Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software Department

Mobility and the Spread of Corona Virus Pandemic: Forecasting, Discovering, and Visualizing Based on Machine Learning Techniques

A Thesis

Submitted to the Council of the College of Information Technology at University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology/Software

By

Dhuha Hussein Mohamed Jawad

Supervised by

Prof. Dr. Eman Salih Sagban

2022 A.C. **1443 A.H.**

بِسْمِ اللَّهِ الرحمن ا لرَّحِيمِ

﴿ اللَّه نُورُ السَّمَاوَاتِ وَالْأَرْضِ مَثَلُ نُورِهِ كَمِشْكَاةٍ فِيهَا مِصْبَاحٌ الْمِصْبَاحُ فِي زُجَاجَةٍ

الزُّجَاجَةُ كَأَنَّهَا كَوْكَبٌ دُرِّيٌّ يُوقَدُ مِنْ شَجَرَةٍ مُبَارَكَةٍ زَيْتُونَةٍ لَا شَرْقِيَّةٍ وَلَا غَرْبِيَّةٍ

يَكَادُ زَيْتُهَا يُضِيءُ وَ لَوْ لَمْ تَمْسَسْهُ نَارٌ نُورٌ عَلَىٰ نُورٍ يَهْدِي اللَّه لِنوُرِهِ منَ يشَ اءُ

وَيَضرْبِ اللَّه ا لْأَمْثَالَ لِلنَّاسِ وَاللَّه بِكُلِّ شَيْءٍ عَلِيم ﴾

صدق الله العلي العظيم

سورة النّور \ آية 35

Supervisor Certification

I certify that this thesis entitled "***Mobility and the Spread of Corona Virus Pandemic: Forecasting, Discovering, and Visualizing Based on Machine Learning Techniques"***

was prepared under my supervision at the Department of Software / College of Information Technology/ University of Babylon, by **Dhuha Hussein Al-Jubory** as partial fulfillment of the requirements for the degree of **Master in Information Technology.**

Signature:

Name: **Dr. Eman Salih Al-Shamery**

Title: **Professor.**

Date: / / 2022

The Head of the Department Certification

In view of available recommendations, I forward this thesis entitled "***Mobility and the Spread of Corona Virus Pandemic: Forecasting, Discovering, and Visualizing Based on Machine Learning Techniques "*** for debate by the examination committee.

Signature:

Name: **Assistant Prof. Dr. Ahmed Saleem Abbas**

Title: **Head of Software Department, College of Information Technology, University of Babylon.**

Date: / / 2022

Certification of the Examination Committee

We the undersigned, certify that (**Dhuha Hussein Mohamed Jawad**) candidate for **Master Degree in Information Technology - Software**, has presented her thesis of the following title (***Mobility and the Spread of Corona Virus Pandemic: Forecasting, Discovering, and Visualizing Based on Machine Learning Techniques***) as it appears on the title page and front cover of the thesis that the said thesis is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: May 12,2022.

Signature:

Name:**Dr. Yahya Mahdi Hadi Al-Mayali**

Title: **Professor.**

Date**: / / 2022**

(Chairman)

Signature:

Name: **Dr.Ali Hadi Hasan AL-Najar**

Title: **Assistant Professor**

Date**: / / 2022**

(Member)

Signature:

Name: **Dr. Ayad Rodhan Abbas**

Title: **Assistant Professor**

Date**: / / 2022**

(Member)

Signature:

Name: **Dr. Eman Salih Al-Shamery**

Title: **Professor.**

Date**: / / 2022**

(Supervisor)

Approved by the Dean of the College of information technology, University of Babylon.

Signature:

Name: **Dr. Hussain Ateya Al-Khalidi**

Title: **Professor.**

Date**: / / 2022**

(Dean of the College of Information Technology)

Dedications

To the savior of mankind …

Al-Imam Al-Mehdi (peace be upon him)

Who will fill the earth with justice and equity, after it has been filled with injustice and oppression. The survival of Allah in the earth.

To the biggest and kindest heart my father-in-law (Razzaq), who has gone to Allah's paradise, but remain among us as long as we live (rest in peace).

To My Father …

how taught me to stand up for what I believe in

To My Mother …

The dearest, sweetest angel, the tower of strength to me

To My Husband …

My greatest support, my strongest motivation, my truest smile

To My Faithful Sisters …

Who have always helped me and believed that I could do it

Special thanks to my family-in-law,

For their continuous support and encouragement during the period of my study

Acknowledgement

It is a genuine pleasure to express my deep sense of thanks to the Almighty Allah who helping me to accomplish this work and presenting it in the best way. When I thought of giving up. I would like to express my sincere gratitude and appreciation to my supervisor ***Prof. Dr. Eman S. Al-Shamery*** for her invaluable guidance, supervision, and untiring efforts during the course of this work. I have lucky enough to be advised and guided by her.

I must express my very profound gratitude to my family, family in law, my husband, my friends and all the kind, helpful, and lovely people who helped me directly or indirectly to complete this work and apologize to them for not being able to mention them by name here, but they are in my heart.

Finally, I thank my colleagues from the Science Course and the staff of the department of Software for the help they have introduced to me.

Dhuha Kussein Al-Jubory

Abstract

Infectious disease spread modeling and forecasting are interesting topics in healthcare systems that have attracted the attention of researchers during the last few decades. In particular, the epidemic of the era COVID-19.

The proposed methodology comprises three major models: The Forecasting Model, the Knowledge Discovery Model, and the Visualization Model. The Forecasting Model has been achieved by the Random Forest Regressor employed to forecast the complex and irregular behavior of COVID-19 new infections as well as human mobility in terms of (driving and walking), the Knowledge Discovery Model is used to find the relationship between human mobility and COVID-19 spread during various periods of time using Pearson's Correlation. Through the Visualization Model, the choropleth map is established based on the geographical locations of the study countries. The choropleth map creation introduces a useful view through the GIS map. The system is applied to two datasets: the World Health Organization (WHO) dataset for daily COVID-19 infections for seven countries (United Arab Emirates, China, South Africa, United Kingdom, Germany, Denmark, and Brazil), and the Apple Company dataset for daily human mobility for six countries (United Arab Emirates, South Africa, United Kingdom, Germany, Denmark, and Brazil).

The Forecasting model is applied to two types of Covid-19 datasets (the original data and the weighted data), The evaluation of the forecasting model has been performed depending on two measures of prediction Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Furthermore, the forecasting technique has been compared with other common techniques: Linear Regression (LR), Auto-Regressive Integrated Moving Average model (ARIMA), and Long-Short Term Memory (LSTM).

The Knowledge Discovery Model based on inference rules introduced a reasonable interpretation to the relationship between the disease spread and human mobility through three periods of time. The Visualization Model has been employed through two types of choropleth maps: the disease spread and human mobility map; and predictions of COVID-19 and human mobility map.

