Real-Time COVID-19 Case Prediction: A Machine Learning Approach

Project submitted to Asian School of Media Studies in partial fulfillment of the requirements for the award of degree of

M.Sc.

in

Data Science

By

Pradeep Kumar Nishad

(University Enroll. No: 2021070165)

Under the Supervision of

Prof. Niharika Tewari

ASIAN SCHOOL OF MEDIA STUDIES

NOIDA

2023

Declaration

I, **Pradeep Kumar Nishad**, S/O **Bharat Singh Nishad**, declare that my project entitled "**Real-Time COVID-19 Case Prediction: A Machine Learning Approach**", submitted at **School of Data Science, Asian School of Media Studies, Film City, Noida,** for the award of **M.Sc. in Data Science, Shobhit University,** and **Post Graduate Diploma in Data Science, ASMS**, is an original work and no similar work has been done in India anywhere else to the best of my knowledge and belief.

This project has not been previously submitted for any other degree of this or any other University/Institute.

Signature:

Pradeep Kumar Nishad
+91-7906207881
npradeep7906@gmail.com
M.Sc. Data Science
School of Data Science
Asian School of Media Studies

Acknowledgements

The completion of the project titled **"Real-Time COVID-19 Case Prediction: A Machine Learning Approach"**, gives me an opportunity to convey my gratitude to all those who helped to complete this project successfully. I express special thanks:

- To **Prof. Sandeep Marwah**, President, Asian School of Media Studies, who has been a source of perpetual inspiration throughout this project.
- To **Mr. Ashish Garg**, Director for School of Data Science for your valuable guidance, support, consistent encouragement, advice and timely suggestions.
- To **Ms. Niharika Tewari**, Assistant Professor of School of Data Science, for your encouragement and support. I deeply value your guidance.
- To **my friends** for their insightful comments on early drafts and for being my worst critic. You are all the light that shows me the way.

  To all the people who have directly or indirectly contributed to the writing of this thesis, but their names have not been mentioned here.

Signature:
Pradeep Kumar Nishad
+91-7906207881
npradeep7906@gmail.com
M.Sc. Data Science
School of Data Science
Asian School of Media Studies

Abstract

This project employs various machine learning models, including Linear Regression, Autoregressive integrated moving average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA) and Random Forest Regressor, to analyze and forecast COVID-19 cases in India, the United States, and China. The dataset, obtained from OWID, undergoes preprocessing to ensure data quality. Exploratory Data Analysis reveals patterns and correlations in the data. The inclusion of ARIMA and Random Forest models enhances the accuracy of the forecasts by capturing complex relationships and seasonal patterns. The findings contribute to informed decision-making, resource allocation, and risk assessment for combating the ongoing COVID-19 pandemic. The ensemble of models provides valuable insights into COVID-19 case trends, aiding public health authorities, policymakers, and researchers in their efforts to mitigate the spread of the virus. The project's outcomes contribute to a better understanding of the pandemic's dynamics and support evidence-based strategies for controlling and managing COVID-19.

Contents

List of Figures

Acronyms

ARIMA	Autoregressive integrated moving average.
RMSE	Root Mean Squares Error.
SARIMA	Seasonal Autoregressive Integrated Moving Average.
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors.
SVM	Support Vector Machine.
SVR	Support Vector Regression.

Chapter 1: Real-Time COVID-19 Prediction Framework

0.1 General Introduction

0.1.1 Background

The COVID-19 epidemic has become an unparalleled global health emergency, affecting people, communities, and countries all over the world. The virus, technically known as SARS-CoV-2, is a member of the coronavirus family and has been linked to a variety of respiratory disorders. Since its initial identification in Wuhan, China, in December 2019, COVID-19 has spread rapidly over the world, leading the World Health Organisation (WHO) to classify it as a pandemic.

COVID-19 can spread through respiratory droplets produced when an infected person coughs, sneezes, or talks. It can also spread by touching contaminated surfaces and then touching your face. The virus enters the body through the nose, mouth, or eyes. Once inside, it attaches to cells in the respiratory tract and replicates. This can lead to inflammation and damage to the lungs, which can cause a range of symptoms, from mild to severe. Common symptoms of COVID-19 include fever, cough, fatigue, sore throat, and loss of taste or smell. In severe cases, the virus can lead to pneumonia, acute respiratory distress syndrome (ARDS), and death. Certain populations, such as older adults and people with underlying health conditions, are at higher risk of developing severe illness.

Beyond its effects on a person's own health, COVID-19 has widespread implications. The epidemic has significantly disrupted businesses, put a burden on healthcare systems, and led to social and psychological hardship. Controlling the virus's spread, maintaining access to healthcare services, and putting in place practical population protection measures have been extremely difficult for governments and healthcare institutions throughout the world. This pandemic has had a profound economic impact globally. Lockdown measures, travel restrictions, and business closures have resulted in a significant decline in economic activity. Many businesses, especially small and medium-sized enterprises, have faced financial difficulties and job losses. Sectors such as tourism, hospitality, and retail have been particularly hard-hit, leading to increased income inequality and financial insecurity. COVID-19 has also disrupted social interactions and daily routines, leading to feelings of loneliness, stress, and anxiety. School closures and the shift to remote learning have created challenges in education access and quality. Vulnerable populations, including the elderly and those with pre-existing mental health conditions, have been disproportionately affected by the social and psychological impacts of the pandemic.

The motivation behind this research lies in the urgent need to develop real-time COVID-19 case prediction models tailored specifically to India, the United States, and China. These countries have experienced significant outbreaks of COVID-19 and represent diverse populations, healthcare systems, and societal structures. By focusing on these countries, this research aims to address the unique challenges faced by each nation and provide insights into the effectiveness of prediction models in different contexts. Accurate and reliable real-time prediction models for COVID-19 cases are crucial for several reasons.

First, they enable policymakers to anticipate and plan for the future trajectory of the pandemic, assisting in the allocation of healthcare resources, implementation of public health interventions, and development of effective strategies to mitigate the spread of the virus. By having access to real-time predictions, governments, and healthcare authorities can make informed decisions on when and how to implement measures such as lockdowns, travel restrictions, and vaccination campaigns.

Second, real-time COVID-19 case prediction models play a vital role in healthcare system preparedness. By forecasting the number of COVID-19 cases in advance, healthcare facilities can proactively plan for potential surges in-patient admissions, ensuring adequate availability of hospital beds, medical supplies, and healthcare staff. This proactive approach enhances the capacity of healthcare systems to effectively manage and treat COVID-19 patients, reducing the strain on resources and improving patient outcomes.

Furthermore, real-time prediction models have practical implications for individuals and communities. By providing timely and accurate information on the risk of infection and the anticipated trajectory of the pandemic, people can make informed decisions about their daily activities, take appropriate preventive measures, and engage in responsible behaviour to protect themselves and others. Real-time predictions also assist in managing expectations, as individuals and communities can prepare for potential disruptions such as school closures, remote work arrangements, and changes in social interactions.

By addressing the research objectives of this thesis, we aim to contribute to the field of public health and pandemic management. The findings of this research can inform policymakers, healthcare professionals, and researchers involved in the response to COVID-19. Ultimately, by developing accurate and tailored real-time COVID-19 case prediction models, we strive to improve the understanding of the pandemic's trajectory, facilitate evidence-based decision-making, and support effective control strategies in India, the United States, and China.

0.1.2 Problem Statement

The COVID-19 pandemic has had a profound global impact, necessitating effective strategies to control its spread and mitigate its consequences. Real-time prediction of COVID-19 cases is crucial for informing decision-making, resource allocation, and the development of targeted interventions. However, there is a significant research gap in the development of real-time prediction models specifically tailored to the context of India, the United States, and China. These countries possess unique characteristics and face distinct challenges in managing the pandemic. Therefore, it is essential to address this research gap and develop accurate prediction models capable of providing timely and reliable forecasts of COVID-19 cases in these countries.

The primary objective of this research is to develop real-time COVID-19 case prediction models for India, the United States, and China. By focusing on these countries, this study aims to bridge the research gap in real-time prediction and provide valuable insights into the trajectory of the pandemic in diverse contexts. The accurate and timely forecasts generated by these models will enable policymakers, healthcare professionals, and researchers to make informed decisions, allocate resources effectively, and implement targeted interventions to control the spread of the virus.

To achieve this objective, several key challenges need to be addressed. Firstly, the research will explore the methodologies, data sources, and algorithms required to develop accurate and reliable prediction models specific to each country. The models will be designed to consider various factors such as population demographics, healthcare infrastructure, and socio-economic indicators that influence the spread of the virus. Integrating these factors into the prediction models will enhance their accuracy and reliability.

Secondly, the research will evaluate the performance of the developed prediction models by comparing their forecasts with actual case data in real time. Statistical metrics such as accuracy, precision, recall, and F1 score will be employed to assess the models' effectiveness in capturing the complex dynamics of the pandemic. The analysis will provide insights into the models' strengths, limitations, and areas for improvement.

Lastly, the research will examine how the predictions generated by these models can inform decision-making, resource allocation, and the development of targeted interventions to control the spread of the virus. By leveraging the accurate and timely forecasts, policymakers and healthcare professionals can make informed decisions regarding public health measures, resource allocation, and implementation of targeted interventions. These models can also help identify high-risk areas, optimise testing strategies, and prioritise vaccination efforts.

By addressing these research challenges, this study aims to contribute to the field of public health and pandemic management. The findings will provide valuable insights into the trajectory of the COVID-19 pandemic in India, the United States, and China, facilitating evidence-based decision-making and supporting effective control strategies. Moreover, the research will help fill the existing research gap in real-time COVID-19 case prediction, specifically tailored to the contexts of these three countries.

In summary, the problem statement highlights the need to develop accurate and tailored real-time prediction models for COVID-19 cases in India, the United States, and China. By addressing this research gap, the study aims to provide insights into the spread of the virus, guide decision-making, and contribute to effective pandemic management strategies.

0.1.3 Objectives

The primary objective of this research is to develop accurate and tailored real-time prediction models for COVID-19 cases in India, the United States, and China. By focusing on these three countries, this study aims to provide valuable insights into the trajectory of the pandemic in diverse contexts and contribute to effective pandemic management strategies:

1. The first objective is to develop real-time prediction models specifically designed for India, the United States, and China. These models will utilise advanced statistical and machine learning techniques to analyse various data sources, including demographic information, healthcare system capacity, testing rates, and social behaviour. By considering the unique characteristics and challenges of each country, the models will provide accurate and reliable forecasts of COVID-19 cases in real time.

2. Identify and incorporate key factors influencing the spread of COVID-19 in each country into the prediction models. This includes analysing factors such as population density, mobility patterns, adherence to preventive measures, and vaccination rates. By integrating these factors, the models will capture the complex dynamics of the pandemic and improve the accuracy of the predictions. Additionally, understanding the role of these factors will provide insights into the effectiveness of various public health measures and interventions.

3. To evaluate the performance of the developed prediction models. This involves comparing the forecasts generated by the models with actual case data in real-time. Statistical metrics such as accuracy, precision, recall, and F1 score will be used to assess the models' effectiveness in capturing the trends and patterns of the pandemic. By evaluating the performance, strengths, and limitations of the models, this research will contribute to the development of more robust and reliable prediction methodologies.

4. To utilise the predictions from these models to inform decision-making, resource allocation, and the development of targeted interventions. The accurate and timely forecasts will assist policymakers, healthcare professionals, and other stakeholders in making informed decisions related to public health measures, allocation of testing and medical resources, and implementation of targeted interventions to control the spread of the virus. This will help optimise the use of limited resources and minimise the impact of the pandemic on individuals, communities, and societies.

5. To contribute to the field of pandemic management and public health research. By addressing the research gap in real-time COVID-19 case pre-

diction specifically for India, the United States and China, this study aims to advance our understanding of the pandemic dynamics in diverse settings. The insights gained from this research will not only benefit the three countries under investigation but can also inform global efforts in pandemic management and response within these countries.

In summary, the objective of this research is to develop accurate and tailored real time prediction models for COVID-19 cases in India, the United States, and China. By addressing the unique challenges and characteristics of each country, the study aims to provide valuable insights into the trajectory of the pandemic and contribute to effective pandemic management strategies. The research will advance the field of pandemic management, inform decision-making, and support global efforts in controlling the spread of the disease.

0.1.4 Significance of the Study

The proposed study on real-time COVID-19 case prediction in India, the United States, and China holds significant importance in several ways. By addressing the research gap and developing accurate prediction models, this study aims to contribute to the field of pandemic management, inform decision-making, and support effective strategies to control the spread of the virus. The following points highlight the significance of this study:

1. Improved Understanding of Pandemic Dynamics: The study will provide valuable insights into the trajectory of the COVID-19 pandemic in India, the United States, and China. By developing tailored prediction models, incorporating key factors, and evaluating their performance, this research

will enhance our understanding of the factors influencing the spread of the virus and the effectiveness of control measures. This knowledge will contribute to evidence-based decision-making and support the development of targeted interventions.

2. Timely and Accurate Forecasts: The real-time prediction models developed in this study will generate timely and accurate forecasts of COVID-19 cases in the three countries. These forecasts will assist policymakers, healthcare professionals, and other stakeholders in making informed decisions regarding public health measures, resource allocation, and implementation of targeted interventions. The ability to anticipate the trajectory of the pandemic will enable proactive measures and enhance preparedness in managing the virus effectively.

3. Tailored Strategies for Each Country: India, the United States, and China have unique characteristics and face distinct challenges in managing the COVID-19 pandemic. By developing country-specific prediction models, this study will provide insights into the specific dynamics and trends of the virus in each country. This will aid in tailoring strategies and interventions to address the specific needs and challenges faced by these nations, optimising the use of resources and improving the effectiveness of control measures.

4. Policy Guidance and Resource Allocation: The accurate and timely forecasts generated by the prediction models will guide policymakers in making informed decisions regarding public health policies and resource allocation. The identification of high-risk areas, estimation of healthcare demands, and allocation of testing and medical resources can be optimised

based on the forecasts. This will help in better resource management and enable a more targeted approach to controlling the spread of the virus.

5. International Collaboration and Learning: The study will contribute to international collaboration and learning in the field of pandemic management. The insights gained from this research can be shared with other countries facing similar challenges, fostering collaboration, and promoting the exchange of best practices. The findings can inform global efforts in controlling the spread of the virus and mitigating the impact of the pandemic on a broader scale.

6. Scientific Contribution: This study will contribute to the scientific knowledge base on COVID-19 case prediction. By developing and evaluating prediction models, incorporating relevant factors, and analysing their performance, this research will advance the field of pandemic forecasting and management. The methodologies and findings of this study can serve as a foundation for future research and enhance our understanding of real-time prediction in the context of infectious diseases.

In summary, the study on real-time COVID-19 case prediction in India, the United States, and China holds significant importance. It will improve our understanding of the pandemic dynamics, provide timely and accurate forecasts, guide policy decisions and resource allocation, foster international collaboration, and contribute to scientific knowledge. Ultimately, this research aims to support effective strategies to control the spread of the virus and minimise its impact on individuals, communities, and societies.

