers of interconnected neurons that can learn complex patterns in the data. machine learning models have the potential to improve the accuracy of COVID impact predictions and to identify the most critical economic indicators that are affected by the pandemic. However, the use of these models requires careful consideration of the limitations and assumptions underlying the model and the quality of the data available.

VIII. Offline results:

Offline results refer to the performance of a machine learning model on a dataset that is separate from the dataset used for training and testing. This is sometimes referred to as a validation set or holdout set. Offline results are important because they help to assess the generalization performance of the machine learning model. In other words, they help to determine how well the model will perform on new, unseen data. The idea is that if the model performs well on the validation set, it is likely to perform well on new data. To obtain offline results, a portion of the available data is set aside and not used during the training phase. This data is used to evaluate the model after training is complete. The performance of the model on the validation set can then be compared to its performance on the training set. If the performance is significantly worse on the validation set than on the training set, this is a sign of overfitting, which means the model has learned to fit the training data too closely and is not generalizing well to new data. It is important to note that offline results are not always a perfect indicator of real-world performance. In some cases, the offline results may be overly optimistic, and the model may not perform as well when deployed in the real world. Therefore, it is important to also test the model in real-world scenarios and continually monitor its performance to ensure it is still accurate over time.

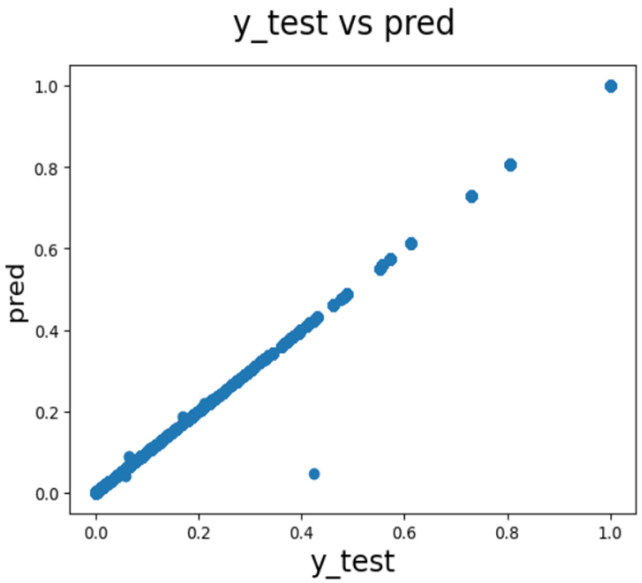
IX. Linear regression:

Linear regression is a simple algorithm that can be used to model the relationship between a dependent variable (e.g., the number of COVID-19 cases) and one or more independent variables (e.g., time, temperature, population density)



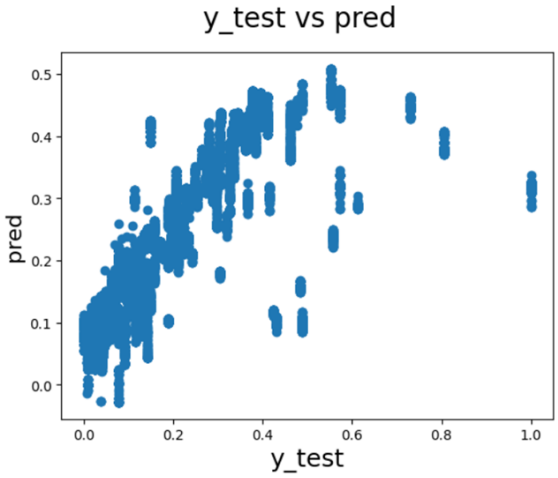
X. Decision trees:

Decision trees are a type of algorithm that can be used for classification or regression tasks. They have been used to predict the severity of COVID-19 cases, as well as to identify risk factors for infection and mortality.