The experimental results show that the performance of RF technique outperforms its comparable ones. In particular, the best RMSE and MAE values for Covid-19 daily infections of the proposed model have reached (0.004 and 0.002) for the weighted new cases of South Africa, and (0.009 and 0.007) for human mobility (driving), and (0.011 and 0.007) for human mobility (walking) of United Arab Emirates. On the other hand, the results of Knowledge Discovery Model indicate a clear and strong relationship between disease spread and mobility, thus reaching two types of patterns in accordance to the periods of the study in addition to introducing a clear explanation for the presence of other factors involved during the periods of the study that affected the results. The Visualization Model plays a role in the evaluation process of the whole work through the comparison of the predictions to the actual values, as well as introducing another view of the relationship patterns between the human mobility and disease spread.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

Abbreviation	***Meaning***
ARE	United Arab Emarat
ARIMA	Auto Regressive Integrated Moving Average
Bagging	Bootstrap Aggregating
BRA	Brazil
DEU	United Kingdom
DM	Data Mining
DNK	Denmark
DSC	Dataset of Covid-19
DSM	Dataset of Mobility
GBR	Germany
GIS	Geographic Information System
ICU	Intensive Care Units beds
KDD	Knowledge Discovery in Database
LR	Linear Regression
LSTM	Long short-term memory
MAE	Mean Absolute Error
MSE	Mean Square Error
R^2	R-Squared
RMSE	Root Mean Square Error
WHO	World Health Organization
ZAF	South Africa

Chapter One
General Introduction

1.1. Introduction

With Global Positioning System trajectories (GPS), many people have started recording and sharing their movements within web-based communities. It has enabled individuals to share their route tracing, travel experiences, sports activity analysis, lifelogging, etc. These trajectories can produce enormous business opportunities in geographical navigation and recommendation systems, and in fact, trajectories of human mobility also play a significant role in numerous research fields such as epidemic modeling, traffic planning, mobile computing, and disaster response [1].

A sound understanding of mobility through different economies might be helpful in the creation of public dynamics to enhance human development in fields such as modeling the epidemic spread, social sciences, and transportation alternatives. In recent years, a wide range of studies using different applications have resulted in various models, like modeling the spread of viruses [2].

The authors of [3] analyzed the relationship between human mobility and urban socio-economic status in which they presented indices of mobility that are best associated with socio-economic levels, and they created a model for predicting the socio-economic level from cell phone traces.

Although exploring human mobility is based on difficult scientific methods, the collection of data involved tedious manual travel surveys, interviews in space-time, and diaries. The development of sensors that can detect location has significantly modified the opportunities for accurate statistics on the movements of human beings. This has led to the promotion of various methodological advances that can recognize human movement patterns and their effects on the spread, persistence, and evolution

of infectious diseases [4]. Human culture has ever been in combat with infectious diseases. Diseases such as the Ebola virus, AIDS, influenza, pestilence, plague, SARS, avian flu, and cholera spread had killed more than tens of thousands of people in many regions around the world [5].

A majority of studies have acknowledged the impact of mobility on epidemics' spread [6][7]. Specifically, COVID-19 that first registered at the end of 2019 in Wuhan, China. The World Health Organization (WHO) announced it as a “Public Health Emergency of International Concern" on January 30, 2020, and declared it a pandemic on March 11, 2020. The disease has been reported in more than 200 countries and continues to spread around the world. According to WHO, up to March 23, 14,509 people had succumbed to death, with 1,727 new deaths, while 332,930 confirmed cases of Coronavirus contamination, with 40,788 affirmed cases were reported. This pandemic has demonstrated the interconnected nature of the world and that no one is safe until everyone is safe. As the number of cases increases, policymakers and authorities have begun to use mobility restrictions that were, and still are, the best way to monitor viral transmission.

Due to the rapid and dangerous spread of the pandemic in the USA, the absence of a centralized policy, and only ‘Stay-at-home’ State Authorities' Instructions, the author of [8] investigated how the prevalence of new infections in 25 counties in the USA was influenced by social distance. Badr found that the decreasing COVID-19 case growth rate patterns were strongly correlated with human mobility for the USA's most affected counties. Searchers of [9] uncovered the behavioral parameters of change in mobility patterns of major cities in Canada and the people’s contentment about how infectious the disease is at the level of compliance with public orders of stay at home, using the concept of an individual’s activity space (encompassing all the locations that an individual interacts with over

time). The research discussed how mobility patterns show the effectiveness of lockdown to control disease transmission. Even under strict restrictions, the degree of social distancing was inferred as being bound by choice, which is related to people's beliefs about the authorities' public order regarding how grim the crisis is.

Thus, enforcing restrictions on gatherings and canceling public events, which limit human movement in numerous and crowded locations, has the greatest impact on pandemic containment, in terms of both statistics and levels of effect. Additionally, school and workplace closures, as well as stay-at-home requirements, have contributed to a decrease in the occurrence of infections. Closures of public transport and controls of international travel had also contributed to a decline in new infections [10].

Apple and Google as global communities have responded to COVID-19 by cooperating with public health officials through the use of products such as Google and Apple Maps that could facilitate critical decision-making to combat the disease.

Due to the randomness and irregular nature of the virus, there is greater instability around the choice of ideal time of disappearance of this illness. In addition, the increasing number of confirmed cases has led to an increase in the demand for hospital beds, and in the most severe cases, the demand for Intensive Care Unit beds[11]. Therefore, for the better management of the economical, societal, cultural, and public health issues it's important to get an estimate about what would be the daily new cases at least short term forecasting for the upcoming few weeks ultimately also help the healthcare systems and governments to prepare in advance for the expected number of new cases [12].

Time series forecasting is a relevant task in various fields of science, like finance, economics, health, engineering, meteorology, and sciences. [13]

Recently, several studies that employ machine learning and deep learning techniques that are capable of forecasting the trends of diseases based on the reported time-series data to estimate the spread of the viruses. In [14] Auto-Regressive Integrated Moving Average model was used to forecast the daily new cases of Covid-19 from April 27, 2020, to May 3, 2020, using South Korean and Japanese data. The dependent structure of the daily time series of newly confirmed cases was well captured by the estimated model of the Auto-Regressive Integrated Moving Average. Other models were introduced to handle forecasting by using Regression .The authors in [15] used Linear Regression in their analysis study which aims to track trends related to the expected number of deaths of the coronavirus in India. As well as Logistic Regression model was used in [16] to estimate the spread size of Covid-19 in Kuwait, by forecasting the daily cases, total infected cases, and the expected dates of the start and ending phase of the pandemic, besides that, simulations of (susceptible-infectious-recovered) stochastic individual contact model were performed to examine the impact of the simultaneous changes on the number of the susceptible and infected people.

This study introduces a knowledge Discovery Model to investigate how effective restrictions of mobility trends could control the transmission of SARS-CoV-2 by finding the correlation degree of human mobility into the pandemic outbreak and providing a reasonable interpretation of relationship patterns also using the chronology of time series and cumulative errors to improve the forecasting accuracy of the Random Forest model in predicting COVID-19 new infections as well as human mobility, and produces a comparative study of the performance of three models with the adopted model , in addition to visualizing the results of those models.

1.2. Problem Statement

The whole world has suffered from the consequences of the era pandemic COVID-19. The disease is spreading rapidly with the inability of authorities and medical staff to contain or control it, thus finding non-medical interventions has become a global priority for many scientific researchers that might help in curtailing the disease's spread and rescuing the globe.