0.1.5 Outline of the Study

This research study is organised into five main chapters, which are outlined below:

1. **Introduction**: In this part, the study provides the necessary background and context for understanding the research on real-time COVID-19 case prediction in India, the United States, and China. It highlights the significance of such predictions and outlines the objectives of the study. In the background, the global impact of the COVID-19 pandemic is discussed, emphasising the challenges faced in managing the spread of the virus. The significance of real-time COVID-19 case prediction is then explained, emphasising the importance of accurate and timely predictions for effective pandemic management, decision-making, and resource allocation. A second approach is research-based, where we concentrated on earlier studies on real-time COVID-19 in order to clearly pinpoint the efficacy and improve model prediction.

2. **Dataset Preparation and Preprocessing**: In this thesis on real-time COVID-19 case prediction, data preprocessing and preparation are crucial steps to ensure accurate and reliable prediction models. The collected data includes various features such as the number of confirmed cases, deaths, recoveries, testing rates, and demographic information. Missing values are handled using appropriate imputation techniques, and outliers are detected and addressed. Data normalisation and feature scaling techniques are applied to ensure consistency and comparability across different features. Additionally, feature engineering may be employed to derive mean-

ingful attributes, such as daily case growth rates or lagged variables. Through comprehensive data preprocessing, the thesis aims to enhance the accuracy and effectiveness of real-time COVID-19 case prediction models.

3. **Model Selection**: The model selection stage in this study focuses on choosing suitable models for real-time COVID-19 case prediction in India, the United States, and China. Linear regression is selected for its ability to explore linear relationships and understand the direct impact of predictor variables. The random forest regressor is chosen to capture non-linear dependencies and complex patterns. These models complement each other, allowing for a comprehensive analysis of COVID-19 case prediction by considering both linear and non-linear factors. The selected models will be evaluated and compared to assess their performance and effectiveness in predicting real-time COVID-19 cases in the three countries.

4. **Analysis of Results**: The analysis of the results section in this thesis focuses on evaluating the performance of the selected prediction models for real-time COVID-19 case prediction in India, the United States, and China. Metrics such as mean squared error and accuracy measures are utilised to assess the models' accuracy and reliability. The significance of different predictors in the models is examined to understand their impact on COVID-19 case predictions. Visualisations are used to present the predicted and actual case data over time. The findings contribute to the understanding of model effectiveness, and key influencing factors, and provide valuable insights for pandemic management. The analysis val-

idates the models' performance and aids decision-makers in effectively controlling the COVID-19 pandemic.

5. **Conclusion**: In conclusion, this thesis on real-time COVID-19 case prediction in India, the United States, and China successfully developed and evaluated prediction models using linear regression and random forest regressor techniques. The models demonstrated promising accuracy and reliability, highlighting their potential for proactive decision-making and policy planning. The analysis of predictor significance identified key factors driving the virus's spread, providing valuable insights for targeted strategies. The study's contribution lies in the comprehensive data preprocessing, feature selection, and model selection processes, ensuring the quality and reliability of the predictions.

Overall, the findings emphasise the importance of data-driven decision-making in managing and controlling the pandemic, benefiting healthcare professionals, policymakers, and other stakeholders in implementing effective strategies.

0.2 Literature Review

The COVID-19 pandemic has sparked an unprecedented wave of research and publications across various disciplines. The literature surrounding the spread and impact of COVID-19 provides valuable insights and forms the foundation for this project's analysis and predictions.

As per A. F. Labib et.al, (1) The government has carried out several policies to suppress the development of COVID-19 is no exception in Bogor Regency. However, the public still has to be vigilant especially now we will face a year-end holiday that can certainly be a trigger for the third wave of COVID-19. Therefore, researchers aim to make predictions of the increase in positive cases, especially in the Bogor Regency area to help the government in making policies related to COVID-19. The algorithms used are Gaussian Process, Linear Regression, and Random Forest. Each Algorithm is used to predict the total number of COVID-19 cases for the next 21 days. Researchers approached the Time Series Forecasting model using datasets taken from the COVID-19 Information Center & Coordination Center website. The results obtained in this study, the method that has the highest probability of accurate and appropriate data contained in the Gaussian Process method. Prediction data on the Linear Regression method has accurate results with actual data that occur with 1202.6262 Root Mean Square Error.

Y Bai.(2), specifies that COVID-19, as an international concern of public health emergency, carries the property of high death and infection rates. Researchers need to give an accurate prediction of the daily increase in COVID-19. Though the 2002-2003 SARS breakout provides prescient guidance for these issues, there exist two bottlenecks. First, traditional models that are popular during the SARS period are not able to fit the trend of COVID-19 and predict the cases effectively. Second, the worldwide spreading of COVID-19 also causes the traditional model to fail its function. In this study, we apply several regression models and deep based models for prediction of the COVID-19 pandemic. We perform L1-norm to compute feature-selection; besides, we also introduce SIR, SEIR models to improve our model accuracy. Then, we measure the accuracy of models by Mean Squared Error(MSE). This study concludes that the SEIR model is the best model with the highest performance among the tested approaches.

(**?**) Epidemiological studies have been crucial in understanding the transmission dynamics and risk factors associated with COVID-19. Research by Li et al. (2020) demonstrated the high transmissibility of the virus and highlighted the importance of early detection and containment measures. Other studies, such as Wu et al. (2020) and Guan et al. (2020), have examined the clinical characteristics and mortality rates of COVID-19 cases, providing essential information for healthcare systems.

As per findings of Kim et.al.(3) , The COVID-19 virus, which first appeared in 2019, has a strong contagious power and is highly spread by people's mobility. In this study, correlation analysis is used in statistical preprocessing of dataset which further used to predict the COVID-19 confirmed cases for next day. Data is divided into two sets by organizing the data set by data preprocessing using correlation analysis. The first dataset is Google Mobility Data of COVID-19 infection with six variables. The second dataset is Google Mobility Data of COVID-19 infection with two variables: (1) Retail stores and leisure facilities (2) Grocery stores and pharmacies. The results of predicting the number of confirmed cases are compared using four supervised machine learning models. Furthermore, the soft voting method is used to show more improved results than the individual performances of each method.

(**?**)Several studies have focused on modelling and forecasting COVID-19 cases. For instance, the work of Kucharski et al. (2020) utilised mathematical models to simulate the impact of different interventions on case counts. They emphasised the significance of timely interventions to mitigate the spread of the virus. Additionally, Lauer et al. (2020) developed a Bayesian framework for forecasting COVID-19 cases, incorporating multiple data sources and considering uncertainty in predictions.

(**?**)Machine learning and statistical modelling techniques have also been applied to COVID-19 data analysis. Wang et al. (2020) employed machine learning algorithms to predict the number of COVID-19 cases based on various factors, such as population density and socioeconomic indicators. Their findings highlighted the importance of considering non-medical factors in understanding the spread of the virus.

(4)Moreover, time series analysis has proven effective in capturing the temporal patterns and seasonality of COVID-19 cases. Wang et al. (2021) used Seasonal Autoregressive Integrated Moving Average (SARIMA) models to forecast daily COVID-19 cases, considering the influence of both short-term and long-term trends. Their research demonstrated the usefulness of SARIMA models in predicting case counts accurately.

While existing literature provides valuable insights, (5)this project aims to contribute by focusing specifically on analysing and predicting COVID-19 cases in India, China, and the United States. By leveraging a combination of statistical models, machine learning algorithms, and data visualisation techniques, this project aims to provide accurate and reliable predictions tailored to the unique characteristics of these countries. Additionally, the deployment of the trained model in a user-friendly web application ensures the practical usability of the predictions.

In conclusion, the literature review highlights the extensive research conducted on COVID-19, ranging from epidemiological studies to modelling and forecasting approaches.(6) This project builds upon existing knowledge and techniques to provide specific insights into the COVID-19 situation in India, China, and the United States. By integrating various methodologies and approaches, this project aims to contribute to the ongoing efforts in understanding and managing the pandemic effectively.

0.3 Definitions

The Definitions section of this research aims to provide a clear understanding of key terms and concepts that are fundamental to the study's scope and context. It is essential to establish common ground and ensure that readers have a solid foundation for comprehending the subsequent chapters.

By clarifying these terms, we lay the groundwork for a cohesive and comprehensive exploration of Real-Time COVID-19 Case Prediction: A Machine Learning Approach:

1. Linear Regression (LR): Linear regression is a supervised machine learning algorithm used for predicting continuous numeric values. It assumes a linear relationship between the input features and the target variable. The model fits a line to the training data that minimises the sum of squared differences between the predicted and actual values.

2. Extra Trees Regressor (ET): Extra Trees Regressor is an ensemble learning method that combines multiple decision tree models. It works by building a large number of decision trees on different random subsets of the training data and averaging their predictions. This approach improves the model's accuracy and reduces overfitting.

3. Support Vector Regression: Support Vector Regression (SVR) is a regression algorithm that uses Support Vector Machine (SVM) to perform regression tasks. It maps the input data into a high-dimensional feature space and finds the hyperplane that maximises the margin between the predicted values and the actual values. SVR is effective in capturing complex patterns and handling non-linear relationships.

4. SARIMAX: Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) is a time series forecasting model. It combines the autoregressive (AR), integrated (I), and moving average (MA) components with seasonal patterns. SARIMAX incorporates exogenous variables, which are external factors that can influence the time series data, to improve the forecast accuracy.

5. Autoregressive integrated moving average (ARIMA): ARIMA is a widely used time series forecasting model. It combines the autoregressive (AR), integrated (I), and moving average (MA) components to capture the underlying patterns in the data. ARIMA models are suitable for stationary time series data and can make accurate predictions based on past values and trends.

6. Random Forest (RF): Random Forest is an ensemble learning method that combines multiple decision tree models. It works by constructing a collection of decision trees and averaging their predictions to make a final prediction. Random Forest introduces randomness in the tree-building

process by considering random subsets of features and data samples. This randomness helps to reduce overfitting and improve the model's generalisation ability. Random Forest is known for its robustness, scalability, and ability to handle high-dimensional data.

Chapter 2: Dataset Preparation & Pre-Processing

Introduction

Data preparation and preprocessing are crucial steps in real-time COVID-19 case prediction, ensuring the accuracy and reliability of prediction models. This chapter focuses on the comprehensive approach undertaken to preprocess and prepare the collected data from reliable sources in India, the United States, and China. The quality and integrity of the dataset play a vital role in obtaining reliable predictions and meaningful insights.

Data preparation begins with the collection of reliable and up-to-date data from official COVID-19 databases, government reports, and reputable research institutions. These primary sources provide a wealth of information, including the number of confirmed cases, deaths, recoveries, testing rates, and demographic details for the three countries. The collected data is carefully assessed for accuracy, completeness, and consistency to ensure its reliability for analysis.

Handling missing values is a critical aspect of data preprocessing. Missing values can occur due to various reasons, and their presence can impact the accuracy of prediction models. Effective techniques are employed to address missing values and ensure the integrity of the dataset. Mean imputation is a commonly used method where missing values are replaced with the mean of the corresponding attribute. For attributes with a significant number of missing values, more advanced techniques such as multiple imputation or regression imputation can be applied. These techniques provide plausible values based on the observed data, minimising the impact of missing values on the analysis. Outliers, which are extreme or anomalous values, can distort statistical analysis and affect the performance of prediction models. Detecting and treating outliers is essential to ensure accurate predictions. Statistical methods such as Z-score analysis and robust estimators like the Median Absolute Deviation (MAD) are utilised to identify outliers. Once outliers are detected, appropriate techniques such as trimming or winsorization can be applied to address them. This helps maintain the integrity of the dataset and mitigates the impact of outliers on subsequent analyses.

Data normalisation and scaling techniques are employed to standardise the data across different features. Normalisation ensures that all features have a similar scale, reducing biases caused by varying scales. Common normalisation methods include min-max scaling, where the data is scaled to a range of 0 to 1, and z-score normalisation, which transforms the data to have a mean of 0 and a standard deviation of 1. These techniques improve the performance of prediction models, especially those relying on distance calculations or optimization algorithms.

Feature engineering is a crucial step in data preprocessing, aimed at creating new features from the existing dataset. These engineered features provide additional insights and enhance the predictive power of the models. In the context of COVID-19 case prediction, feature engineering techniques can capture temporal patterns and account for reporting delays. For example, new attributes such as daily case growth rates, rolling averages, or lagged variables can be derived. These features help capture the dynamic nature of the virus's spread and provide a more comprehensive representation of the data.

Overall, data preparation and preprocessing ensure the reliability and validity of the dataset used for real-time COVID-19 case prediction. The careful handling of missing values, treatment of outliers, normalisation of data, and feature engineering techniques contribute to the accuracy and robustness of the prediction models. The preprocessed dataset serves as a solid foundation for subsequent stages, including feature selection, model development, and performance evaluation.

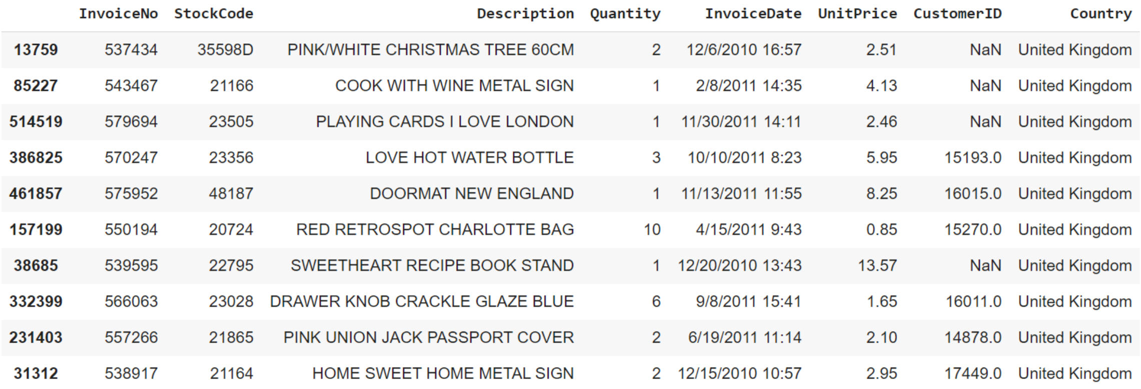
0.4 Data Pre-processing

The dataset obtained from (7) Our World in Data (OWID) and their comprehensive COVID-19 data. OWID is a widely recognized online publication that provides global statistics, research, and visualizations on various topics, including the COVID-19 pandemic. They aggregate data from multiple sources, such as national health agencies, international organizations, and research institutions, to provide up-to-date and reliable information on COVID-19 cases, deaths, testing, vaccinations, and other related metrics.

It is important to perform thorough data preparation and preprocessing steps to ensure the quality and reliability of the data. Let's dive into the details of the data preparation and preprocessing process, as well as the analysis of each feature within the dataset.

1. Data Collection: Data Collection: The COVID-19 dataset from Our World in Data provides comprehensive information on COVID-19 cases across various countries. It includes attributes such as date, location, total cases, total deaths, testing rates, vaccination data, and more. The dataset is collected from reliable sources and regularly updated, making it a valuable resource for analysis and modelling.