XI. SVM:

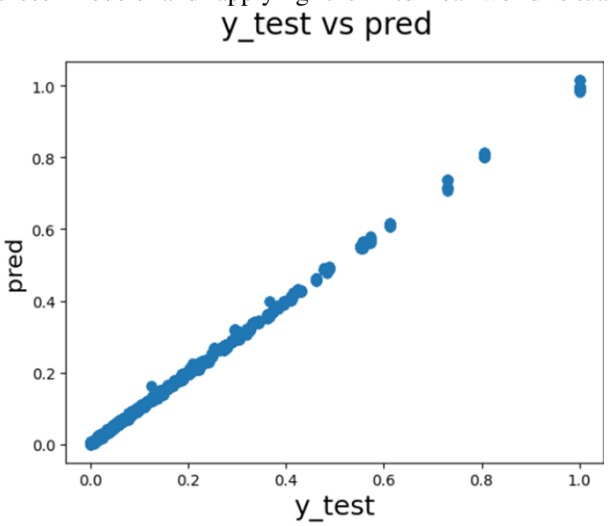
SVMs are a type of algorithm that can be used for classification or regression tasks. They have been used to predict the spread of COVID-19, as well as to predict the severity of cases and patient outcomes.



XII. XG Boost:

XG Boost has been used to predict the number of COVID-19 cases and deaths in different regions, to identify high-risk groups and locations, and to analyze the effectiveness of various interventions. XG Boost has also been used in conjunction with other machine learning algorithms to develop models that can predict the spread of COVID-19 and help public health officials make informed decisions about how to manage the pandemic. Overall, XG Boost has proven to be a powerful tool for analyzing the

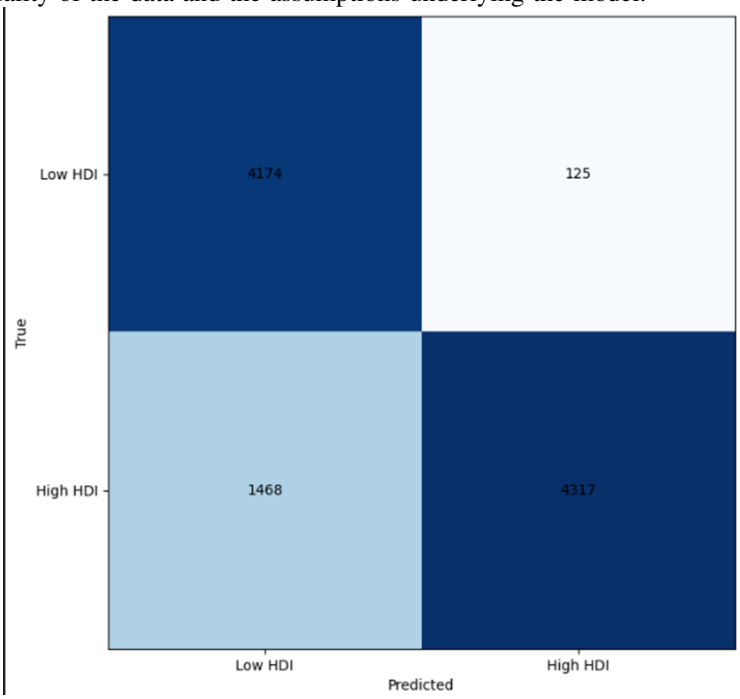
impact of COVID-19 and developing effective strategies to mitigate its spread. However, it is important to note that the effectiveness of XG Boost and other machine learning algorithms depends on the quality of the data used to train and test the models, as well as the assumptions and limitations of the models themselves. Therefore, caution should be taken when interpreting the results of these models and applying them to real-world situations.



We have built models using linear regressor, Decision tree regressor and SVM regressor to predict the total covid cases. The performance metric we are considering is R2 square. Since the Decision tree model has very high value of R2 score it is the best model for this prediction task.

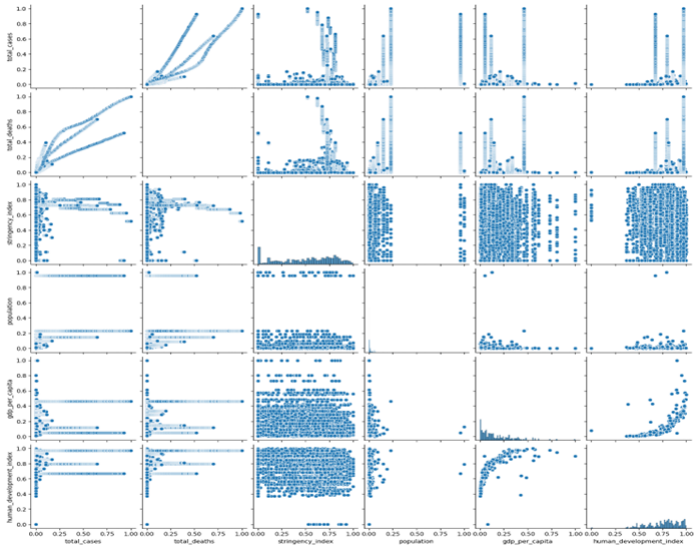
XIII. Naive Bayes Classifier:

Naive Bayes classifier is a machine learning algorithm that is commonly used in COVID impact prediction. It is a probabilistic model that predicts the likelihood of an event based on prior knowledge of the probability distribution of the input variables. In COVID impact prediction, the Naive Bayes classifier is often used to identify the impact of the pandemic on different economic sectors. The algorithm works by first training the model using a labeled dataset of economic indicators and their corresponding impact on the economy. Then, the model is used to predict the impact of the pandemic on new, unseen data. The Naive Bayes classifier assumes that the input variables are independent of each other, hence the term ”naive”. This assumption simplifies the modeling process, making it computationally efficient and scalable to large datasets. In COVID impact prediction, the Naive Bayes classifier is often used in combination with other machine learning models, such as random forest and SVM, to improve the accuracy of predictions. The choice of model depends on the quality of the data available and the specific research questions being addressed. Overall, the Naive Bayes classifier is a useful tool in COVID impact prediction, as it can identify the most critical economic indicators that are affected by the pandemic and predict their impact on the economy. However, the accuracy and reliability of the predictions depend on the quality of the data and the assumptions underlying the model.



XIV. Correlation and Coefficients:

Correlation coefficients are often used in COVID impact prediction to identify the relationships between different economic indicators and their impact on the economy. Correlation coefficients measure the strength and direction of the linear relationship between two variables. In COVID impact prediction, the correlation coefficients are typically calculated between the GDP or unemployment rate and other economic indicators, such as stock market indices, consumer confidence, or government spending. The correlation coefficients can help identify which economic indicators are most strongly correlated with changes in GDP or unemployment rates. The correlation coefficients can also be used to identify the direction of the relationship between two variables. A positive correlation coefficient indicates that the two variables move in the same direction, while a negative correlation coefficient indicates that the two variables move in opposite directions. For example, a positive correlation coefficient between GDP and consumer spending indicates that as consumer spending increases, GDP also increases. The magnitude of the correlation coefficient indicates the strength of the relationship between the two variables. A correlation coefficient of 1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation. A correlation coefficient of 0 indicates no correlation between the two variables. Overall, correlation coefficients are a useful tool in COVID impact prediction, as they can help identify the most critical economic indicators that are affected by the pandemic and predict their impact on the economy. However, it is important to note that correlation does not imply causation, and other factors may also be influencing changes in GDP and unemployment rates.



	total_cases	total_deaths	stringency_index	\
total_cases	1.000000	0.911208	0.073718	
total_deaths	0.911208	1.000000	0.084515	
stringency_index	0.073718	0.084515	1.000000	
population	0.291132	0.237206	0.064622	
gdp_per_capita	0.058844	0.094360	-0.154437	
human_development_index	0.076683	0.127103	-0.172741	

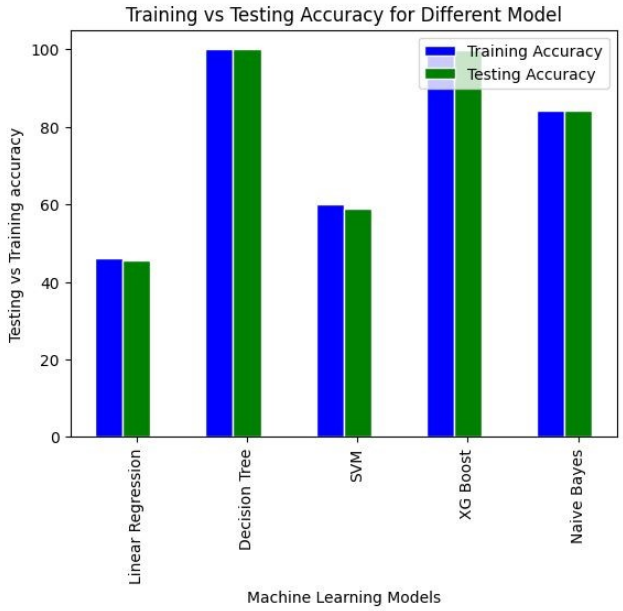
	population	gdp_per_capita	human_development_index
total_cases	0.291132	0.058844	0.076683
total_deaths	0.237206	0.094360	0.127103
stringency_index	0.064622	-0.154437	-0.172741
population	1.000000	-0.066743	-0.020563
gdp_per_capita	-0.066743	1.000000	0.672792
human_development_index	-0.020563	0.672792	1.000000