One of the important aspects that play a major role in many studies is discovering the relationship between human movements and disease transition, thus helping policy makers to consider that point when they make decisions about locking down or limiting people's mobility and social isolation. The forecasting field also took a place in dealing with the pandemic to help governments to make preparations and gain an indicator of how to handle the upcoming waves or infections by predicting future cases.

1.3. Aims of Study

1- Discovering the correlation degree between human mobility and the number of people infected with the epidemic in countries and examining how lockdown policy measured by the relative change in mobility can affect the rate of infections.

2- Improving the accuracy of forecasting as much as possible for COVID-19 infections and mobility rates.

3- Constructing Choropleth Maps:

- Disease outbreak map with daily human movements of driving and walking.
- Forecasting map.

1.4. Challenges of the Study

1- Mobility data for Apple company covers most countries around the world since some countries or cities have no accurate mobility information because the internet services are not active throughout the year, thus the researcher has limited information concerning mobility.

2- Difficulties to search for regular periods where mobility has decreased or increased regularly to study the influence of human movements on an epidemic's spread whereas each country varies from the other in the dates of declaring closure or urging people's compliance with public orders of safety and social distancing, as well as take into consideration the most important factor, which is the commitments of people to those orders.

3- The spread of the epidemic is not only dependent on mobility; other factors, such as the disease's incubation time or people's compliance with local authorities' public commands of mobility restrictions, can also play a role.

4- The number of newly reported cases may not reflect the true number of infections, because many infected people don't do the Polymerase Chain Reaction (PCR) test.

5- Covid-19 forecasting is difficult due to the virus's novelty and the indiscriminate behavior of the virus spread and many countries experienced multiple fluctuation waves that vary in complexity and pattern with faster acceleration in new cases, this making accurate prediction is a challenging process.

1.5. The Study Contribution

1- Proposing a Knowledge Discovery Model to find the relationship between mobility restrictions and Covid-19 spread.

2- Improving the prediction results for Random Forest regressor by re-weighting Covid-19 infections as well as introducing mobility predictions that are very close to the truth.

3- Providing cartographic representations that clearly explain the aforesaid findings.

1.6. Related Work:

This section reviews the most recent prediction studies for COVID-19 time series forecasting, where some rely on statistical methods and others on machine learning. But the researcher has noticed that limited studies are using random forest regressor to predict Covid-19 cases. Also, some studies are introduced about exploring the relationship between human movement and the disease transition.

Many researchers work in the field of COVID-19 new cases, recovery, or death forecasting:

1- In [17] compartmental model uses the distribution of the Bayesian non-linear approach to get a posterior distribution for case trajectory, and a Random-Forest algorithm trained on COVID-19 data to estimate the daily new infections and mortality in the U.S. states. The model predicts COVID-19 cases and deaths well based on the mean absolute error criterion for two weeks, where the MAE register 0.32 for new cases and 0.40 for deaths.

2- The authors in [18] examined Random-Forest, Linear model, Decision Tree, Neural Network, and Support Vector Regressor for estimating the confirmed, death, and recovery cases of COVID-19 in different states of India. The dataset was taken from the Ministry of Health and Family Welfare of India from 30 January 2020 to the end of May 2020. All five models were evaluated on an

accuracy measure calculated as the percentage deviation of the predicted values to the actual values. The experimental results show that the Random Forest regressor outperforms the other five models according to the accuracy of predictions.

3- In another study [19], the authors utilized the Machine Learning model in an attempt to forecast the disease trend in Indonesia within 30 days from April 22, 2020, to May 21, 2020, using a dataset obtained from the Kaggle website containing the confirmed, recovered and deaths cases, Facebook's Prophet (FB) and ARIMA models were introduced to compare their performance. They used (MSE), (MAE), and (MFE) as un evaluation metrics. The results show that the Prophet outperforms the ARIMA model.

4- The Long Short-Term Memory (LSTM) network has been used in [20] to predict the stopping time of COVID-19 spread based on a Canadian historical dataset that was taken from Johns Hopkins University and the Canadian Health authority from the first appearance of the disease in the country till March 31, 2020. They introduced two models, the first LSTM model was trained and tested on Canadian data with 34.83 RMSE error for short-term predictions and 92.67% for long-term predictions, the second model is trained on an Italian dataset to predict cases in Canada. the RMSE for short-term predictions was 51.46, within those results the authors expected to see exponential growth of new infections in Canada and they predicted the potential ending of the outbreak will take place through June 2020. This study was the first study in Canada to predict the gravity of Covid-19 and model the transmission of the diseases using deep learning methods.

5- Authors in [21] present a Linear Regression model to predict daily active cases of Coronavirus based on daily positive cases in Odisha and India. The data was collected from WHO daily reports from March 22, 2020, to the end of Jun 2020.

. They first present Correlation research between a dependent variable (active cases) and independent variables (positive) to find the next active cases number on the test set they found that every unit of positive cases increases, the active cases increase by 68% in Odisha and by 57% active cases in the opposite situation and that indicates a powerful forecasting model. After that, they used Linear Multiple Regression with more independent variables (Recovered, deceased) which also served as powerful as linear regression where R-Square tends to be 0.99 and 1.0 for the first and second model respectively.

Many researchers are focusing on human mobility and the current COVID-19 transmission.

1. Authors of [22] in their modeling study had built a modified susceptible–exposed–infectious–recovered compartmental transmission model for COVID-19 to predict the epidemic curve shape of the disease in Shenzen, China by combining the effect of human mobility with the spread of infection. Data from mobile phones were obtained from the providers of mobile phone services in Shenzhen, covering working days (Monday to Friday) from January 10 to March 10, 2019. The daily cases of Coronavirus in the city were obtained from the Shenzhen Local Government website. The study was distributed through ten regions of Shenzhen. The policy of mobility restriction leads to a decrease in the virus transmissibility by 25% to 50%. The model demonstrated the impact of different types of mobility restrictions in mitigating the outbreaks of the disease. According to the authors, the model could assist policymakers to determine the best combinations of human mobility restrictions during the outbreak for assessing the potential positive impact of mobility restriction on public health in the light of the potential negative societal and economic impacts. The results were presented graphically in their study.

2. In [23], the authors developed a multiple linear regression model to identify the effect of mobility behaviors in the COVID-19 spread in Italy. The study aimed to examine the disease transmission in the country and to discuss the influence of citizen movement on the spread of the epidemic. The daily reports of the positive cases between February 21 and May 5, 2020, were acquired from the Italian Ministry of Health along with the data of the Italian national census relative to 2019, while the COVID-19 mobility data were obtained from the Italian Transport Ministry. The findings indicated that the number of new infections in a day was associated with the trips performed three weeks prior. For this case study, a 21-day threshold is accepted as a kind of positivity detection time measure. In other words, quarantine for mobility restriction is commonly set at 14 days. Underestimating the policy of containment and slowdown in implementing restrictive actions can result in more infections and deaths by COVID-19.
3. The authors of [24] organized a study to investigate the impact of lockdown during the COVID-19 outbreak on human mobility using the change of spatial time series over various union territories in the States of India. This study also figured out the difference before and after the lockdown. Data were collected from February 15 to April 30, 2020, from Google COVID-19 Mobility Reports, 2020; the reports included various places like groceries and pharmacies, transit stations, retail and recreation, parks, groceries and pharmacies, residential and workplaces. During their study, they used conditional formatting techniques such as color ramps (red, yellow, and green) where the red color indicates a high percentage of decreased mobility from baseline, while yellow designates a moderate percentage of decreased mobility and green implies a very low percentage of mobility decreased from baseline. They have also employed the interpolation mapping techniques of "spatial inverse distance weighted" to show

pre- and post-lockdown human mobility trends due to COVID-19. The results showed that grocery and pharmacy, workplaces, transit stations, retail, and recreation, as well as visits to parks mobility dropped by -51.2%, -56.7%, -66%, -73.4%, and -46.3% respectively. As people mostly opted to stay at home during the lockdown, the residential places mobility visits raised by 23.8%. Hence, the result of the mobility trends over time can be used in public health strategies to reduce COVID-19 transmission.