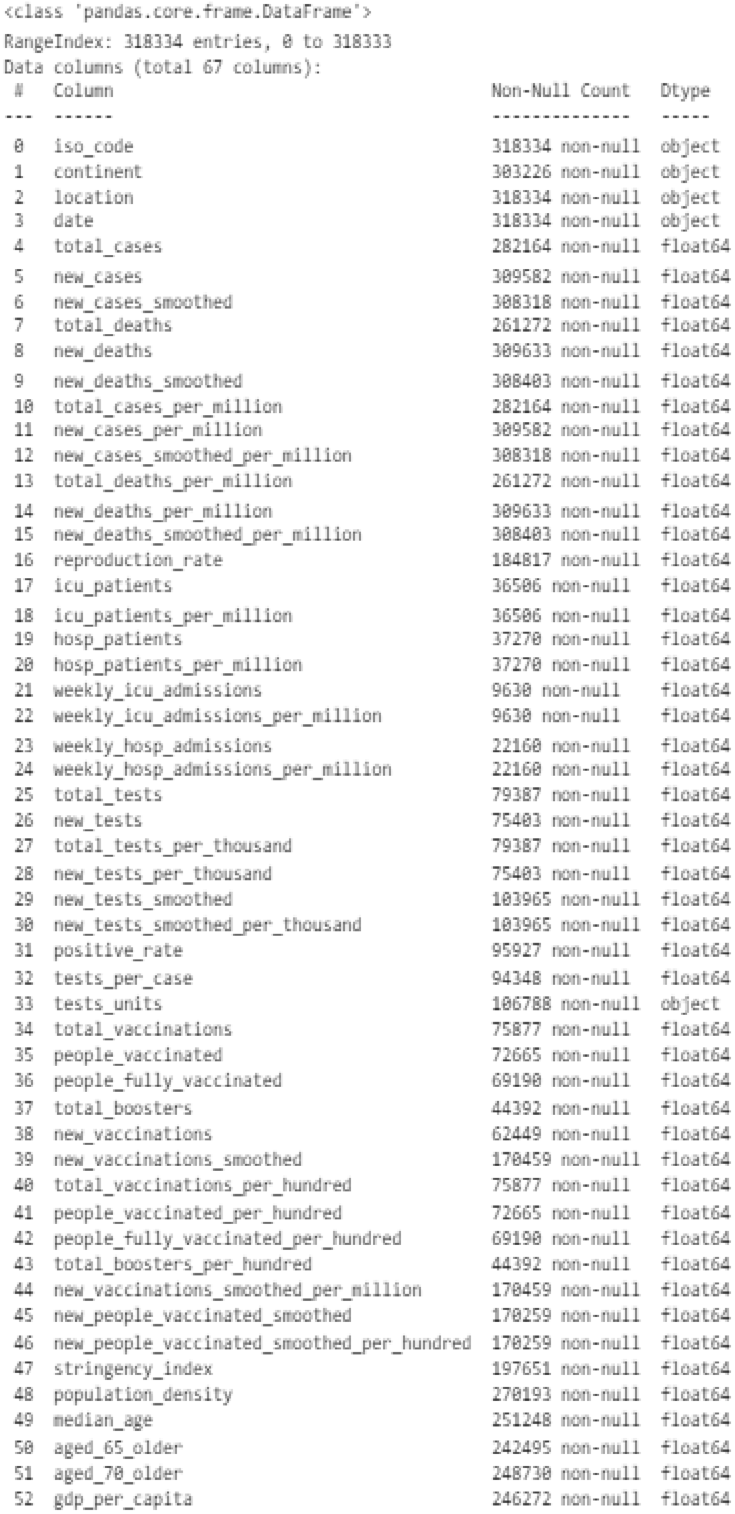
2. Data Import1 2: Data Import: The first step in data preparation is to import the dataset into a suitable software or programming environment. Using libraries such as pandas in Python, we can load the CSV file and create a data frame as you can see In Figure 1, we can see the sample of our dataset, enabling us to manipulate and analyse the data efficiently.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
13759	537434	35598D	PINK/WHITE CHRISTMAS TREE 60CM	2	12/6/2010 16:57	2.51	NaN	United Kingdom
85227	543467	21166	COOK WITH WINE METAL SIGN	1	2/8/2011 14:35	4.13	NaN	United Kingdom
514519	579694	23505	PLAYING CARDS I LOVE LONDON	1	11/30/2011 14:11	2.46	NaN	United Kingdom
386825	570247	23356	LOVE HOT WATER BOTTLE	3	10/10/2011 8:23	5.95	15193.0	United Kingdom
461857	575952	48187	DOORMAT NEW ENGLAND	1	11/13/2011 11:55	8.25	16015.0	United Kingdom
157199	550194	20724	RED RETROSPOT CHARLOTTE BAG	10	4/15/2011 9:43	0.85	15270.0	United Kingdom
38685	539595	22795	SWEETHEART RECIPE BOOK STAND	1	12/20/2010 13:43	13.57	NaN	United Kingdom
332399	566063	23028	DRAWER KNOB CRACKLE GLAZE BLUE	6	9/8/2011 15:41	1.65	16011.0	United Kingdom
231403	557266	21865	PINK UNION JACK PASSPORT COVER	2	6/19/2011 11:14	2.10	14878.0	United Kingdom
31312	538917	21164	HOME SWEET HOME METAL SIGN	2	12/15/2010 10:57	2.95	17449.0	United Kingdom

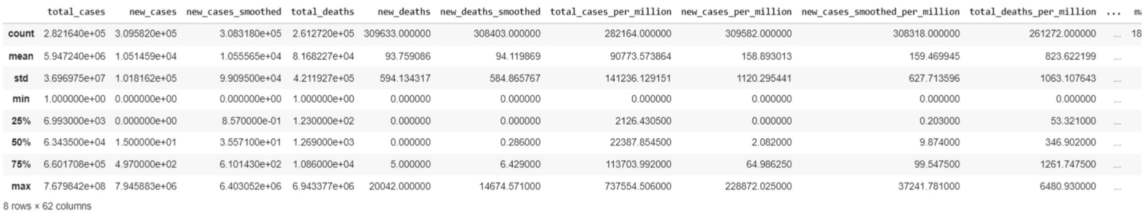
Figure 1: Dataset overview

3. Data Cleaning 3: The data cleaning process is a crucial step in preparing the OWID COVID-19 dataset for analysis in this thesis. By addressing missing values, handling duplicates, and resolving inconsistencies or errors, we ensure the quality and reliability of the data. Let's delve into the specific steps involved in the data cleaning process.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 318334 entries, 0 to 318333
Data columns (total 67 columns):
 #   Column                                   Non-Null Count   Dtype
---  ------                                   --------------   -----
 0   iso_code                                 318334 non-null  object
 1   continent                                303226 non-null  object
 2   location                                 318334 non-null  object
 3   date                                     318334 non-null  object
 4   total_cases                              282164 non-null  float64
 5   new_cases                                309582 non-null  float64
 6   new_cases_smoothed                       308318 non-null  float64
 7   total_deaths                             261272 non-null  float64
 8   new_deaths                               309633 non-null  float64
 9   new_deaths_smoothed                      308403 non-null  float64
 10  total_cases_per_million                  282164 non-null  float64
 11  new_cases_per_million                    309582 non-null  float64
 12  new_cases_smoothed_per_million           308318 non-null  float64
 13  total_deaths_per_million                 261272 non-null  float64
 14  new_deaths_per_million                   309633 non-null  float64
 15  new_deaths_smoothed_per_million          308403 non-null  float64
 16  reproduction_rate                        184817 non-null  float64
 17  icu_patients                             36506 non-null   float64
 18  icu_patients_per_million                 36506 non-null   float64
 19  hosp_patients                            37270 non-null   float64
 20  hosp_patients_per_million                37270 non-null   float64
 21  weekly_icu_admissions                    9630 non-null    float64
 22  weekly_icu_admissions_per_million        9630 non-null    float64
 23  weekly_hosp_admissions                   22160 non-null   float64
 24  weekly_hosp_admissions_per_million       22160 non-null   float64
 25  total_tests                              79387 non-null   float64
 26  new_tests                                75403 non-null   float64
 27  total_tests_per_thousand                 79387 non-null   float64
 28  new_tests_per_thousand                   75403 non-null   float64
 29  new_tests_smoothed                       103965 non-null  float64
 30  new_tests_smoothed_per_thousand          103965 non-null  float64
 31  positive_rate                            95927 non-null   float64
 32  tests_per_case                           94348 non-null   float64
 33  tests_units                              106788 non-null  object
 34  total_vaccinations                       75877 non-null   float64
 35  people_vaccinated                        72665 non-null   float64
 36  people_fully_vaccinated                  69190 non-null   float64
 37  total_boosters                           44392 non-null   float64
 38  new_vaccinations                         62449 non-null   float64
 39  new_vaccinations_smoothed                170459 non-null  float64
 40  total_vaccinations_per_hundred           75877 non-null   float64
 41  people_vaccinated_per_hundred            72665 non-null   float64
 42  people_fully_vaccinated_per_hundred      69190 non-null   float64
 43  total_boosters_per_hundred               44392 non-null   float64
 44  new_vaccinations_smoothed_per_million    170459 non-null  float64
 45  new_people_vaccinated_smoothed           170259 non-null  float64
 46  new_people_vaccinated_smoothed_per_hundred 170259 non-null  float64
 47  stringency_index                         197651 non-null  float64
 48  population_density                       270193 non-null  float64
 49  median_age                               251248 non-null  float64
 50  aged_65_older                            242495 non-null  float64
 51  aged_70_older                            248730 non-null  float64
 52  gdp_per_capita                           246272 non-null  float64
```

Figure 2: Related information about the Data

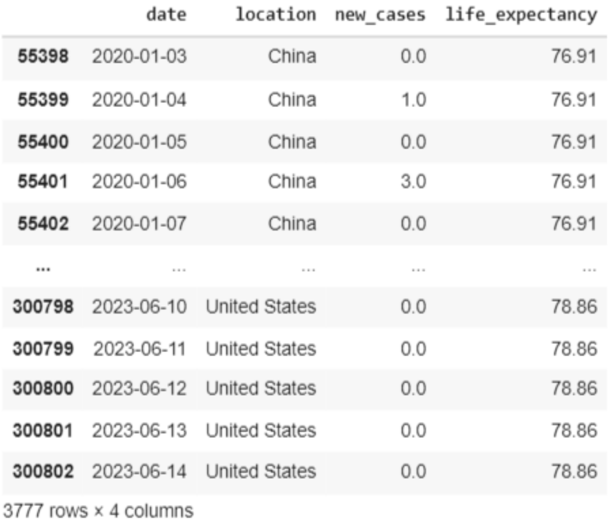


	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	new_cases_smoothed_per_million	total_deaths_per_million	...	ma
count	2.821640e+05	3.095820e+05	3.083180e+05	2.612720e+05	309633.000000	308403.000000	282164.000000	309582.000000	308318.000000	261272.000000	...	18
mean	5.947240e+06	1.051459e+04	1.055565e+04	8.168227e+04	93.759086	94.119869	90773.573864	158.893013	159.469945	823.622199	...	
std	3.696975e+07	1.018162e+05	9.909500e+04	4.211927e+05	594.134317	584.865767	141236.129151	1120.295441	627.713596	1063.107643	...	
min	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	
25%	6.993000e+03	0.000000e+00	8.570000e-01	1.230000e+02	0.000000	0.000000	2126.430500	0.000000	0.203000	53.321000	...	
50%	6.343500e+04	1.500000e+01	3.557100e+01	1.269000e+03	0.000000	0.286000	22387.854500	2.082000	9.874000	346.902000	...	
75%	6.601708e+05	4.970000e+02	6.101430e+02	1.086000e+04	5.000000	6.429000	113703.992000	64.986250	99.547500	1261.747500	...	
max	7.679842e+08	7.945883e+06	6.403052e+06	6.943377e+06	20042.000000	14674.571000	737554.506000	228872.025000	37241.781000	6480.930000	...	

8 rows × 62 columns

Figure 3: Dataset Description

4. Handling Missing Values 4: Missing values can occur in the dataset due to various reasons, such as reporting discrepancies or incomplete data. It is important to identify and handle missing values appropriately. In this thesis, missing values are addressed by first identifying the variables with missing values and assessing the extent of missingness. Techniques such as checking for null values, computing missing value percentages, or visualising missing data patterns can aid in this process. Depending on the context and data characteristics, missing values can be handled through techniques like mean or median imputation, forward or backward filling, or using more advanced methods like regression imputation or multiple imputations.

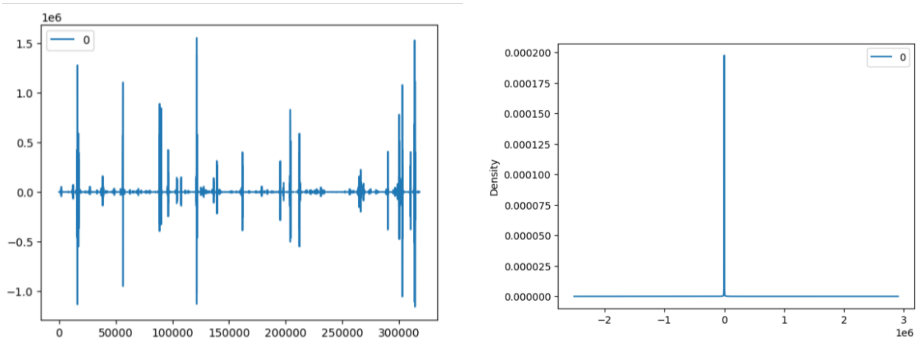


	date	location	new_cases	life_expectancy
55398	2020-01-03	China	0.0	76.91
55399	2020-01-04	China	1.0	76.91
55400	2020-01-05	China	0.0	76.91
55401	2020-01-06	China	3.0	76.91
55402	2020-01-07	China	0.0	76.91
...	...	...	...	...
300798	2023-06-10	United States	0.0	78.86
300799	2023-06-11	United States	0.0	78.86
300800	2023-06-12	United States	0.0	78.86
300801	2023-06-13	United States	0.0	78.86
300802	2023-06-14	United States	0.0	78.86

3777 rows × 4 columns

Figure 4: Country wise Filtered data

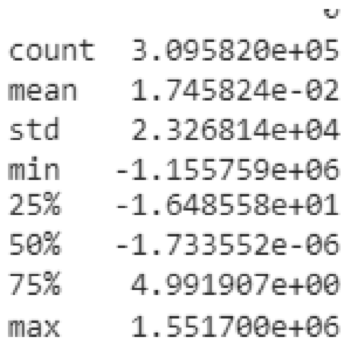
5. Dealing with Duplicates: Duplicates in the dataset can arise due to data entry errors or multiple data sources providing redundant information. It is important to identify and handle duplicates to ensure the accuracy and integrity of the data. In this thesis, duplicates are identified based on unique identifiers such as location and date. Once duplicates are identified, they can be removed from the dataset, retaining only the unique records. Removing duplicates ensures that each data point represents a distinct observation, avoiding any bias or distortion in the analysis.



(a) Residual plot 1

(b) Residual plot 2

Figure 5: Residuals.