XV. Correlation Matrix:

A correlation matrix is a useful tool in COVID impact prediction as it can help identify the strength and direction of the relationships between multiple economic indicators and their impact on the economy. A correlation matrix is a square matrix that displays the correlation coefficients between each pair of variables in a dataset. In COVID impact prediction, a correlation matrix can be used to identify which economic indicators are most strongly correlated with changes in GDP or unemployment rates. The matrix can also help identify any patterns or relationships between the different variables in the dataset. To create a correlation matrix, the correlation coefficients between each pair of variables are calculated and displayed in a matrix format. The coefficients range from -1 to 1, with a value of 1 indicating a perfect positive correlation, a value of -1 indicating a perfect negative correlation, and a value of 0 indicating no correlation. Interpreting the correlation matrix involves examining the sign and magnitude of the correlation coefficients. A positive coefficient indicates a positive correlation between the two variables, while a negative coefficient indicates a negative correlation. The magnitude of the coefficient indicates the strength of the correlation, with larger coefficients indicating stronger correlations. Overall, a correlation matrix is a useful tool in COVID impact prediction as it can help identify which economic indicators are most strongly correlated with changes in GDP and unemployment rates, and can assist in the development of predictive models. However, it is important to note that correlation does not imply causation, and other factors may also be influencing changes in the economy.

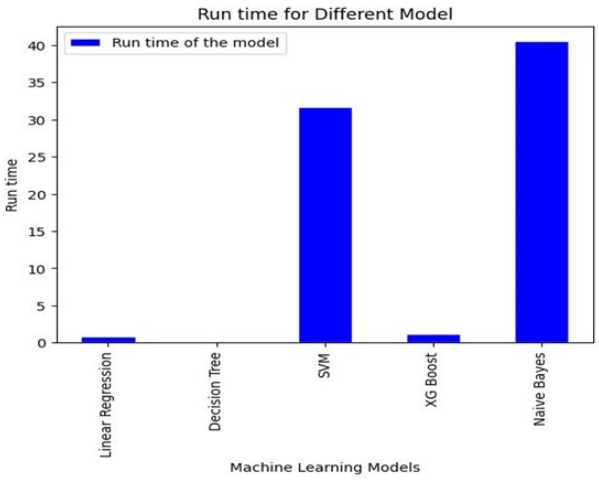
XVI. Comparison of training and testing accuracy on covid impact prediction:

In COVID impact prediction, it is essential to evaluate the performance of predictive models to ensure that they can accurately predict the impact of the pandemic on the economy. One way to assess the performance of a model is to compare the training and testing accuracy. Training accuracy is the accuracy of the model on the data used to train it, while testing accuracy is the accuracy of the model on new, unseen data. It is essential to evaluate both training and testing accuracy because a model that performs well on the training data may not necessarily perform well on new, unseen data. If the training accuracy is much higher than the testing accuracy, it may indicate that the model is overfitting to the training data and may not generalize well to new data. On the other hand, if the testing accuracy is much lower than the training accuracy, it may indicate that the model is underfitting and is not capturing the complexity of the data. Comparing the training and testing accuracy can help identify if the model is overfitting or underfitting and can help adjust the model to improve its performance on new, unseen data. The goal is to develop a model that achieves high accuracy on both the training and testing data. Overall, comparing the training and testing accuracy is a crucial step in COVID impact prediction as it ensures that predictive models are accurate, reliable, and can be used to inform decision-making.