4. In [25], the author had collected data on SARS-CoV-2 between February 24 and April 5 from the Italian and Spanish Ministries of Health websites. He used an interrupted time series to evaluate the change in trends of incident diagnosed cases based on the ICU admissions before and after the two countries' respective lockdowns, and analyzed with quasi-Poisson regression using Stata Release 16. He found that the daily percentage of all the incidence outcomes before the lockdown had increased and Spain recorded higher percentages (38.5% for diagnosed cases, 26.5% for ICU admissions, and 59.3% for deaths) compared to Italy (21.6%, 16.7%, and 32.8% respectively). In Italy during the first lockdown, the diagnosed cases were reduced by 42.1%, ICU admissions by 77.8%, and deaths by 58.2%, whereas in Spain, the reduction was even higher; diagnosed cases had dropped by 69.1%, ICU admissions by 66.8% and deaths by 77.8%. During the second lockdown, both countries showed a reduction in daily diagnosed cases, ICU admissions, and deaths.
5. The authors in [26], had also analyzed the association of mobility trends with the number of infected people with Coronavirus for 144 countries. Schengen countries were coded as 1 and 0 for otherwise using Negative Binomial Regression Analysis based on the Poisson-gamma mixture distribution. They focused on airports number, the Schengen system, and the volume of air travel. Data on Coronavirus infection were extracted from WHO and mobility data were

obtained from the World Development Indicators. The results showed a positive correlation between the increasing number of infections was positively correlated with the magnitude of airline travel. They also found that Schengen countries which have a higher percentage of elderly people and a higher population density were reported with higher COVID-19 cases compared to other countries. In addition to that, countries with a greater number of airports were linked to a higher number of infections due to the epidemic.

Table 1.1 Summary of the related works concerning forecasting of COVID-19 cases

No	Method Name	Forecasting Type	Data set	Country	Period	Evaluation measures	results
1	Bayesian nonlinear method and Random Forest regressor	Regression	los Alamos National Laboratory USA	U.S	Three months period	MAE	0.32 for new cases 0.40 for death
2	Random Forest, linear model, decision tree neural network, support vector regressor	Regression	Ministry of Health and Family Welfare of India	India	Four months	the percentage deviation of the predicted values to the actual values	Random Forest Outperforms Other models to predict confirm, recovered, and death
3	Facebook's Prophet (FB) and ARIMA	regression	Kaggle website	Indonesia	30-day time period to	(MSE), (MAE)	Prophet generally outperfor

					predict two weeks	and (MFE)	ms ARIMA
4	Long short-term memory (LSTM) network	Regression	Johns Hopkins University and Canadian Health authority	Canada	from the first appearance of the disease in the country till March 31, 2020.	RMSE	the exponential growth of new infections
5	Linear Regression model and linear multiple regression	Regression	WHO daily reports	India	Three and a half months	R-Square	0.99 and 1.0 for the first and second model respectively to predict the positive cases

Table 1.2 Summary of related works concerning the effect of mobility on Covid-19 transmission

No.	**method**	**Country**	**Mobility data**	**Covid-19 data**	**result**
1	Modified susceptible–exposed–infectious–recovered compartmental transmission model	Shenzen, China	providers of mobile phone services in Shenzhen, from January 10 to March 10, 2019	Shenzhen Local Government website.	Mobility restriction decreases virus transmissibility by 25% to 50%.

2	multiple linear regression	Italy	mobility data were obtained from the Italian Transport Ministry	Covid-19 new cases between February 21 and May 5, 2020, from the Italian Ministry of Health along with the data of the Italian national census relative to 2019,	the number of new infections in a day was associated with the trips performed three weeks prior
3	the change of spatial time series over various union territories (UTs) and the interpolation mapping techniques of "spatial inverse distance weighted" (IDW)	India	Google COVID-19 Mobility Reports, 2020 from February 15 to April 30, 2020		Grocery, pharmacy, workplaces, transit stations, retail recreation, and visits to parks, mobility dropped by -51.2%, -56.7%, -66%, -73.4%, -46.3% respectively. And that reduces COVID-19 transmission during the lockdown
4	Poisson regression	Italy and Spain		Italian and Spanish Ministries of Health websites between February 24 and April 5.	both countries showed a reduction in daily diagnosed cases, ICU admissions, and deaths during the lockdown.

5	negative binomial regression (NBR) analysis based on the Poisson-gamma mixture distribution	Schengen countries	World Development Indicators.	WHO	Schengen countries were reported with higher COVID-19 cases compared to other countries. In addition to that, countries with a greater number of airports were linked to a higher number of infections.

1.7. Outlines of the Study

After the general introduction presented in chapter one, the rest of the chapters of this thesis are organized as follows:

4- Chapter 2: presents a description of the theories and methods of Data Mining (DM), Artificial Intelligent (AI) Forecasting techniques, Knowledge Discovery, and choropleth Maps Visualization that were utilized in this study.
5- Chapter 3: clarifies the basic steps of designing the study methodology concerning building a Knowledge Discovery Model to find out the correlation degree of human mobility into the epidemic spread and Forecasting COVID-19 new infections and mobility as well as the Visualization procedure.
6- Chapter 4: discuss the experimental results conducted through implementing the methodology on Covid-19 new infections and human mobility in terms of driving and walking.
7- Chapter 5: list the conclusion reached through this study and give suggestions for future work.

Chapter Two
Research Background

2.1. Overview

This chapter presents the concepts and methodologies that were followed in this study. It shows the theoretical background of DM and AI principles, these principles are described by providing a brief overview of data preprocessing, forecasting of Covid-19 new infections as well as mobility prediction, the correlation between mobility and dieses spread and data visualization.

At the beginning, an overview of the relationship between infectious diseases spread and human movements are introduced.

2.2. Apple's Mobility Data of COVID-19

On 13 January 2020 Apple released from Apple Maps a mobility data tool. [1]

The data covered a large number of countries and cities around the world. The reports reflected the requests for directions in Apple Maps for daily mobility trends of walking, driving, and transit. The data are indexed to (100) for each city on January 13^{th,} 2020, Apple's data 'Days' are defined as Midnight to Midnight. There is an existing privacy setting; Apple does not have a search history for individual movements. Figure (2.1) below shows an example of Apple data-.