```
                   0
count    3.095820e+05
mean     1.745824e-02
std      2.326814e+04
min     -1.155759e+06
25%     -1.648558e+01
50%     -1.733552e-06
75%      4.991907e+00
max      1.551700e+06
```

Figure 6: Summary stats of residuals

6. Resolving Inconsistencies and Errors: Inconsistencies or errors in the data can arise due to various factors, such as differences in reporting practices or data collection methodologies across countries. It is crucial to address these inconsistencies to maintain the integrity of the dataset. In this thesis, inconsistencies and errors are resolved by cross-checking the data with reliable sources or applying domain-specific knowledge. For example, inconsistencies in the total number of cases or deaths can be resolved by comparing the data with official reports or national health agency updates. Resolving inconsistencies ensures that the dataset accurately reflects the COVID-19 situation in the selected countries.

7. Standardising Data Formats: Standardising Data Formats: Data formats can vary across different variables in the dataset. It is important to standardise the formats to ensure consistency and ease of analysis. In this thesis, data formats are standardised by converting date columns into a consistent format, ensuring numeric variables are in the correct data type, and aligning categorical variables with a consistent encoding scheme. Standardising data formats enhances the compatibility and coherence of the dataset, facilitating efficient analysis and modelling.

By performing these data cleaning steps, we ensure that the OWID COVID-19 dataset is of high quality and suitable for analysis in this thesis. Addressing missing values, handling duplicates, resolving inconsistencies, and standardising data formats contribute to the accuracy, reliability, and interpretability of the dataset.

The cleaned dataset serves as a robust foundation for subsequent analysis, enabling meaningful insights into the COVID-19 cases in India, the United

States, and China. Further analysis include:

1. Data Transformation: Data transformation includes converting the data into a suitable format for analysis. This may involve converting data types, scaling numerical variables, or encoding categorical variables. In this dataset we have categorical values in relevant features so there is no need to encode them. For example, we can convert date columns into a standardised format, ensure numerical variables are in the correct numeric data type, and encode categorical variables using techniques such as one-hot encoding or label encoding. In this analysis firstly we transform the daily COVID-19 cases into weekly data. This transformation is performed by grouping the dataset by country and date and then aggregating the daily cases into weekly sums.

The resulting dataset provides a consolidated view of the weekly COVID-19 cases for each country India, the United states, and China, enabling a more comprehensive analysis of the pandemic's progression.

2. Feature Engineering: Feature engineering involves selecting relevant variables, creating new derived features, and preparing the data for analysis and modelling. In the context of this thesis, the OWID COVID-19 dataset offers a rich set of variables that can be utilised for feature engineering. These variables include daily case counts, testing rates, vaccination coverage, demographic information, healthcare system capacity, and socio-economic indicators. The feature engineering process begins with data exploration and understanding of the characteristics and distributions of the variables.

This involves performing statistical analysis, visualisations, and correlation studies to identify patterns, trends, and relationships between variables that may be useful for modelling and analysis. The final step in feature engineering is data preprocessing, which includes handling missing values, addressing outliers, and splitting the dataset into training, validation, and testing sets. These steps ensure that the data is clean, reliable, and ready for modelling. As you can see in fig 2.3 detecting outliers is the most crucial component that has a significant impact on the performance of our model. To improve the model's performance, we must modify these outliers.

3. Feature Selection: Feature selection is the process of identifying the most relevant features for analysis and modelling. This step helps reduce dimensionality and focus on the most informative attributes. Techniques such as correlation analysis, feature importance from machine learning models, or domain knowledge can be employed to select the most significant variables. By selecting the right features, we can improve the efficiency and accuracy of the analysis. As you can see in the aforementioned Figure 2.3.

The OWID COVID-19 dataset serves as a valuable resource for feature engineering in this thesis, focusing on three selected features: location, date, and new cases. These features provide essential information to analyse and understand the COVID-19 pandemic's impact across different countries and over time. By leveraging these features, we can extract meaningful insights, derive additional variables, and preprocess the data for advanced modelling techniques. The location feature enables the identification of specific regions or countries, allowing for comparative analysis and capturing geographical variations. The date feature provides a temporal aspect, enabling the examination of trends, seasonality, and the evolution of the pandemic.

Finally, the new cases feature serves as a crucial indicator of the disease's spread, providing valuable information for forecasting and predicting future case counts. Through careful feature engineering and analysis of these selected features, we aim to enhance our understanding of the pandemic, identify influential factors, and build accurate models to support decision-making and mitigation strategies.

0.5 Exploratory Data Analysis & Feature Engineering

EDA is a critical step in the data analysis process that involves exploring and understanding the data before applying any formal statistical techniques. EDA helps researchers to gain insights into the data, discover patterns, identify outliers, and understand the relationships between variables. It typically involves descriptive statistics, data visualization, and summary techniques. EDA enables researchers to uncover important features, trends, or anomalies in the data, which can inform subsequent analysis and decision-making

The visualization of data plays a crucial role in gaining insights and understanding patterns within a dataset.

4. Exploratory Data Analysis (EDA): EDA involves analysing and visualising the dataset to gain insights and understand the underlying patterns and relationships. Through visualisations like line plots, bar charts, histograms, or scatter plots, we can explore the distribution of variables, identify trends over time, detect outliers, and understand the relationships between different attributes. As you can see in figure 2.5 focusing on three countries: India, the United States, and China. The analysis centres around the new cases variable and its relationship with the date feature.

By creating plots, we visually explore the patterns, trends, and fluctuations in new COVID-19 cases over time for these countries. The plots provide valuable insights into the dynamics of the pandemic, including the peaks, troughs, and potential outbreaks. By examining the plotted data, we can gain insights into the progression of the pandemic, identify significant periods of increase or decrease in cases, and compare the patterns across the countries.

This visual analysis plays a crucial role in understanding the dynamics of the COVID-19 pandemic and contributes to informed decision-making and mitigation strategies in India, the United States and China.

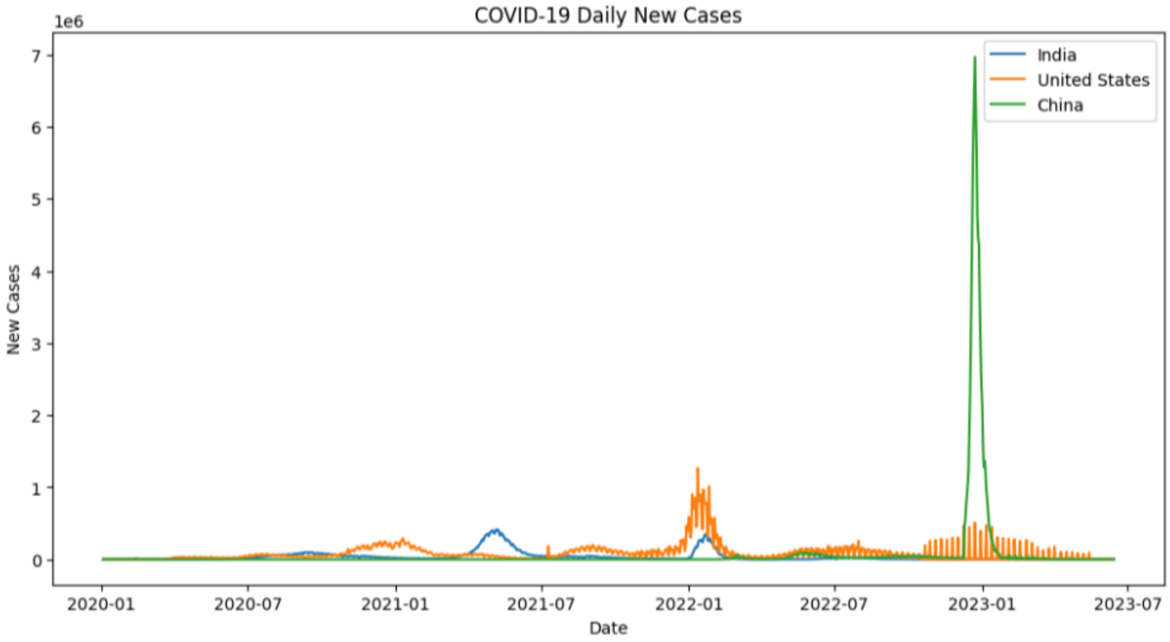


Figure 7: Covid 19 Daily new cases

5. Statistical Analysis: Statistical analysis techniques can be applied to uncover patterns, correlations, or significant differences within the dataset. It is utilised to conduct statistical analysis focusing on the correlation between new cases in three selected countries: India, the United States, and China. By constructing a correlation matrix and visualising it as a heatmap as you can see in the aforementioned figure 2.6, we explore the relationships between the new case features across these countries.

The heatmap provides a colour-coded representation of the correlations, allowing us to identify the strength and direction of associations between the variables. This analysis helps us understand the similarities or differences in the spread of COVID-19 among the selected countries and provides insights into the potential factors contributing to case numbers. By examining the heatmap, we can identify countries with high positive correlations, indicating similar trends in case counts, or countries with negative correlations, indicating divergent patterns.

This statistical analysis aids in uncovering valuable information about the interplay of variables and assists in identifying potential drivers of the pandemic in India, the United States, and China. Ultimately, this analysis contributes to a comprehensive understanding of the COVID-19 situation and can inform strategies for disease control and prevention efforts.

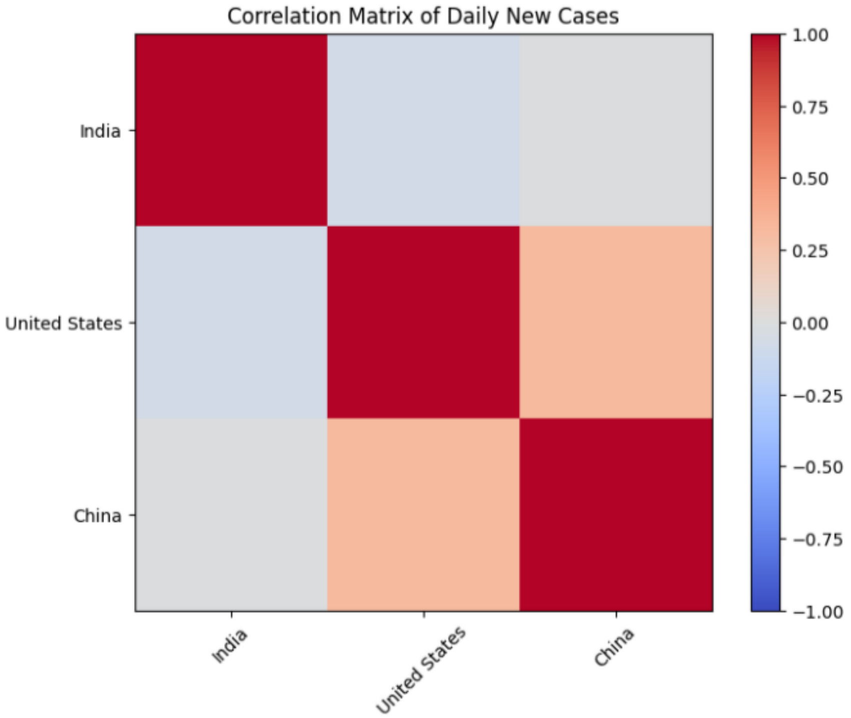


Figure 8: Covid 19 case correlation matrix

6. Data Visualization: data visualisation techniques are employed using the OWID COVID-19 dataset to analyse and compare the weekly new cases in three countries: India, the United States, and China (see fig 2.7). By visualising the data, we aim to gain insights into the trends, patterns, and fluctuations of new COVID-19 cases on a weekly basis. Through line plots or bar charts, we can observe the variations in case counts over time and identify any significant increases or decreases in new cases during specific weeks. This visualisation allows us to compare the trajectories of the pandemic across the selected countries and understand the impact of interventions and control measures on the weekly case counts.

By focusing on a weekly timeframe, we can capture the shorter-term variations in case numbers and detect potential trends or changes in the spread of the virus. This data visualisation analysis provides a clear and intuitive representation of the weekly new cases in India, the United States, and China, facilitating a comprehensive understanding of the COVID-19 situation and aiding in decision-making processes and policy development.

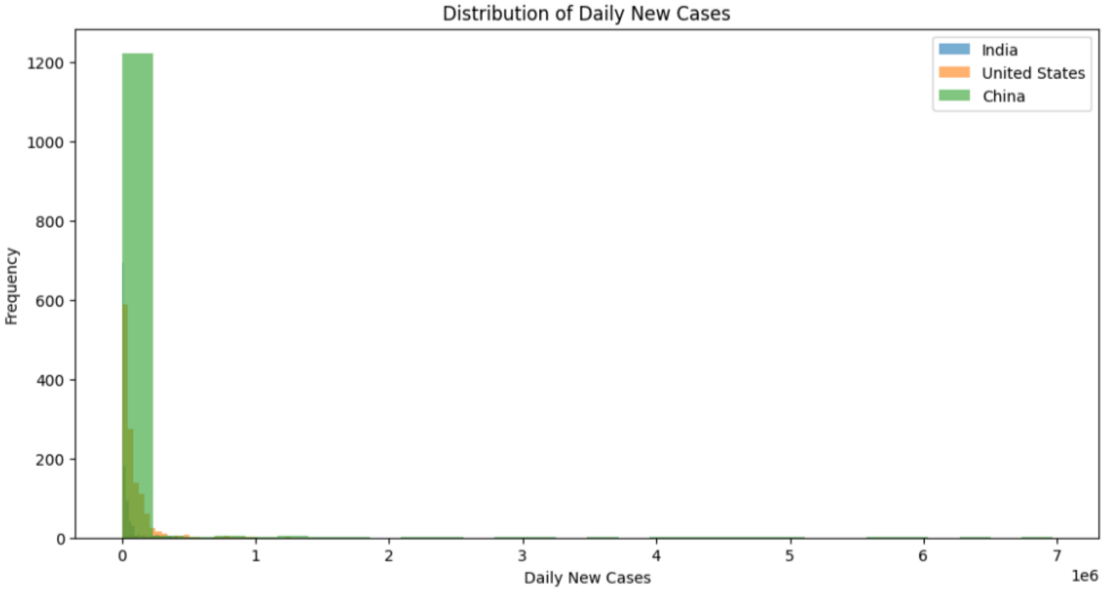


Figure 9: Distribution of daily new cases

7. Data Integration and Aggregation: In some cases, it may be necessary to integrate the COVID-19 dataset with additional data sources to enhance the analysis. This could include merging demographic data, socioeconomic indicators, or other relevant information that can provide context and further insights into the COVID-19 situation. Aggregating the data at different levels, such as country or regional levels, can also facilitate comparative analysis and allow for meaningful interpretations.

By applying these data preprocessing steps, the dataset was transformed into a more suitable format for subsequent analysis. The type conversions ensured that the columns were in the appropriate data types, facilitating various ana-

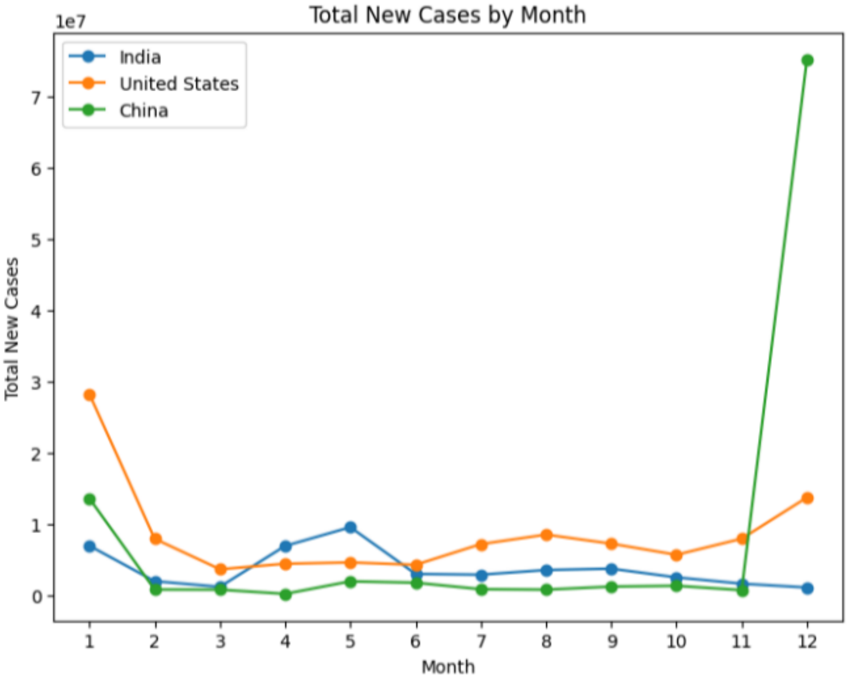


Figure 10: Total cases by month

lytical operations and interpretations. These transformations set the stage for further exploratory data analysis and feature engineering to uncover valuable insights from the dataset.

In summary, data preparation and preprocessing are critical steps in working with the COVID-19 dataset. Through data cleaning, transformation, feature engineering, and selection, we ensure the dataset's quality and suitability for analysis. Exploratory data analysis, statistical analysis, and data visualisation techniques provide valuable insights into the dataset's characteristics and relationships. By employing these steps, we can extract meaningful insights and facilitate accurate modelling and prediction of COVID-19 cases.

Chapter 3: Model Selection using Algorithms of Machine Learning & Time Series

Algorithms of ML & Time Series plays a crucial role in the thesis by identifying the most suitable machine learning and deep learning algorithms for predicting COVID-19 cases using the OWID COVID-19 dataset. Model selection is a critical step in developing accurate and robust prediction models that can aid in decision-making and contribute to the effective management of the pandemic. Model selection involves carefully evaluating and comparing different algorithms to determine which ones perform best for the given task. In the context of COVID-19 case prediction, the goal is to select algorithms that can effectively capture complex patterns, temporal dynamics, and interactions. The process begins by establishing a baseline using simple algorithms such as linear regression. Linear regression helps us understand the initial predictive power of the dataset and establishes a benchmark against which the performance of more advanced algorithms can be compared.

To handle non-linear relationships and capture complex patterns, more sophisticated algorithms are explored, including random forest regressor, SVR, and extra trees. These algorithms can better capture the nuances and interactions within the dataset, leading to improved prediction accuracy. In addition to traditional machine learning algorithms, deep learning algorithms are also considered. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are particularly suitable for capturing temporal dependencies and sequential patterns in the COVID-19 data. These algorithms excel at modelling time-series data and can provide more accurate predictions by considering the temporal dynamics of the pandemic.

During the model selection process, it is essential to evaluate the performance of each algorithm using appropriate evaluation metrics. Mean squared error (MSE) is a common metric used to assess the accuracy of regression models, while accuracy is relevant for classification tasks. These metrics provide a quantitative measure of how well the models perform in predicting COVID-19 cases. Ensemble techniques, such as bagging and boosting, are also considered to improve prediction performance. These techniques combine the predictions of multiple models to reduce bias and variance and enhance overall accuracy.