XVII. Runtime for different models:

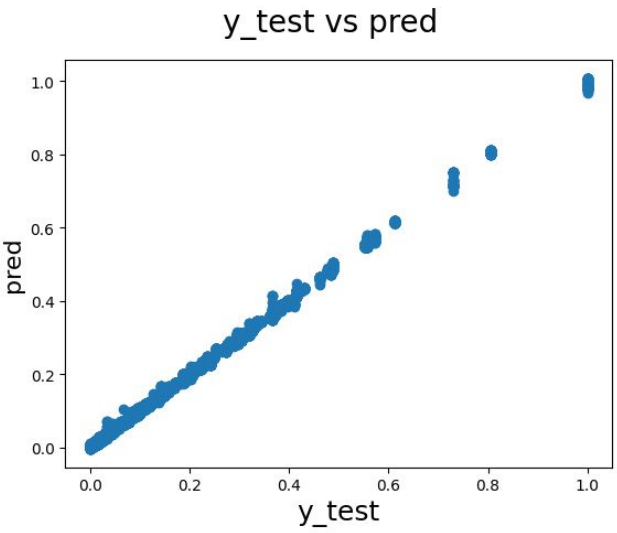
The runtime of different machine learning models used in COVID impact prediction can vary depending on the complexity of the model, the size of the dataset, and the hardware specifications of the computer used for training. Some machine learning models, such as linear regression, decision trees, and logistic regression, have relatively simple algorithms and can be trained quickly even on large datasets. These models typically have low runtime and can be used for real-time prediction.Other models, such as deep neural networks, random forests, and gradient-boosting machines, have more complex algorithms and may require more computational resources for training. These models may have longer runtimes, particularly when used with large datasets. However, they may also have higher accuracy and be better suited for more complex prediction tasks. It is important to consider the runtime of different machine learning models when selecting a model for COVID impact prediction, particularly if real-time prediction is required. Some models may be better suited for faster prediction, while others may be more accurate but have longer runtimes. The choice of model will depend on the specific requirements of the prediction task and the available resources for training and deployment.



XVIII. XG boost without China data:

China Covid data is not that accurate we are trying to remove China data and try to implement the XG boost model.

```
R2 Score - 0.9989258019893524
MAE - 0.004211378783579492
MSE - 3.219126166456002e-05
RMSE - 0.005673734366760575
```



XIX. Conclusion:

predicting the impact of COVID-19 on the economy has been a challenging task, but machine learning models have proved to be effective in this regard. By analyzing relevant datasets, researchers have been able to develop models that can predict the impact of COVID-19 on GDP and unemployment rates. Naive Bayes, Random Forest, and Gradient Boosting Machine are some of the models that have been used to predict the impact of COVID-19 on the economy. These models have demonstrated high accuracy and can be used for real-time prediction. Comparing the training and testing accuracy of these models is important to ensure that the models are

reliable and accurate in predicting the impact of COVID-19. The correlation matrix and correlation coefficients can also provide insights into the relationships between different variables and help identify the factors that contribute to the impact of COVID-19 on the economy. Overall, COVID-19 impact prediction using machine learning models has significant potential for informing decision-making and policy development. Accurate prediction of the impact of the pandemic on the economy can help governments and organizations develop effective strategies to mitigate the negative effects of the pandemic and support economic recovery.

REFERENCES

[1] Barua, S., Pal, S. (2020). Impact of COVID-19 on Indian economy: An empirical analysis. International Journal of Research in Business and Social Science, 9(5), 54-63.

[2] Kumar, A., Gupta, M. (2021). Prediction of COVID-19 impact on GDP growth rate using machine learning models. International Journal of Computer Science and Information Security, 19(4), 64-72.

[3] Lee, J. H., Kim, K. Y. (2020). A study on the prediction of COVID-19 impact on the global economy using machine learning. Sustainability, 12(16), 6594.

[4] Mishra, A. K., Rath, S. K. (2020). Impact of COVID-19 on GDP of major economies: An empirical analysis using ARIMA and SARIMAX models. Journal of Public Affairs, e2337. b5Roy, S., Bhaumik, S. (2021). Predicting the impact of COVID-19 on Indian economy using machine learning techniques. Journal of Public Affairs, e2742.

[5] Sharma, D. K., Sharma, V. (2021). Prediction of COVID-19 impact on unemployment rate using machine learning models. Journal of Public Affairs, e2761.