2.3. World Health Organization COVID-19 dataset

The World Health Organization (WHO) oversees and maintains a wide range of data sets concerning global health and well-being. Covid-19 data is freely available on the WHO international database that provides daily data on newly, confirmed, deaths, and cumulative cases for all countries around the world. The WHO Covid-19 dataset was explored on January 3, 2020. Figure (2.2) shows an example of the COVID-19 dataset. [2] China serves as a good example of how they

[1] https://www.apple.com/covid19/mobility

[2] https://covid19.who.int

control the diseases transmission through the lockdown. Thus, the disease spreads decreased as a result of mobility restrictions and reached only one or two new cases daily, as shown in Figure (2.2) [53].

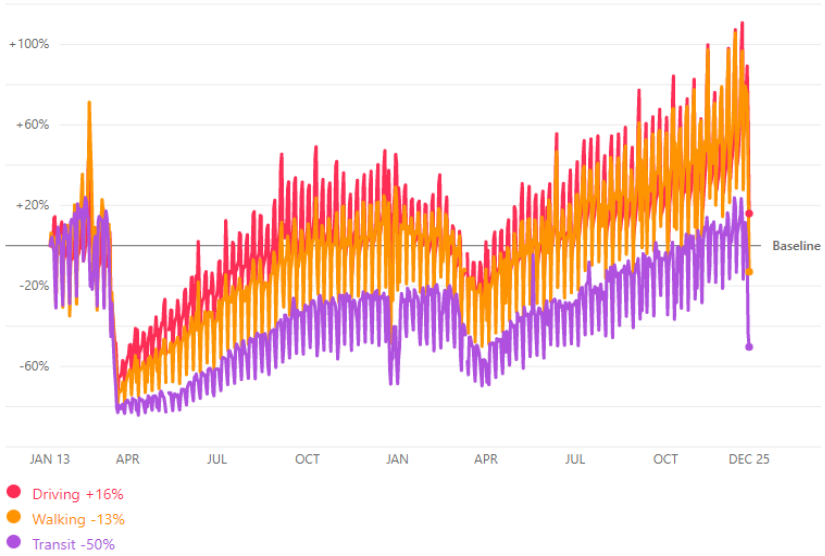


Figure (2.1) Sample of Apple mobility trends for Brazil

Table (2.1) Sample of WHO COVID_19 Dataset for (Brazil)

Date - reported	Country - code	Country	WHO - region	New - cases	Cumulative - cases	New - deaths	Cumulative - deaths
3/11/2020	BR	Brazil	AMRO	9	38	0	0
3/12/2020	BR	Brazil	AMRO	18	56	0	0
3/13/2020	BR	Brazil	AMRO	25	81	0	0
3/14/2020	BR	Brazil	AMRO	44	125	0	0
3/15/2020	BR	Brazil	AMRO	0	125	0	0
3/16/2020	BR	Brazil	AMRO	79	204	0	0
3/17/2020	BR	Brazil	AMRO	34	238	0	0
3/18/2020	BR	Brazil	AMRO	57	295	1	1
3/19/2020	BR	Brazil	AMRO	133	428	3	4
3/20/2020	BR	Brazil	AMRO	193	621	0	4
3/21/2020	BR	Brazil	AMRO	283	904	7	11
3/22/2020	BR	Brazil	AMRO	0	904	0	11
3/23/2020	BR	Brazil	AMRO	0	904	0	11
3/24/2020	BR	Brazil	AMRO	642	1546	14	25
3/25/2020	BR	Brazil	AMRO	655	2201	21	46

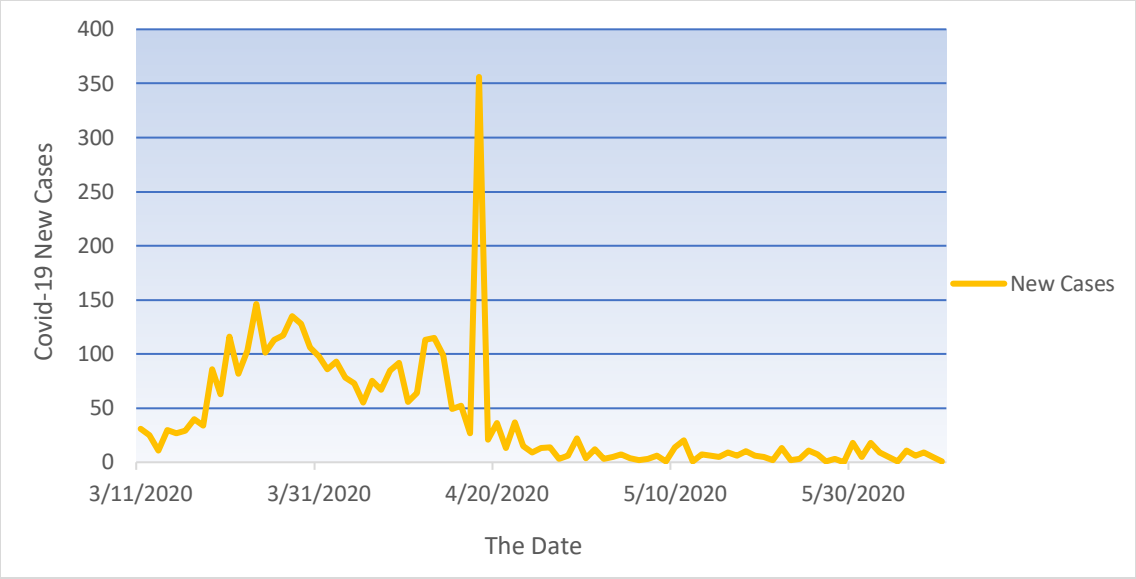


Figure (2.2) The daily trend line of the disease spread in China

2.4. Data Mining

This section introduces the most relevant terms in the DM terminology that were utilized in the research:

2.4.1. Data Preprocessing

To make data more suitable for DM and Machine Learning methods. Data preprocessing techniques are required [27]. Data preprocessing consists of data preparation, which includes data integration, normalization, cleaning, and transformation, as well as data reduction activities that include discretization, feature selection, and instance selection. The result of data preparation operations is a final dataset that may be considered acceptable for future algorithms of DM [28]. Figure (2.3) shows some forms of data preprocessing. This work needs preprocessing techniques as described below:

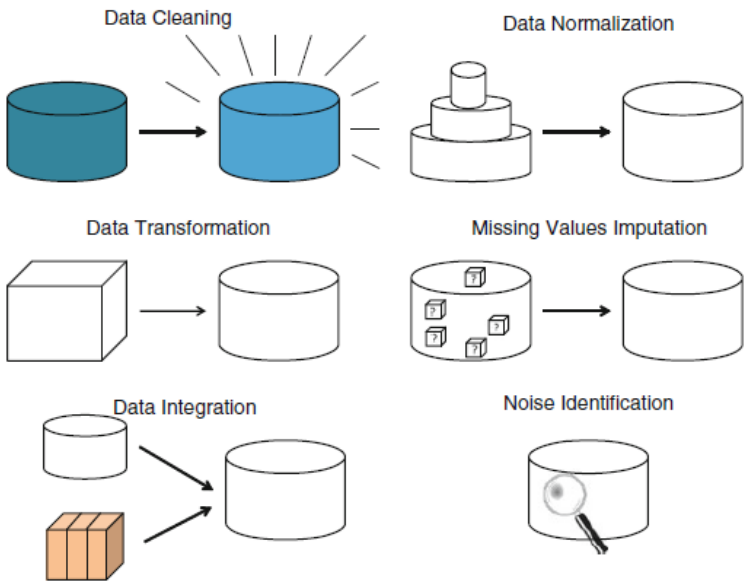


Figure (2.3) Forms of data pre-processing.