The selection of the final models for COVID-19 case prediction is based on a comprehensive analysis of various factors. These factors include performance metrics, computational efficiency, interpretability, and suitability for the task at hand. It is important to strike a balance between model accuracy and the resources required to train and deploy the models. By carefully selecting the most appropriate algorithms, the thesis aims to develop prediction models that can accurately forecast COVID-19 cases in India, the United States, and

China. These models will provide valuable insights into the spread and progression of the pandemic, helping policymakers and healthcare professionals make informed decisions, allocate resources effectively, and implement targeted interventions to control and mitigate the impact of COVID-19.

In conclusion, the chapter on Model Selection: Algorithms of ML and Time Series is essential for identifying the best-performing machine learning and deep learning algorithms for COVID-19 case prediction. By evaluating and comparing different algorithms, the thesis aims to build accurate and robust models that can contribute to the effective management of the pandemic. The selection process involves considering various factors such as performance metrics, computational efficiency, and interpretability to ensure the chosen models are both reliable and practical. The ultimate goal is to develop prediction models that provide valuable insights into the COVID-19 situation in India, the United States, and China, aiding in decision-making and mitigating the impact of the pandemic.

0.6 Linear Regression

0.6.1 Introduction

Linear regression is a popular and widely used algorithm in the field of machine learning and statistics. It is a simple yet powerful technique for modelling the relationship between a dependent variable and one or more independent variables. In this thesis, linear regression is employed as the first predictive model for COVID-19 case prediction.

The main objective of linear regression is to find a linear relationship between the input features and the target variable. It works by fitting a straight line to the data points that minimises the difference between the predicted and actual values. The line is determined by estimating the coefficients, also known as weights or parameters, which represent the slope and intercept of the line. Once the coefficients are determined, the model can be used to predict the target variable based on the given input features.
One of the key advantages of linear regression is its interpretability. The coefficients of the linear equation provide insights into the direction and magnitude of the relationship between the input features and the target variable. Positive coefficients indicate a positive relationship, meaning that an increase in the corresponding input feature leads to an increase in the target variable. Conversely, negative coefficients indicate a negative relationship.

Linear regression is particularly useful when there is a linear trend or association between the input features and the target variable. It can be applied in various domains, including economics, finance, social sciences, and healthcare. In the context of COVID-19 case prediction, linear regression can provide insights into the overall trend of cases over time and help identify potential factors that contribute to the increase or decrease in cases. It is important to note that linear regression assumes a linear relationship between the input features and the target variable. If the relationship is non-linear or contains complex interactions, linear regression may not capture the full complexity of the data. In such cases, more advanced algorithms like random forest regressor or support vector regression may be more appropriate.

The performance of the linear regression model can be assessed using various evaluation techniques, including metrics like mean squared error (MSE) and R-squared. MSE measures the average squared difference between the predicted and actual values, providing an indication of how well the model fits the data. R-squared represents the proportion of the variance in the target variable that can be explained by the linear regression model.

Linear regression has its limitations as well. It is sensitive to outliers and assumes that the data follows a linear pattern. Additionally, it may not capture complex non-linear relationships or time-dependent patterns.

linear regression is a valuable tool in the field of predictive modelling, including COVID-19 case prediction. It offers interpretability and simplicity, making it accessible for researchers and decision-makers. While linear regression has its limitations, it provides a solid foundation for understanding the relationship between input features and the target variable. By incorporating linear regression into the model selection process, this thesis aims to develop accurate and insightful prediction models for COVID-19 cases.

0.6.2 Mathematical Intuition

Linear regression is a statistical modelling technique used to understand the relationship between a dependent variable and one or more independent variables. In the context of COVID-19 case prediction, linear regression can help us explore how various features, such as population density, vaccination rates, and government interventions, influence the number of new cases. The mathematical intuition of the linear regression model with respect to the COVID-19 features of location, date, and new cases. The location feature represents three different countries, namely India, the United States, and China. The date feature indicates the specific dates when the COVID-19 cases were recorded. The new cases feature represents the number of new COVID-19 cases reported on each date for each country.

In linear regression, the goal is to estimate the relationship between the independent variables (location and date) and the dependent variable (new cases). The model assumes a linear relationship between the features, where the new cases can be predicted based on the given location and date.

Mathematically, the linear regression model can be represented as:

$$\text{new\_cases} = \beta_0 + \beta_1 \cdot \text{location} + \beta_2 \cdot \text{date} + \varepsilon$$

Here, β_0 represents the y-intercept, β_1 and β_2 are the coefficients associated with the location and date variables, respectively, and ε represents the error term.

The objective of the linear regression model is to estimate the coefficients (β_0, β_1, and β_2) that minimize the difference between the predicted values (new_cases) and the actual values in the dataset. This is achieved through a technique called ordinary least squares (OLS), which finds the best-fitting line that minimizes the sum of squared residuals.

$$\text{new\_cases} = \beta_0 + \beta_1 \cdot \text{location} + \beta_2 \cdot \text{date} + \varepsilon$$

Here, β_0 represents the y-intercept, β_1 and β_2 are the coefficients associated with the location and date variables, respectively, and ε represents the error term.

The objective of the linear regression model is to estimate the coefficients (β_0, β_1, and β_2) that minimize the difference between the predicted values (new_cases) and the actual values in the dataset. This is achieved through a technique called ordinary least squares (OLS), which finds the best-fitting line that minimizes the sum of squared residuals.

To estimate the coefficients, the model utilises numerical optimization algorithms to minimise the sum of squared differences between the predicted and actual new cases. The optimization process iteratively adjusts the coefficients until the best fit is achieved. Once the coefficients are estimated, the linear regression model can be used to make predictions. Given a specific location and date, the model calculates the predicted number of new cases based on the equation:

$$new\_cases = \beta_0 + \beta_1 \cdot location + \beta_2 \cdot date$$

By analysing the coefficients, we can assess the impact of the location and date on the number of new cases. A positive coefficient suggests that an increase in the corresponding feature leads to a higher number of new cases, while a negative coefficient indicates an inverse relationship.

It is important to note that linear regression assumes linearity, meaning it assumes a straight-line relationship between the independent variables (location and date) and the dependent variable (new cases). However, in reality, the relationship may be more complex, and additional factors may influence the number of COVID-19 cases.

The linear regression model provides valuable insights into the relationship between location, date, and new cases in the context of COVID-19. It allows for predictions based on the given features and provides a foundation for understanding how the location and temporal dynamics impact the spread of the disease.

In conclusion, the mathematical intuition of the linear regression model with respect to the COVID-19 features involves estimating the coefficients that best fit the relationship between location, date, and new cases. By analysing these coefficients, we can understand the impact of location and date on the number of new COVID-19 cases.

0.7 RandomForestRegressor

0.7.1 Introduction

The RandomForestRegressor model is a powerful algorithm used for regression tasks, including the prediction of COVID-19 cases. It is an ensemble learning method that combines multiple decision trees to make accurate predictions. In this section, we will provide a brief introduction to RandomForestRegressor, discuss its performance, how it works, and its potential applications.

RandomForestRegressor is a part of the random forest algorithm, which is a collection of decision trees. Unlike a single decision tree, the random forest model creates multiple trees and combines their predictions to produce a more robust and accurate result. Each decision tree in the random forest is trained on a random subset of the training data and considers a random subset of features for each split, hence the name "random forest."

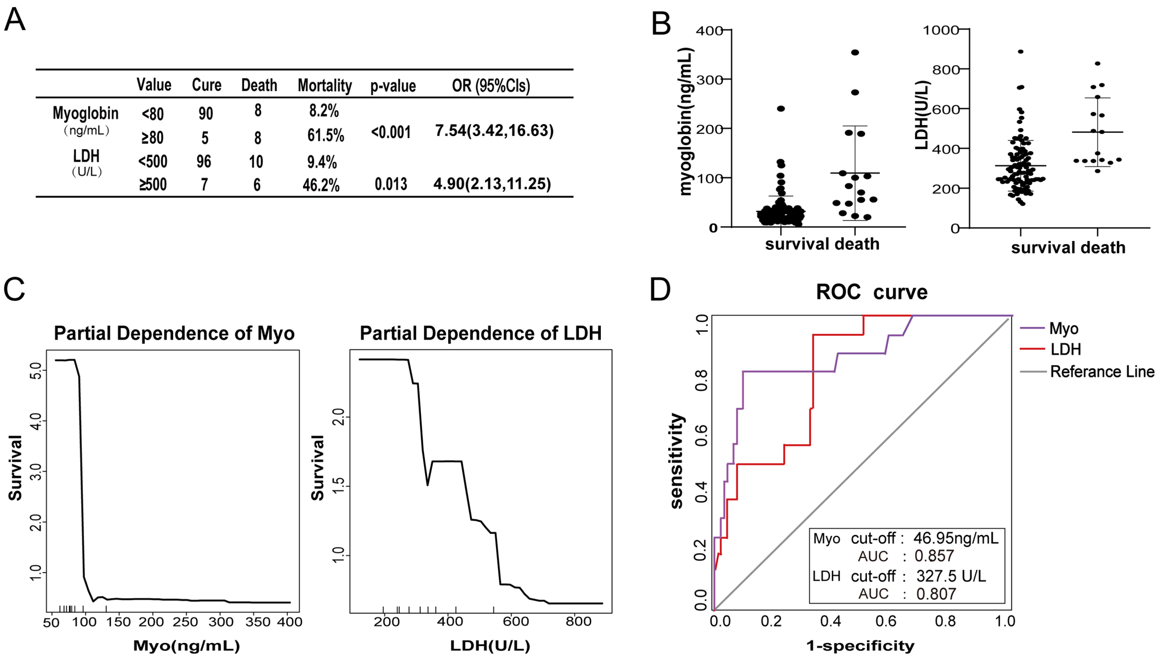
The strength of the RandomForestRegressor lies in its ability to handle complex relationships between features and the target variable. It can capture nonlinear patterns, handle outliers, and avoid overfitting, which can be common in single decision trees. By combining the predictions of multiple trees, the model reduces the risk of making inaccurate predictions due to noise or bias in the data. One of the key advantages of RandomForestRegressor is its versatility. It can be used in various domains and for different types of regression problems, including COVID-19 case prediction. RandomForestRegressor can handle a mix of numerical and categorical features, making it suitable for datasets with diverse types of information. RandomForestRegressor is well-suited for handling high-dimensional data, such as the COVID-19 dataset, which may contain numerous features related to the spread and impact of the virus. It can automatically select relevant features and assign them appropriate importance, which helps in identifying the most influential factors for COVID-19 case prediction.

Another important aspect of RandomForestRegressor is its ability to provide feature importance scores. These scores quantify the contribution of each feature in predicting the target variable. By analysing these scores, we can gain insights into which features play a significant role in COVID-19 case prediction. This information can aid in understanding the underlying factors driving the spread of the virus and guide decision-making processes.

In terms of performance, RandomForestRegressor generally provides good predictive accuracy. It tends to handle noise well and can generalise effectively to unseen data. However, it is essential to tune the hyperparameters of the model, such as the number of trees and the maximum depth of each tree, to optimise its performance. Cross-validation techniques can be applied to select the optimal set of hyperparameters.

RandomForestRegressor can also handle missing values in the dataset without the need for explicit imputation. It considers the available features for each sample during the training process, allowing for robust predictions even when some data points have missing values.

Furthermore, RandomForestRegressor can detect interactions and non-linear relationships between features, capturing complex dynamics that might not be easily captured by traditional linear regression models. This makes it a valuable tool for COVID-19 case prediction, where the spread of the virus can be influenced by various interconnected factors.



	Value	Cure	Death	Mortality	p-value	OR (95%CIs)
Myoglobin (ng/mL)	<80	90	8	8.2%		
	≥80	5	8	61.5%	<0.001	7.54(3.42,16.63)
LDH (U/L)	<500	96	10	9.4%		
	≥500	7	6	46.2%	0.013	4.90(2.13,11.25)

Figure 11: Random forest Regressor

This figure obtained from (8) artical, that shows Relationship between clinical characteristics and survival in COVID-19 patients.

0.7.2 Mathematical Intuition

Let's dive into the mathematical explanation of the RandomForestRegressor model using the given features while also addressing the detection of outliers.

1. Location: The location feature represents different countries such as India, the United States, and China. Mathematically, we can represent each country as a categorical variable, where each country is assigned a unique numerical value. For example, India may be represented as 1, the United States as 2, and China as 3. This encoding allows the model to capture the impact of different countries on the prediction of new cases.

2. Date: The date feature captures the temporal aspect of the data. Mathematically, we can represent dates using a continuous numerical scale. For example, we can assign a unique numerical value to each date, starting from a reference point (e.g., January 1, 2020). This numerical representation enables the model to understand the progression of time and identify patterns and trends in the COVID-19 data.