2.4.1.1. Handling Missing Values

Missing values must be handled correctly to perform effective modeling. There are several methods introduced to handle missing values, such as eliminating the data, ignoring it during analysis, or estimating it. For numerical values, the most important type is an estimation, which can be done in two popular methods. The first one is to replace the missing one with the average of all values of the corresponding attribute. The second method is the Mode method (replace the missing value with the more frequent value of an attribute) [28].

2.4.1.2. Data Normalization

In most cases, data is obtained from different sources and then kept in a data warehouse. If these data are merged for a modeling method, this might cause a lot of problems. In addition, inconsistency is a problem here. This problem can affect model efficiency. So, the goal of normalization is to ensure that all attributes have equal weights. As a result, the smaller numbers have a higher influence on the final

result. Normalization techniques such as minimum-maximum and z-score normalization are well-known. The min-max normalization is calculated using Equation 2.1 [29].

$$\hat{v} = \frac{v - min}{max - min}\ (new\_max - new\_min) + new\_min \quad (2.1)$$

Where:

$\hat{v}$ is the new value of any feature, v is the old value of any feature, Max is the maximum attribute value, Min is the minimum attribute value.

new_min and new_max are the minima and maximum interval of the values respectively.

2.4.1.3. Data Aggregation

Row data can't be processed directly. Sometimes it's important to extract a summary of the features from the original dataset and these features shouldn't be redundant. This makes further processing easier and avoids the curse of dimensionality. many types of data aggregation methods such as count, mean, median, maximum, minimum, range, mode, and sum. This study uses the sum of data aggregation [29].

2.4.1.4. Data Integration

The data integration process consists of combining data from multiple data sources. This operation must be done carefully to avoid inconsistencies and redundancies in the final data set. The most common data integration procedures are the unification and identification of domains and variables, attribute correlation analysis, tuples duplication, and the detection of conflicts in data values from several sources [28].

2.4.2. Knowledge Discovery of Research Problem

Knowledge Discovery in Database (KDD) is the process of extracting ideal and useful hidden patterns or models (knowledge) from large datasets [30]. The KDD process consists of the following five major steps (see also Figure (2.4):

1- selection, creation of a target dataset from the original one.
2- preprocessing, improving the data accuracy by various operations, such as removal of noise or outlier and handling missing data.
3- transformation, Preparation of the data to be suitable for DM algorithm by reducing dimensions with techniques such as (feature selection or extraction) and attributes transformation like (numerical attributes discretization, aggregation, and functional transformation)
4- DM, extraction of useful patterns by applying a specific DM task such as (classification, regression, clustering, etc.) with a suitable algorithm to perform that task, and a proper output representation.
5- evaluation, this step carried out the interpretation of the patterns and extracts knowledge by patterns or model visualization [31].

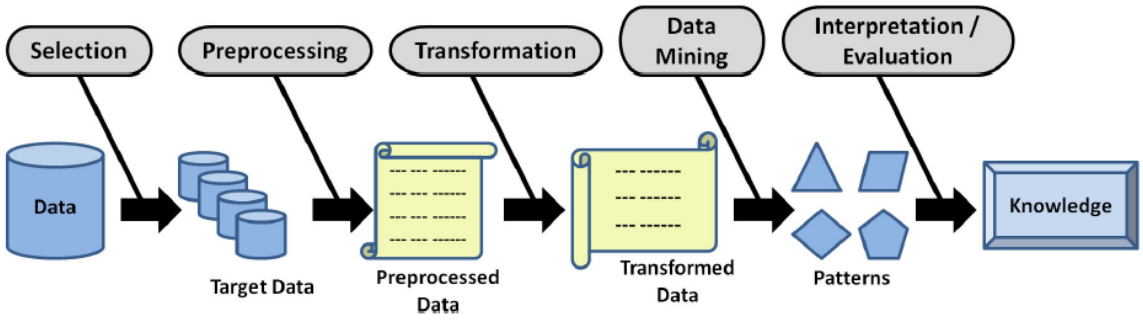


Figure (2.4) Knowledge discovery in database steps

2.5. Pearson's Correlation Coefficient

The Pearson's correlation coefficient formula is used to measure the association level of two data points. It defines the strength of the linear correlation between the two variables. For a size M sample with x and y variables, Pearson's formula is defined as the covariance of x and y divided by their standard deviation product. Correlation coefficients can be high or low (magnitude), and positive or negative (direction). Correlation coefficients vary from -1 to +1: whereas -1 and +1 indicate perfect negative and perfect positive correlation coefficients respectively, a correlation coefficient of 0 implies no correlation (zero relationship). Further, correlation coefficients lower that ± 0.4 (whether negative or positive ± 0.4) are said to be low, between ± 0.4 and ± 0.6 are moderate, and above ± 0.6 are high.[32]

The formula of Pearson's correlation coefficient is defined in Equation (2.2)[33].

$$R_{xy} = \frac{\sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{M}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{M}(y_i - \bar{y})^2}} \qquad (2.2)$$

Where $\bar{x}$ and $\bar{y}$ are the means of sample, R is the Pearson's correlation coefficient between x and y, and $R \in [-1, 1]$, where -1 means there is a strong negative correlation, 0 means there is no association between x and y, and +1 reflect perfect strong correlation.

2.6. Data Mining Categories

The taxonomy of DM methods falls into two categories: predictive (classification, regression, and time series) and descriptive (clustering, association, summarization, and visualization) as shown in Figure (2.5). The main popular techniques are classification, regression, and clustering. The use of these techniques

depends on the targeted attribute Y. If the value of Y is continuous, the regression technique may be employed; if the target value of Y is discrete, the classification technique can be used [34].

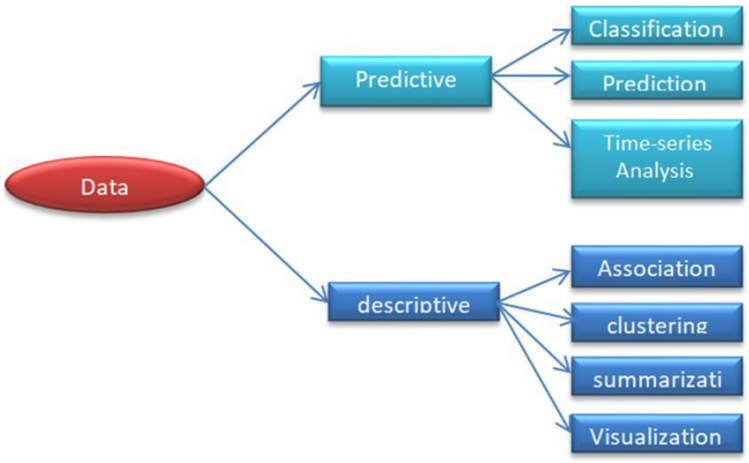


Figure (2.5) Data Mining Taxonomy

2.7. The Prediction Techniques

Machine Learning (ML) and Deep Learning (DL) are two of the most important fields of learning in AI. ML is the use and development of algorithms that can learn and adapt automatically via experience, and by the use of data, algorithms, and statistical models to analyze and draw inferences from patterns in data.[35]

Supervised learning algorithms, in particular, build a model from sample data, referred to as training data, to make predictions or decisions without being explicitly programmed to do so. Deep learning is a subset of a larger family of learning methods based on artificial neural networks that employ numerous layers of processing to extract progressively higher-level features from data. Deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks are examples of deep-learning architectures.