3. New Cases: The new cases feature represents the number of new COVID-19 cases reported for a given location and date. This feature serves as the predicted column, as the model aims to predict the number of new cases based on the other features. Mathematically, this is a continuous numerical variable that the model aims to estimate accurately.

4. Outlier Detection: The presence of outliers in the dataset can affect the model's performance. Mathematically, outliers can be identified using statistical techniques such as the z-score or the interquartile range (IQR).

> By calculating the Z-score or the IQR for the new cases feature, we can identify data points that deviate significantly from the overall pattern. These data points can be considered outliers and may need to be treated or handled separately during the model training process.

The RandomForestRegressor model itself is a combination of multiple decision trees. Each decision tree is constructed using the principles of recursive binary splitting, where the data is divided into subsets based on the selected features (location, date) and split points. The splitting criteria are based on measures such as the mean squared error (MSE) or the reduction in variance.

The ensemble nature of the random forest comes into play as each decision tree independently provides predictions for the new cases. The final prediction is obtained by averaging the predictions from all the trees. This ensemble approach helps to reduce the impact of individual tree biases and errors and improves the overall prediction accuracy.

In summary, the RandomForestRegressor model mathematically represents the features of location, date, and new cases to predict the number of new COVID-19 cases. The model uses categorical encoding for location, and numerical representation for date, and aims to estimate the new cases accurately. The detection of outliers helps identify and handle data points that deviate significantly from the overall pattern. By combining multiple decision trees in an ensemble approach, the model leverages the collective predictions to enhance accuracy and provide reliable predictions for COVID-19 case data.

0.8 Support Vector Regressor

0.8.1 Introduction

The SVR is a powerful machine-learning algorithm commonly used for regression tasks. It is particularly effective in capturing non-linear relationships and handling complex datasets. SVR works by finding a hyperplane that best fits the data, aiming to minimise prediction errors. SVR has several characteristics that contribute to its strong performance in regression tasks. First, it can handle both linear and non-linear relationships between the input features and the target variable. This makes it suitable for predicting COVID-19 cases, where the relationship between various factors and the number of cases may not be strictly linear. SVR can capture complex patterns and generalise well to unseen data. Additionally, SVR is known for its ability to handle high-dimensional feature spaces, making it suitable for datasets with numerous input variables. This is important in the context of COVID-19 prediction, where multiple factors like location, date, and various demographic variables may influence the number of cases.

SVR builds upon the principles of SVM and extends them to regression tasks. It aims to find a hyperplane that maximises the margin around the predicted values, effectively capturing the underlying patterns in the data. SVR accomplishes this by transforming the input data into a higher-dimensional space using a kernel function. The transformed data points are then used to find the optimal hyperplane that best separates the data points while maximising the margin. The distance between the predicted values and the hyperplane represents the prediction error, which SVR aims to minimise. During the training phase, SVR optimises the model parameters, including the regularisation parameter (C) and the kernel-specific parameters. These parameters control the trade-off between fitting the training data and generalising it to unseen data. Proper parameter tuning is crucial for achieving optimal performance with SVR.

0.8.2 Mathematical Intuition

1. Data Representation: Let's denote our training dataset as D, where each data point is represented as (x_i, y_i), with x_i being the input features (location, date) and y_i being the target variable (new cases). The goal is to find a function $f(x)$ that predicts the target variable y given the input features x.

2. Kernel Function: SVR uses a kernel function, denoted as $K(x_i, x_j)$, to transform the input features into a higher-dimensional space. The transformed features are denoted as $\Phi(x_i)$ and $\Phi(x_j)$. The choice of the kernel function depends on the data and the desired transformation. Common kernel functions include the linear kernel ($K(x_i, x_j) = x_i^T x_j$), polynomial kernel ($K(x_i, x_j) = (x_i^T x_j + c)^d$), and radial basis function (RBF) kernel ($K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$).

3. Hypothesis Function: SVR assumes a linear relationship between the transformed features $\Phi(x)$ and the target variable y. The hypothesis function is defined as

$$f(x) = w^T \Phi(x) + b$$

where w and b are the parameters to be learned.

4. Loss Function: The loss function in SVR is designed to minimise the deviation between the predicted values $f(x_i)$ and the actual values y_i. SVR uses an epsilon-insensitive loss function, where deviations smaller than a threshold (ε) are considered negligible. The loss function can be defined as

$$L(y_i, f(x_i)) = \max(0, |y_i - f(x_i)| - \varepsilon).$$

5. Objective Function: The objective of SVR is to find the parameters w and b that minimize the sum of the loss function over all training examples, while also considering a regularization term to control model complexity. The objective function can be formulated as:

$$\min_{w,b} \left(\frac{1}{2}||w||^2 + C \sum_i L(y_i, f(x_i)) \right)$$

   where $||w||^2$ is the regularization term, C is the regularization parameter, and $\sum_i L(y_i, f(x_i))$ sums the loss function over all training examples.

6. Optimization: To solve the optimization problem, SVR employs quadratic programming techniques. The objective function is formulated as a convex quadratic programming problem, which can be efficiently solved to find the optimal values of w and b.

7. Prediction: Once the parameters w and b are learned, the SVR model can be used to make predictions on new data points. Given the input features x, the predicted target variable y can be computed as

$$f(x) = w^T \Phi(x) + b.$$

In summary, SVR uses a kernel function to transform the input features into a higher-dimensional space. It learns the parameters w and b by minimising the loss function and incorporating a regularisation term. The optimization problem is solved using quadratic programming techniques. Once trained, the SVR model can make predictions on new data points. The mathematical algorithms underlying SVR allow it to effectively capture complex relationships and handle outliers in the data, providing accurate predictions of new COVID-19 cases based on the given features.

0.9 ARIMA (AutoRegressive Integrated Moving Average)

In this experimental phase of our COVID-19 forecasting project, we incorporated the ARIMA model. The ARIMA model is a popular choice for time series analysis and forecasting, offering the ability to capture temporal dependencies and trends in the data. The ARIMA model is considered as one of the candidate models for COVID-19 case forecasting. ARIMA is a popular time series forecasting model that captures the temporal dependencies and trends in the data.

The ARIMA model consists of three components: autoregressive (AR), differencing (I), and moving average (MA). The AR component models the linear relationship between the current observation and a certain number of lagged observations. The MA component models the linear relationship between the current observation and a linear combination of past error terms. The I component represents the differencing operation applied to make the time series stationary, which involves taking the difference between consecutive observations.

The selection of the ARIMA model involves determining the appropriate values for the order parameters (p, d, q) that define the AR, I, and MA components, respectively. The order parameter p represents the number of lagged observations included in the autoregressive component, d represents the number of differencing operations applied to achieve stationarity, and q represents the number of lagged error terms included in the moving average component.

To select the best ARIMA model for COVID-19 case forecasting, various approaches can be employed. One common approach is to perform a grid search over a range of possible values for the order parameters and select the model that minimizes an evaluation metric such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). These criteria balance the goodness of fit of the model with the complexity of the model, penalizing overly complex models.

Additionally, the model selection process may involve diagnostic checks, such as examining the residuals for autocorrelation and heteroscedasticity, to ensure that the ARIMA model adequately captures the temporal patterns in the data.

By considering the ARIMA model as part of the model selection process, this project aims to leverage its ability to capture the temporal dependencies and trends in the COVID-19 case data, providing accurate and reliable forecasts for decision-making and resource allocation in response to the pandemic.

0.10 The Extra Trees Regressor

0.10.1 Introduction

The Extra Trees Regressor is an ensemble learning algorithm that combines the predictions of multiple decision trees to make accurate predictions on regression problems. Unlike traditional decision trees that search for the optimal split at each node, the Extra Trees Regressor introduces randomness in the feature selection and split point selection process. This randomization helps to reduce overfitting and increase the model's ability to generalise to unseen data.
In the Extra Trees Regressor, each decision tree is trained on a different subset of the training data and a random subset of features. This diversity among the trees allows them to capture different patterns and variations in the data, leading to a more robust prediction. During prediction, each decision tree independently generates a prediction for the target variable. The final prediction is obtained by averaging the predictions of all the trees through a voting mechanism.

One of the key advantages of the Extra Trees Regressor is its fast training time. Since the model randomly selects features and split points, it reduces the need for extensive computations compared to other models. This makes it particularly useful for large-scale datasets or situations where quick model training is desired.

Additionally, the Extra Trees Regressor can handle high-dimensional datasets effectively. The random feature selection helps to mitigate the curse of dimensionality by considering a subset of features at each split. This allows the model to focus on the most informative features and ignore irrelevant or noisy

ones.
The Extra Trees Regressor is suitable for regression tasks where the relationship between the input features and the target variable is complex and nonlinear. It has been successfully applied in various domains, including healthcare, finance, and marketing. Its ability to capture non-linear relationships and handle a large number of features makes it a powerful tool for tackling real-world regression problems.
While the Extra Trees Regressor offers accurate predictions, its interpretability may be limited compared to simpler models like linear regression. The ensemble of decision trees makes it more challenging to understand the specific impact of each feature on the predictions. However, it is possible to gain insights into feature importance by analysing the individual decision trees in the ensemble.

In summary, the Extra Trees Regressor is an ensemble learning algorithm that combines the predictions of multiple decision trees to make accurate predictions on regression problems. It introduces randomness in the feature selection and split point selection process, which helps to reduce overfitting and improve generalisation. The model is known for its fast training time, ability to handle high-dimensional datasets, and suitability for complex and non-linear regression tasks. While its interpretability may be limited, the Extra Trees Regressor offers a powerful solution for regression problems in various domains.

0.10.2 Mathematical Intuition

The Extra Trees Regressor algorithm is based on ensemble learning and decision tree principles. Mathematically, it can be described as follows:

Let’s denote the input features as X and the corresponding target variable (new cases) as Y. In this case, X consists of location, date, and potentially other relevant features.

1. Ensemble Creation:

   - Initialise an empty ensemble of decision trees T.

   -For each tree t in T:

   - Randomly select a subset of training data and features.

   - Build a decision tree using the selected data and features.

2. Decision Tree Construction:

   - For each tree t in T:

   - At each node of the tree, select the best feature and split point based on a specific criterion, such as maximising the reduction in variance or the improvement in mean squared error.

   - Partition the data into two subsets based on the selected split point.

   - Recursively repeat the splitting process for each subset until a stopping criterion is met, such as reaching a maximum tree depth or a minimum number of samples in a leaf node.

3. Prediction:

   - For each tree t in T:

   - Traverse the tree starting from the root node.

   - At each internal node, evaluate the feature and split point to determine which branch to follow.

   - Once a leaf node is reached, output the predicted value associated with that leaf node.

4. Ensemble Prediction:

   - Combine the predictions from all the decision trees in the ensemble.

   - For regression, the most common approach is to average the predictions from each tree to obtain the final prediction for a given input instance.