In recent years, AI technology has sparked a lot of interest in a variety of disciplines, including medicine, where it may help physicians and authorities with image inspection, surgery, medical data integration, hospital management, and disease-assisted diagnosis, to name a few [35]. Table (2.1) shows A summary of prediction approaches.

Table (2.2) A summary of prediction approaches

	Method	Abbreviation	Description
ML	Statistic	**LR**	**Linear Regression**
		MLR	Multiple Linear Regression
		PR	Polynomial Regression
		LOR	Logistic Regression
		LASSO	Least Absolute Shrinkage and Selection Operator
	Time Series	WMA	Weighted Moving Average
		ES	Exponential
		AR	Autoregressive process
		MA	Moving Average
		ARMA	Auto-Regressive Moving Average
		ARIMA	**Auto-Regressive Integrated Moving Average**
		Prophet	Modular regression model developed a Facebook
	Classification	SVM	Support Vector Machine
		LS-SVM	Least Square Support Vector Machine
		MLP	Multilayer Perceptron
		SVR	Support vector regression
		NB	Naive Bayes
		EL	Ensemble Learning
		XGB	Extreme Gradient Boosting
		HMM	Hidden Markov
		BL	Instance-Based Learning
		KNN	K-Nearest Neighbor
		DT	Decision trees
		CR	Classification via Regression
		RF	**Random Forest**
		Extra Trees	Extremely Randomized Trees

	Artificial Neural Networks	RVFL	Random Vector Functional Link Network
		RNNs	Recurrent Neural Networks
		DNNs	Deep Neural Networks
		LSTMs	**Long Short-Term Memory Networks**
		BiLSTMs	Bidirectional Long short-term Memory Networks
		SLSTMs	Stacked Long Short-Term Memory Networks
		ConvLSTM	Convolutional LSTM
		GRU	Gate Recurrent Unit
		CNN	Convolutional Neural Network
		GAN	Generative Adversarial Network
		VAE	Variational Autoencoder

2.7.1. Regression

Regression analysis is a supervised learning technique based on statistical concepts which allow estimating the relationships between a dependent variable and one or more independent variables and modeling the future relationship between them.[35]

2.7.1.1. Linear Regression

Linear regression models try to attempt the relation among two different parameters by fitting the data in a straight line that is observed between the dependent parameter and independent parameter. The idea at the base of regression analysis for forecasting a time series *Y* is that there is a linear relationship with other time series *X*. *Y* is called forecast or dependent variable, while *X* the regressors, predictors, or independent variables. In the simplest case, the forecast variable has a linear relationship with a single variable Here both parameters have numeric values. In simple regression analysis, the linear regression is expressed

As in equation (2.3). [36]

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon \qquad (2.3)$$

The equation represents a straight line where, $\beta_0$is the *Y* intercept, β_1 is the slope called the regression coefficient (weight of regression), and ε is the error term (the regression residual.) The goal of the prediction is thus to find the values of the coefficients β_i to obtain the best-fit regression line.

2.7.2. Time Series Forecasting

A time series is a sequence of observations listed for a certain period of time in order. The forecasting of time series is an important task in the machine learning field and involves estimating future observations based on historical data. Each record in the dataset that feeds the forecasting model represents a tuple *(X, Y)*, where *X* is the attributes and *Y* is the target of process prediction [34]. analyzing the time series data is important in a wide range of areas like business, agriculture, meteorology, and ecology, to forecast, for example, daily stock prices for closing, hourly speeds of wind, erosion of soil, and the species of an animal, respectively [37]. four features associated with the time series (horizontal, cyclical, seasonal, trend pattern), depending on the information design. The horizontal pattern is found when the change of the data value comes with a constant mean value; a cyclical pattern exists when the data is subject to economic change such as the business period, whereas a seasonal design appears if the series data have a repeated pattern over sequential time periods, the trend pattern appears when the value in the time series data increasing or decreasing in long-term time cycle.[38]

The task of forecasting usually entails two main stages. The first stage (learning model stage) is the model building where the employed methodology trained on the training data, whereas the second stage (Forecasting stage) is to produce forecasting of the target attribute. The model represents the forecasting system as it receives input data and analyzes it to produce a predicted value as an

output. The prediction models are based on a hypothesis about the link between historical data (inputs) and future values (outputs).[38]

2.7.2.1 ARIMA Model

Auto-Regressive Integrated Moving Average (ARIMA) models have widely used approaches to time series forecasting, especially for detecting outbreaks of infectious diseases[39].

An ARIMA model is characterized by 3 terms: p, d, q. Where,'p' is the order of the'Auto-Regressive' (AR) term; it refers to the number of Y lags which should be used as predictors. The 'q' is the order of the 'Moving Average' (MA) term; it refers to the number of lagged errors in the forecast that should go into the ARIMA model, d is the number of differencing required to make the time series stationary. the input time series for an ARIMA model needs to be stationary, i.e., the time series should have a constant mean, variance, and autocorrelation through time. Therefore, the stationarity of the data series needs to be identified first. The most common approach to making a series stationary is to subtract the previous value from the current value. Sometimes more than one differentiation may be required, depending on the complexity of the series. Therefore, the value of d is the minimum amount of differentiation needed to render the sequence stationary, and if the time series is stationary already, then $d = 0$. The principal objective of the fitting ARIMA model is to correctly recognize the stochastic mechanism of the time series and forecast future values. Such approaches have also proven useful in other types of scenarios in which models for discrete-time series and dynamic systems are created [40]. A seasonal ARIMA can be built for the data series as in Equation (2.4) [41]

$$\varphi_q(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)u_t \qquad (2.4)$$

Where P is the number of seasonal autoregressive parameters, Q is the seasonal moving average order, s is the period length, and D denotes the number of differencing passes.

$\varphi_q(B)$ is the regular Auto-Regressive polynomial of order *p*.

$\Phi_p(B^s)$ is the seasonal Auto-Regressive polynomial of order *P*.

$\theta_q(B)$ is the regular Auto-Regressive polynomial of order *p*

$\Theta_Q(B^s)$ is the seasonal Moving-Average polynomial of order *Q*.

u_t follows a white noise

B is the operator of backshift, which shift the *yt* observation (a time series) by one point in time.

Figure(2.6) shows the procedure of applying the ARIMA model.