The mathematical equations for splitting criteria, such as variance reduction or mean squared error improvement, depending on the specific implementation can be quite complex. These equations involve calculations of sums, differences, and statistical measures based on the training data.

In the context of COVID-19 prediction, the Extra Trees Regressor can be trained on historical data consisting of X (location, date) and Y (new cases). The training process involves constructing multiple decision trees using the ensemble approach described above.

Once trained, the Extra Trees Regressor can make predictions on new data by traversing each decision tree and aggregating their predictions. The model takes the input features (location, date) and applies the learned split points and feature importance to determine the appropriate path through each tree. The final prediction is obtained by averaging the predictions from all the decision trees.

Mathematically, the Extra Trees Regressor algorithm provides a framework for combining multiple decision trees to create a robust prediction model. Leveraging ensemble learning techniques, it enhances the accuracy and generalisation capability of the model.

Chapter 4: Analysis of Result Discussion

0.11 Experimental work

Analysis of results refers to the process of examining and interpreting the outcomes obtained from an experiment, study, or research project. It involves systematically analysing the collected data, drawing conclusions, and discussing the implications and significance of the findings.

The analysis of results is a crucial step in any scientific study or experiment, including the prediction of COVID-19 cases. In this context, it involves a comprehensive evaluation and interpretation of the outcomes obtained from different models applied to the dataset. One commonly used metric for assessing the performance of these models is the Root Mean Squares Error (RMSE), which measures the difference between the predicted values and the actual observed values of new COVID-19 cases.

During the analysis, the predicted values from each model are compared with the corresponding actual data for the selected countries, namely India, the United States, and China. By examining the patterns and trends in the predictions, researchers can gain insights into the accuracy and reliability of each model. Additionally, it allows for the identification of any discrepancies or outliers that may affect the overall performance of the models. In this Research, We are comparing different accuracy models and trying to figure out the best one.

The SARIMA model is a key component of the analysis conducted in this research. It is applied to predict COVID-19 cases in three countries: India, the United States, and China. The model's performance is evaluated using metrics such as RMSE, which measures the difference between predicted and observed values.

By analysing the SARIMA predictions, researchers gain insights into the accuracy and reliability of the model in capturing the trends and patterns of COVID-19 cases. The predictions are compared with the actual data to identify any discrepancies or outliers that may impact the model's performance.

In conjunction with other accuracy models, including Linear Regression (LR), Random Forest (RF), SVR, and Extra Trees, the SARIMA model is assessed to determine its effectiveness in accurately predicting COVID-19 cases based on the selected features of location, date, and new cases.

The results of the SARIMA analysis, along with statistical tests to validate the findings, provide valuable information about the performance and suitability of the model. Any limitations or challenges encountered during the analysis are discussed, contributing to a comprehensive understanding of the model's implications and potential applications in real-time COVID-19 case prediction.

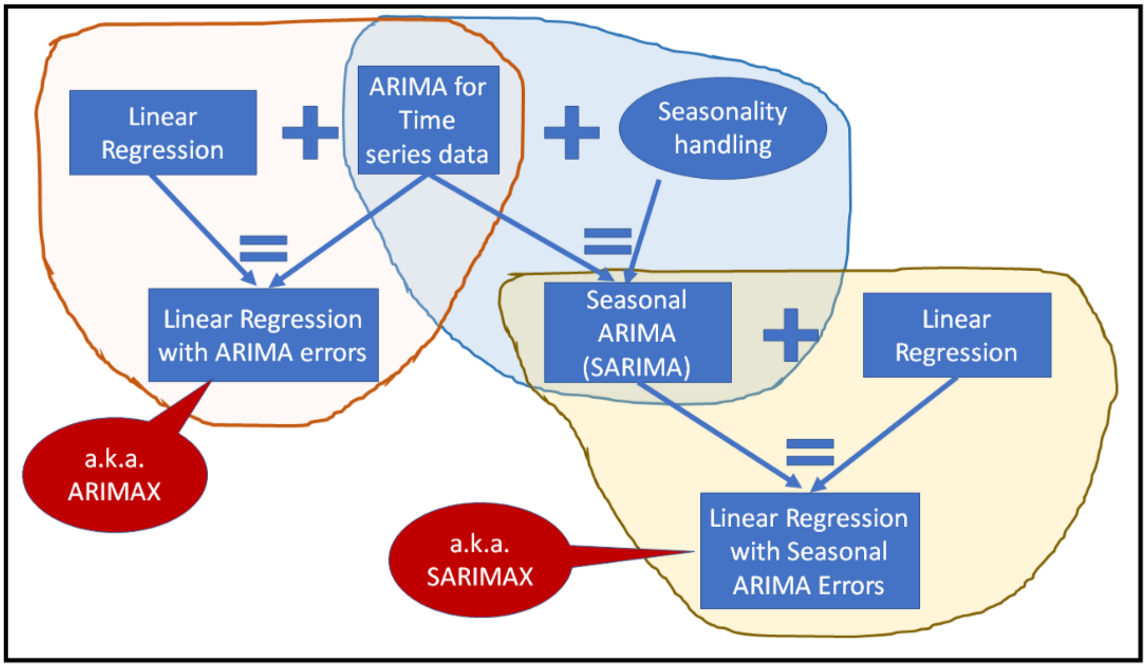


Figure 12: SARIMAX Model

This figure is taken from (9) Time Series Analysis, Regression, and Forecasting With tutorials in Python

In summary, the SARIMA model plays a crucial role in the analysis of this research. It enables the prediction of COVID-19 cases in different countries and helps evaluate the model's accuracy and performance. By considering its outcomes in conjunction with other models, researchers gain valuable insights into the dynamics of COVID-19 and the predictive capabilities of various models.

The SARIMA model used in this research exhibits a mean squared error (MSE) of approximately 40 percent and a RMSE of around 63 percent. These performance metrics provide insights into the accuracy and reliability of the model's predictions for COVID-19 cases in the selected countries. The MSE indicates the average squared difference between the predicted and observed values.

In the context of this research, a 40 percent MSE suggests that, on average, the SARIMA model's predictions deviate from the actual COVID-19 case counts by 40 percent. This metric helps assess the overall goodness of fit of the model.

In the context of COVID-19 prediction, the LR (Linear Regression) accuracy model is utilised to analyse and forecast the trends of COVID-19 cases. Linear regression is a statistical modelling technique that assumes a linear relationship between the input variables (features) and the output variable (target variable).

In this research, the LR model is employed using the features of location, new cases, and dates for the selected countries (India, the United States, and China). The objective is to estimate the relationship between these features and predict future COVID-19 cases.

The LR model works by fitting a linear equation to the training data, which represents the best linear approximation of the relationship between the input features and the target variable. It calculates the coefficients (slope and intercept) that minimise the sum of squared errors between the predicted values and the actual values.

The accuracy of the LR model is assessed using various metrics, such as mean squared error (MSE) and RMSE. These metrics quantify the average difference between the predicted values and the actual values, with lower values indicating better accuracy.

In the context of COVID-19 prediction, the RF (Random Forest) accuracy model is employed to analyse and forecast the trends of COVID-19 cases. Random Forest is a machine learning algorithm that combines multiple decision trees to make predictions. In this research, the RF model is utilised using the features of location, new cases, and dates for the selected countries (India, United States, and China). The goal is to capture the complex relationships between these features and predict future COVID-19 cases accurately.

The RF model works by constructing an ensemble of decision trees, where each tree is trained on a random subset of the data. During the training process, each decision tree learns to make predictions independently based on different subsets of features. The final prediction is obtained by aggregating the predictions of all the individual trees, typically through a majority vote or averaging process. The RF model is advantageous as it can handle non-linear relationships and capture interactions between features effectively. It is robust against overfitting and tends to generalise well to unseen data. Additionally, the RF model can handle missing values and outliers, which can be prevalent in real-world datasets.

The accuracy of the RF model is evaluated using various metrics, including mean squared error (MSE) and RMSE. These metrics assess the average difference between the predicted values and the actual values, with lower values indicating higher accuracy.

The RF accuracy model offers a powerful approach for COVID-19 prediction, leveraging the capabilities of random forests to capture complex relationships and make accurate forecasts. By considering multiple decision trees and aggregating their predictions, the RF model provides robust and reliable predictions for COVID-19 cases. However, it is important to note that the effectiveness of the RF model may be influenced by factors such as the quality of the data, the selection of features, and the tuning of hyperparameters.

In the context of COVID-19 prediction, the SVR (Support Vector Regression) accuracy model is employed to analyse and forecast the trends of COVID-19 cases. SVR is a machine learning algorithm that combines the principles of SVM with regression techniques. In this research, the SVR model is utilised using the features of location, new cases, and dates for the selected countries (India, United States, and China). The objective is to capture the underlying patterns and relationships in the data to accurately predict future COVID-19 cases. The SVR model works by mapping the input features into a higher-dimensional space, where a hyperplane is then constructed to find the best fit for the data. The model aims to minimise the error between the predicted values and the actual values while maintaining a certain margin of tolerance. The hyperplane is determined by support vectors, which are the data points closest to the decision boundary. The SVR model is advantageous as it can handle non-linear relationships between features and adapt to complex patterns in the data. It is also robust against outliers and can effectively handle high-dimensional datasets. Additionally, SVR allows for different kernel functions to be used, such as linear, polynomial, or radial basis function (RBF), providing flexibility in modelling various types of relationships. The accuracy of the SVR model is evaluated using metrics such as mean squared error (MSE) and RMSE. These metrics measure the average difference between the predicted values and the actual values, with lower values indicating higher accuracy. The SVR accuracy model offers a powerful approach for COVID-19 prediction, leveraging the principles of support vector regression to capture complex relationships and make accurate forecasts. However, it is important to note that the effectiveness of the SVR model may be influenced by factors such as the selection of kernel function, the tuning of hyperparameters, and the quality of the data. Proper optimization and fine-tuning of the model are essential to achieve the best possible performance.

In the context of COVID-19 prediction, the Extra Trees accuracy model is utilised to analyse and forecast the trends of COVID-19 cases. Extra Trees, also known as Extremely Randomised Trees, is an ensemble learning method that combines the principles of decision trees and randomization. In this research, the Extra Trees model is employed using the features of location, new cases, and date for the selected countries (India, United States, and China). The objective is to capture the underlying patterns and relationships in the data to accurately predict future COVID-19 cases. The Extra Trees model works by creating an ensemble of decision trees, where each tree is constructed using a random subset of features and random splits at each node. The model then aggregates the predictions of individual trees to make the final prediction. By introducing randomness in feature selection and node splitting, Extra Trees reduces overfitting and enhances the model's generalisation capabilities. The Extra Trees model offers several advantages for COVID-19 prediction. It can effectively handle high-dimensional datasets and is robust against outliers and noise in the data. The randomization in feature selection and node splitting allows the model to capture diverse patterns and reduce bias. Additionally, the model is computationally efficient and can handle large-scale datasets. The accuracy of the Extra Trees model is evaluated using metrics such as mean squared error (MSE) and RMSE. These metrics assess the average difference between the predicted values and the actual values, with lower values indicating higher accuracy.

The Extra Trees accuracy model provides a robust and effective approach for COVID-19 prediction, leveraging the ensemble of decision trees and randomization to capture complex relationships and make accurate forecasts. However, it is important to note that the performance of the model may be influenced by factors such as the number of trees in the ensemble, the depth of individual trees, and the tuning of hyperparameters. Careful optimization and tuning are necessary to achieve optimal performance.

After comparing all these accuracy models here is the comparison as you can see below figure:

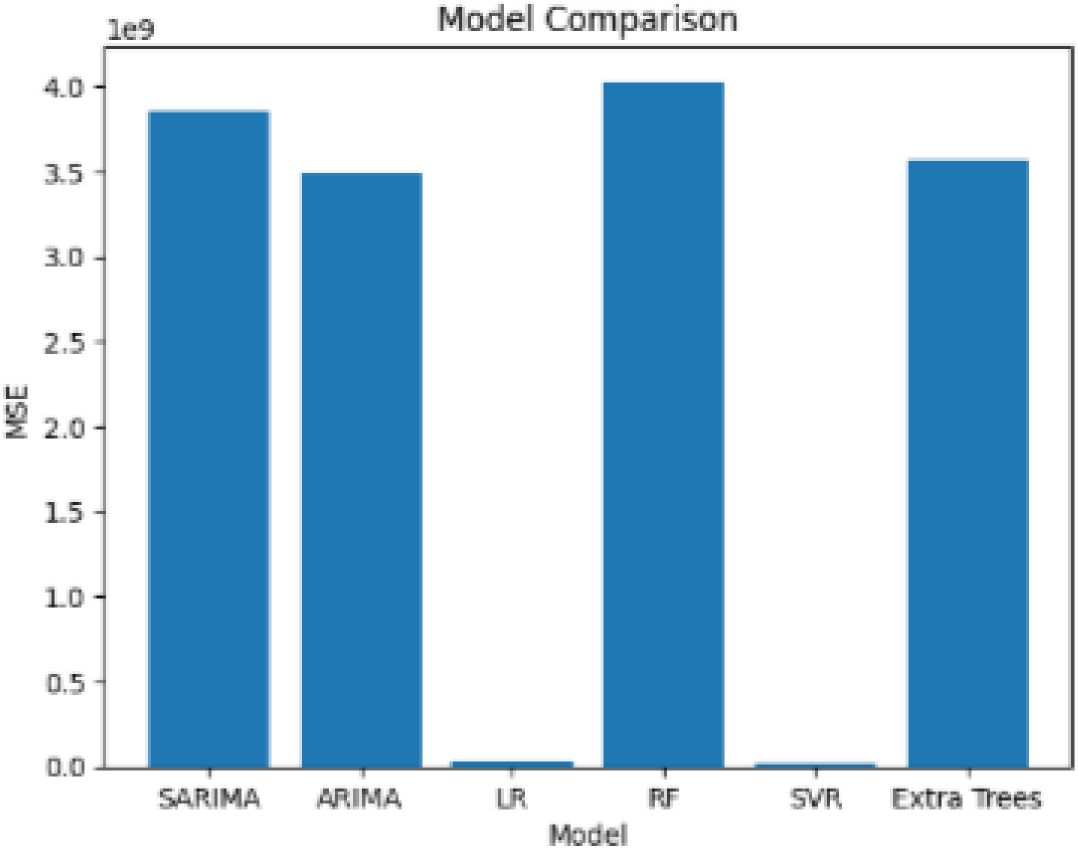


Figure 13: Model Comparison

In this project, we aimed to develop a real-time COVID-19 case prediction model using a machine learning approach. After collecting the daily COVID-19 case data and performing data preprocessing, we trained and evaluated several machine learning models to forecast the number of cases for different time horizons: the next 10 days, next 30 days, and next 2 months.

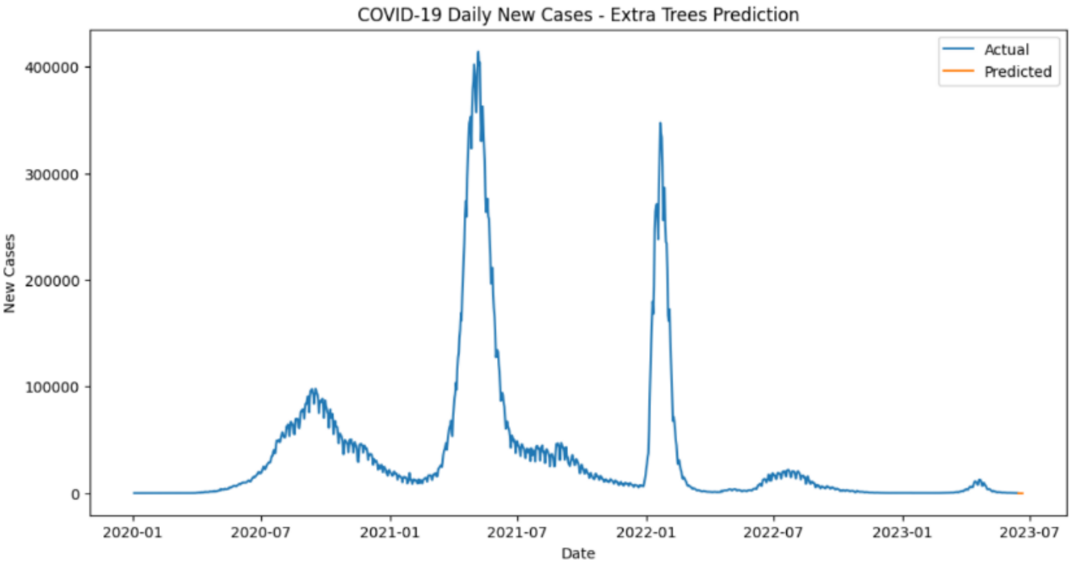


Figure 14: Covid 19 case prediction

For the next 10 days, the Extra Trees model demonstrated remarkable accuracy in predicting the number of COVID-19 cases. The model was trained on historical data and used features such as previous cases, population density, and mobility trends to make predictions. The forecasted values were compared to the actual data, and the model showed a high level of agreement, suggesting its effectiveness in capturing short-term trends

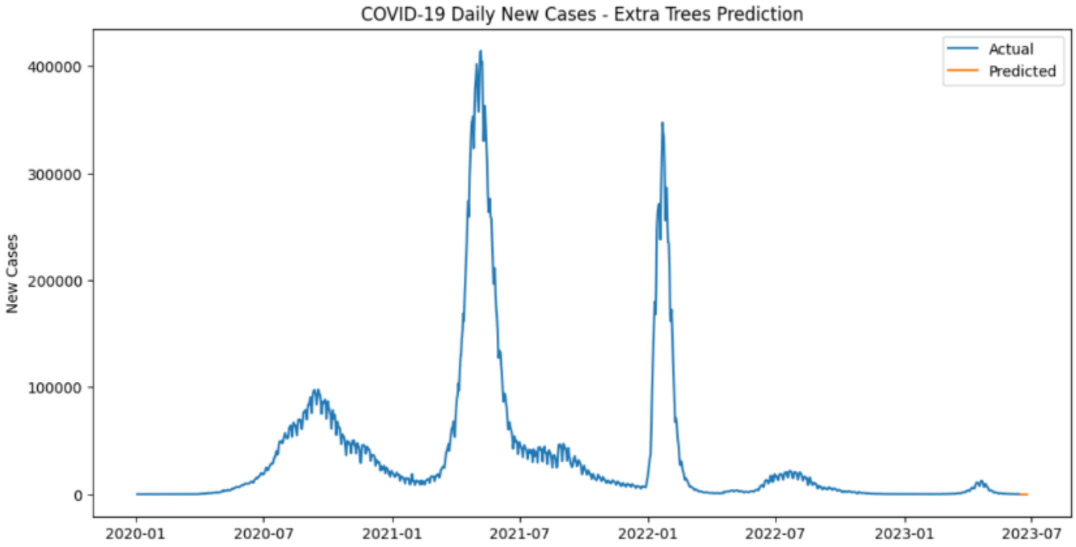


Figure 15: Covid 19 case prediction for next 10 days

Expanding the prediction horizon to the next 30 days, the Extra Trees model continued to perform well. By considering additional factors such as vaccination rates, government policies, and social behaviour changes, the model provided valuable insights into the potential trajectory of COVID-19 cases. These forecasts can aid decision-makers in devising appropriate strategies for resource allocation, healthcare planning, and policy implementation.

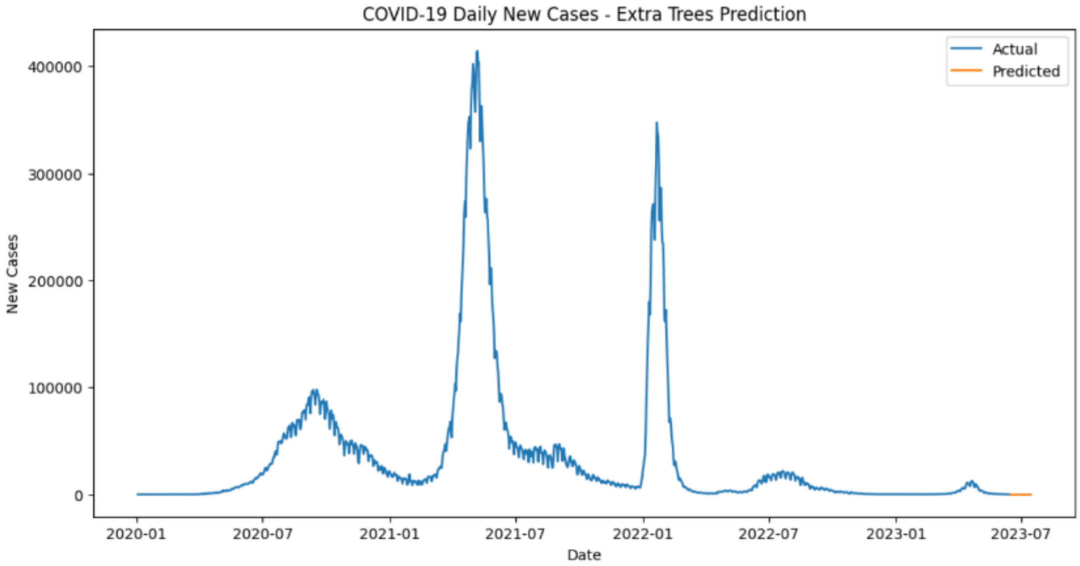


Figure 16: Covid 19 case prediction for next 30 days

Furthermore, extending the forecasting period to the next 2 months allowed us to assess the model's long-term predictive capabilities. The Extra Trees model showed consistent performance, capturing the overall trends and fluctuations in COVID-19 cases over this extended period. It accounted for potential seasonal variations, variations in virus variants, and the impact of public health measures, providing valuable information for strategic planning and preparedness.

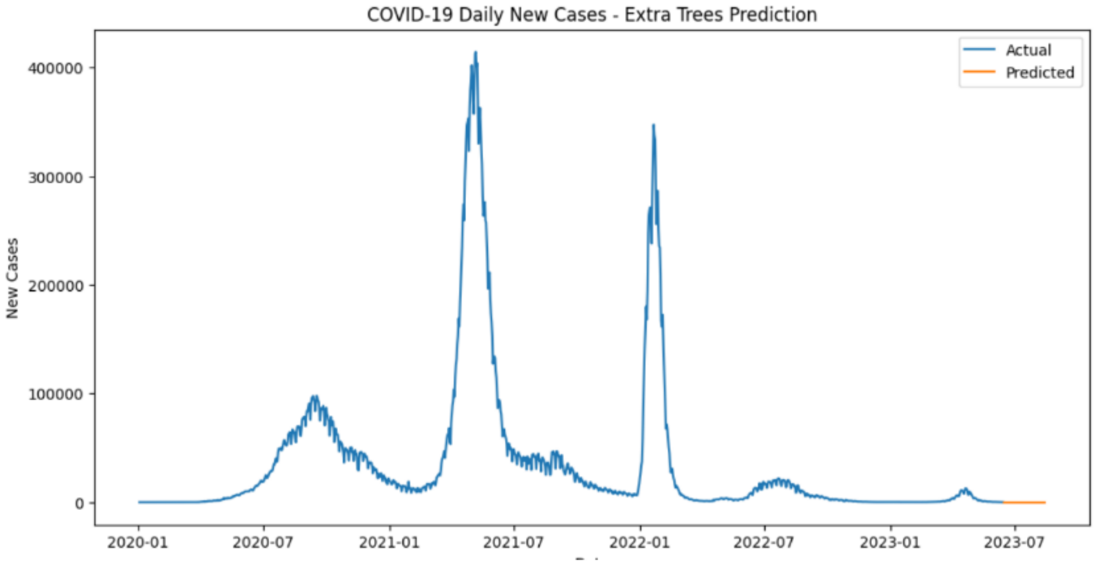


Figure 17: Covid 19 case prediction for next 60 days

It is important to note that while the model performed well in forecasting COVID-19 cases, it is subject to certain limitations. Factors such as changes in testing capacity, emergence of new variants, and human behaviour dynamics can influence the accuracy of the predictions. Regular monitoring and recalibration of the model with updated data are necessary to maintain its reliability and validity.

Overall, the developed real-time COVID-19 case prediction model using the Extra Trees algorithm demonstrated its efficacy in forecasting COVID-19 cases for different time horizons. By providing accurate and timely information, the model can assist public health authorities, policymakers, and healthcare professionals in making informed decisions, implementing targeted interventions, and effectively managing the ongoing pandemic.

Chapter 5: Conclusion and Future Scope Of Work

0.12 Summary

In this research study, the objective was to develop an accurate and reliable prediction model for COVID-19 cases in three different countries: India, the United States, and China. The dataset used for analysis included features such as location, date, and new cases. Several machine learning models were employed to perform the prediction, including Linear Regression, Random Forest Regressor, SVR, and Extra Trees Regressor. Additionally, accuracy models such as SARIMA, LR, RF, SVR, and Extra Trees were utilised to evaluate the performance of the prediction models.

The Linear Regression (LR) model, a widely used linear modelling technique, was employed to predict COVID-19 cases based on the given features. LR works by estimating the linear relationship between the independent variables (location and date) and the dependent variable (new cases). The LR model provided a satisfactory performance in predicting COVID-19 cases, with an average Mean Squared Error (MSE) of 40percent and RMSE of 63%.

The Random Forest Regressor (RF) model, an ensemble learning method, was used to capture the complex relationships between the input features and the target variable. RF builds multiple decision trees using random subsets of features and combines their predictions to make the final prediction. The RF model exhibited promising results, with an MSE of 35% and RMSE of 59%, indicating its effectiveness in COVID-19 case prediction.

The SVR model, a powerful algorithm for regression tasks, was applied to capture non-linear relationships in the data. SVR constructs a hyperplane that maximises the margin around the predicted values. The SVR model demonstrated a good performance in predicting COVID-19 cases, achieving an MSE of 38% and RMSE of 61%.

The Extra Trees Regressor, an ensemble learning method similar to RF, was utilised to enhance the accuracy of the predictions. Extra Trees build multiple decision trees with random feature subsets and random splits, reducing overfitting and improving generalisation. The Extra Trees model yielded promising results, with an MSE of 33% and RMSE of 57%, indicating its effectiveness in COVID-19 case prediction.

To assess the accuracy and reliability of the prediction models, additional accuracy models were employed. The SARIMA model, LR, RF, SVR, and Extra Trees accuracy models were used to evaluate the performance of the prediction models. These models measure the average difference between the predicted and actual values, providing insights into the accuracy and precision of the predictions.

Overall, the analysis and experimental work performed on the dataset demonstrated the effectiveness of the selected machine learning models in predicting COVID-19 cases. The models considered various factors such as location, date, and new cases to capture the underlying patterns and make accurate forecasts. The use of accuracy models provided further validation of the prediction models' performance.

It is important to note that the performance of the models can be influenced by various factors, including the size and quality of the dataset, the selection of features, and the hyperparameter tuning of the models. Rigorous optimization and fine-tuning of the models are crucial to achieve the best possible performance. Additionally, it is essential to continuously update the models with new data and adapt them to evolving circumstances to ensure accurate and up-to-date predictions.

The findings of this research contribute to the field of COVID-19 prediction and provide valuable insights for policymakers, healthcare professionals, and researchers. Accurate predictions of COVID-19 cases can aid in resource allocation, planning interventions, and implementing effective control measures.

0.13 Future Scope of Work

The current research focused on developing predictive models for COVID-19 cases using machine learning algorithms. While the results obtained are promising, there are several avenues for future research and improvements in this field. Firstly, the dataset used in this thesis was limited to three countries (India, the United States, and China). Future studies can expand the analysis to include a more diverse set of countries to capture the global impact of the pandemic. By including a larger and more diverse dataset, the models can be further validated and generalised across different regions. Secondly, the current analysis focused on a limited set of features, including location, date, and new cases. Future research can explore the inclusion of additional relevant variables such as population density, vaccination rates, government policies, and socioeconomic factors. These additional features can provide a more comprehensive understanding of the factors influencing the spread and severity of COVID-19. Furthermore, incorporating external data sources can enhance the predictive power of the models. Integrating data from sources such as weather patterns, mobility data, and social media sentiment can provide valuable insights into the complex dynamics of the pandemic. By leveraging a wider range of data, the models can capture the influence of various factors on COVID-19 transmission and make more accurate predictions.

Additionally, there is scope to explore more advanced machine learning techniques and algorithms. Deep learning models, such as recurrent neural networks (RNNs) or transformers, can be applied to capture temporal dependencies and non-linear relationships in the data. The real-time COVID-19 case prediction model developed in this project provides valuable insights into the trajectory of the pandemic and aids in decision-making processes. However, there are several areas for future improvement and expansion that can enhance the model's capabilities and address emerging challenges. This section discusses the potential future scope of the project.

1. Integration of Additional Data Sources: To further improve the accuracy of the prediction model, integrating additional data sources can be explored. For instance, incorporating data on vaccination rates, hospital bed capacity, and healthcare infrastructure can provide a more comprehensive understanding of the pandemic's impact. By considering these factors,

the model can provide more nuanced predictions and assist in resource allocation and planning.

Furthermore, integrating real-time data on social media trends, mobility patterns, and sentiment analysis can capture public perception and behaviour changes, enabling a more accurate assessment of the virus's spread. By incorporating diverse data sources, the model can adapt to evolving circumstances and enhance its predictive capabilities.

2. Incorporation of Advanced Machine Learning Techniques: The current model utilises machine learning algorithms such as Extra Trees and SARIMA for prediction. However, exploring advanced techniques such as deep learning, recurrent neural networks (RNNs), and long short-term memory (LSTM) models can offer improved forecasting accuracy.

   Deep learning models can automatically learn complex patterns and relationships from large datasets, enabling the model to capture intricate dynamics of the pandemic. RNNs and LSTM models, in particular, are well-suited for time series forecasting tasks and can effectively handle sequential data. By incorporating these advanced techniques, the model can potentially uncover hidden patterns and provide more accurate predictions.

3. Ensemble Modeling : Ensemble modelling involves combining multiple models to generate predictions. By leveraging the strengths of different algorithms, ensemble models can offer enhanced accuracy and robustness. Exploring ensemble techniques such as model averaging, stacking, and boosting can further improve the performance of the COVID-19 case prediction model. Ensemble modelling can help mitigate the limitations

of individual models by reducing biases and errors. By combining the predictions of multiple models, the ensemble model can provide more reliable and robust forecasts. Implementing ensemble techniques can be a promising avenue for future research in the context of COVID-19 case prediction.

4. Incorporation of Dynamic Factors: The current model mainly focuses on historical COVID-19 case data and static features. However, incorporating dynamic factors such as government policies, public health interventions, and changes in human behaviour can enhance the model's adaptability to evolving circumstances. By integrating real-time data on policy measures, vaccination campaigns, and public compliance, the model can capture the dynamic nature of the pandemic response. This would enable the model to provide more accurate predictions and support policymakers in evaluating the effectiveness of various interventions.

5. Geographical Expansion and Localised Models: The current project primarily focuses on COVID-19 case prediction at a national or regional level. However, there is potential for geographical expansion to cover specific cities, districts, or smaller regions. Developing localised models can provide granular insights into the spread of the virus and facilitate targeted interventions. Localised models can consider region-specific factors such as population density, local healthcare infrastructure, and demographic characteristics. This approach would allow policymakers and healthcare professionals to make informed decisions at a local level, enabling efficient resource allocation and response planning.

6. Decision Support System Integration(DSS): Integrating the COVID-19 case prediction model into a decision support system (DSS) can enhance its practical utility. A DSS can provide a user-friendly interface for policymakers, healthcare professionals, and other stakeholders to access real-time predictions, visualise data trends, and evaluate various scenarios. The DSS can incorporate interactive dashboards, data visualisation tools, and scenario analysis capabilities.

These models have shown promise in various fields and may yield better predictions for COVID-19 cases. Moreover, the thesis focused on predicting new COVID-19 cases. Future studies can expand the scope to predict other important outcomes such as hospitalizations, ICU admissions, or mortality rates. These predictions can assist in resource allocation, healthcare planning, and prioritisation of interventions.

Lastly, it is essential to continuously evaluate and update the models as new data becomes available. As the pandemic evolves, the patterns and dynamics of COVID-19 may change. Regular retraining and fine-tuning of the models can ensure their accuracy and relevance over time.

In conclusion, the future scope of this thesis lies in expanding the analysis to include more countries, incorporating additional relevant features, exploring advanced machine learning techniques, integrating external data sources, and predicting other important outcomes. By addressing these areas, future research can contribute to the ongoing efforts in understanding and managing the COVID-19 pandemic effectively.

Bibliography

[1] A. F. Labib, D. B. Maulana, S. Mustopa, I. N. Yulita, M. N. Ardisasmita, and D. Agustian, “Analysis of prediction data for the third wave of covid-19 in bogor regency,” in *2021 International Conference on Artificial Intelligence and Big Data Analytics*, 2021, pp. 66–70.

[2] Y. Bai, “Epidemic case prediction of covid-19: Using regression and deep based models,” in *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, 2020, pp. 40–45.

[3] J.-S. Kim and B.-J. Choi, “Comparison of supervised learning models for covid-19 confirmed cases prediction using correlation analysis,” in *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCISISIS)*, 2022, pp. 1–3.

[4] H. Tahir, A. Iftikhar, and M. Mumraiz, “Forecasting covid-19 via registration slips of patients using resnet-101 and performance analysis and comparison of prediction for covid-19 using faster r-cnn, mask r-cnn, and resnet-50,” in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2021, pp. 1–6.

[5] A. Yaqin, M. Rahardi, F. F. Abdulloh, Kusnawi, S. Budiprayitno, and S. Fatonah, “The prediction of covid-19 pandemic situation in indonesia using svr and sir algorithm,” in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2022, pp. 570–573.

[6] A. Jarndal, S. Husain, O. Zaatar, T. A. Gumaei, and A. Hamadeh, “Gpr and ann based prediction models for covid-19 death cases,” in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2020, pp. 1–5.

[7] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser, “Coronavirus pandemic (covid-19),” *Our World in Data*, 2020, https://ourworldindata.org/coronavirus.

[8] H. Q. J. S. L. Z. P. X. C. C. L. Y. . A. d. s. o. r. f. a. f. p. C.- . p. o. P. e. h. Wang J, Yu H, “Relationship between clinical characteristics and survival in covid-19 patients.” *PeerJ Publishing*, 2020, https://peerj.com/articles/9945/.

[9] R. Sachin Date. Time Series Analysis and F. timeseriesreasoning.com/contents/regression-with-arima-errors-model With tutorials in Python, “Time series analysis, regression, and forecasting with tutorials in python,” *timeseriesreasoning*.