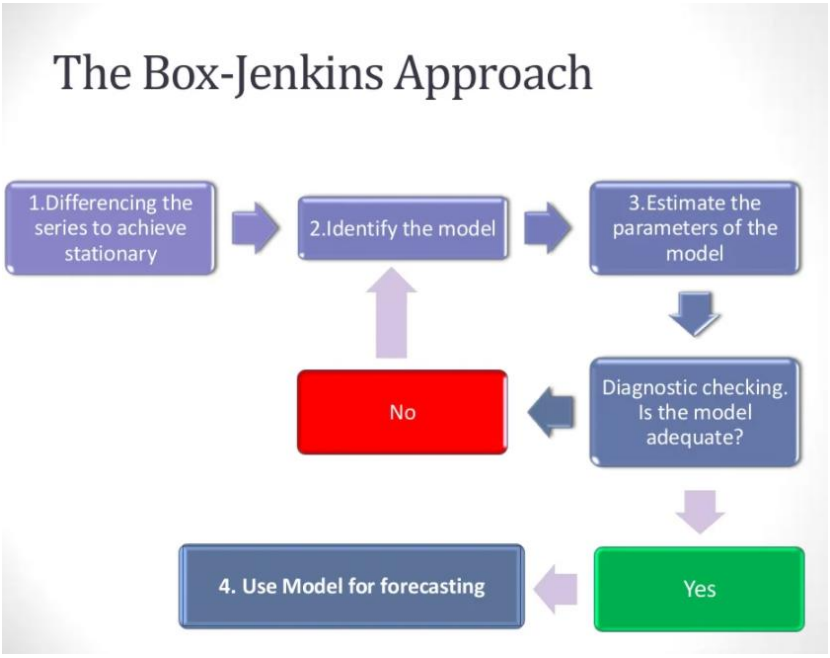


Figure (2.6) The procedure of applying ARIMA model

A partial autocorrelation function (PACF) can be utilized to find the AR parameter value and the correlogram graphs of the autocorrelation (ACF) functions can be used to achieve the value of the MA. In order to obtain the most appropriate parameter in the ARIMA approach, the model performance is usually measured by the Akaike Information Criteria (AIC) expression, where the model with the lowest AIC criterion is considered more successful than the others. It can be calculated as in Equation (2.5)[42] :

$$\mathrm{AIC} = -2\ln \mathrm{L} + 2\mathrm{K} \qquad (2.5)$$

Here: L represent the function of likelihood.

K is the number of free parameters (k=p+q).

n is the residuals number that can be calculated for the time series.

2.7.3. Ensemble Methods

Ensemble methods are techniques that combine multiple models to produce an improved and accurate decision. The set of models (ensembles) also called base models are more powerful than a single model. Each model is obtained by applying a machine-learning algorithm to training data such as decision trees, logistic regression, etc. to produce a prediction. The final prediction is constructed by taking the vote of each classifier (ensemble) for Classification problems or taking the average of the ensembles for Regression problems. Figure (2.7) reflects the ensemble technique idea.

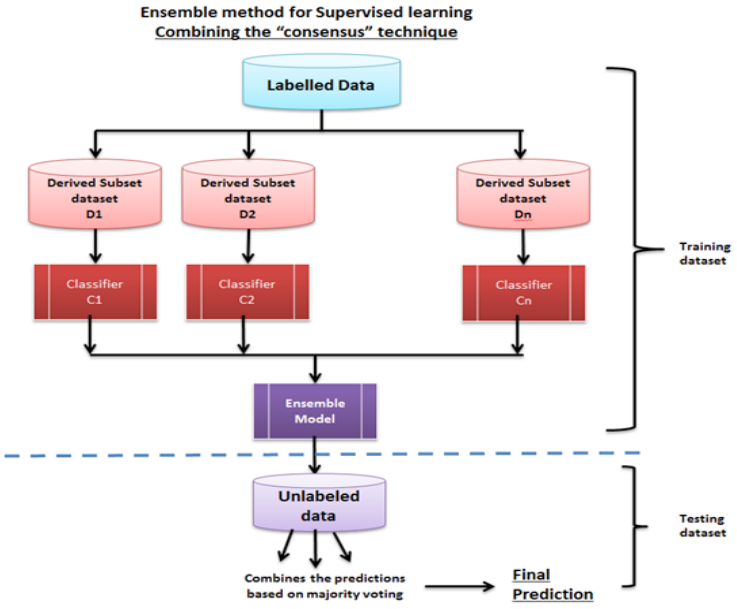


Figure (2.7) Ensemble technique

2.7.3.1. Bootstrap Aggregating (Bagging)

Bagging is a method for creating classifiers that each one learns from the other in parallel, independently of each other, by sampling the dataset periodically with replacement using a uniform probability distribution. The size of each "bootstrap" sample is the same size as the original dataset. Because the sampling is done with replacement, some examples (instances) may recur several times in the same training set, while others may be removed from it. The final predictions are constructed based on the combination of all predictions. This strategy reduces variance and aids in preventing overfitting when employed in classification and regression problems. Bagging is commonly used in decision tree approaches [43]. The bagging process is shown in Figure (2.8).

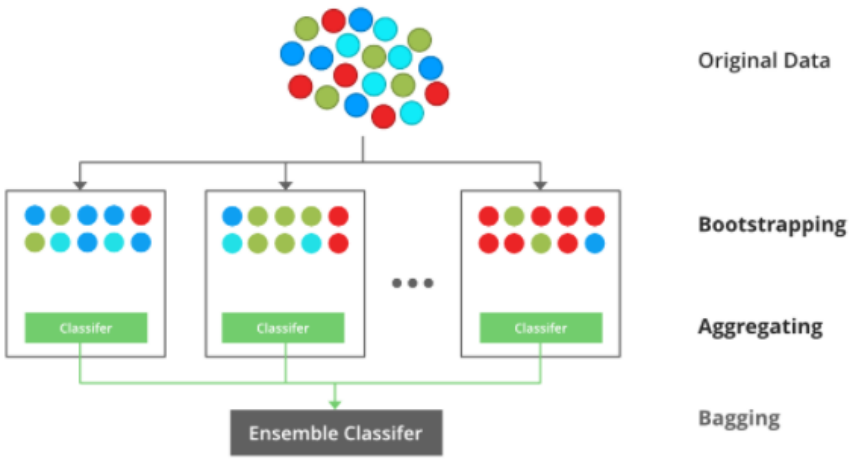


Figure (2.8) Bagging process

2.7.3.2. Random Forest

A Random Forest is an ensemble predictor that combines multiple decision trees that are used for Classification and Regression, for Classification, the predictions are combined using the scheme of a majority vote, whereas, for regression, the average prediction of all trees is returned as a result (see Figure (2.9)). For every tree in the forest, the samples are picked randomly using bootstrap sampling and trained independently. Then all the decision trees in the forest predict each sample in the testing set. With a dataset that has M samples, (Equation (2.6) shows the probability (P) of an example not being chosen :[38]

$$P = 1 - \frac{1}{M} \qquad (2.6)$$

Equation (2.7) shows the probability of a sample to be in the testing set

$$P = \left(1 - \frac{1}{M}\right)^{M} exp^{-1} = 0.3673 \qquad (2.7)$$

Each tree has a test set that consists of 63.27% of the data and differs from other trees. The test set samples are called “out-of-bag” data. every tree has randomly selected features. The tree selects a subset of features and chooses the best one to split the node from this set. This process is repeated recursively until a specific error

rate is reached. To achieve the specified error rate, each tree is grown individually [38].

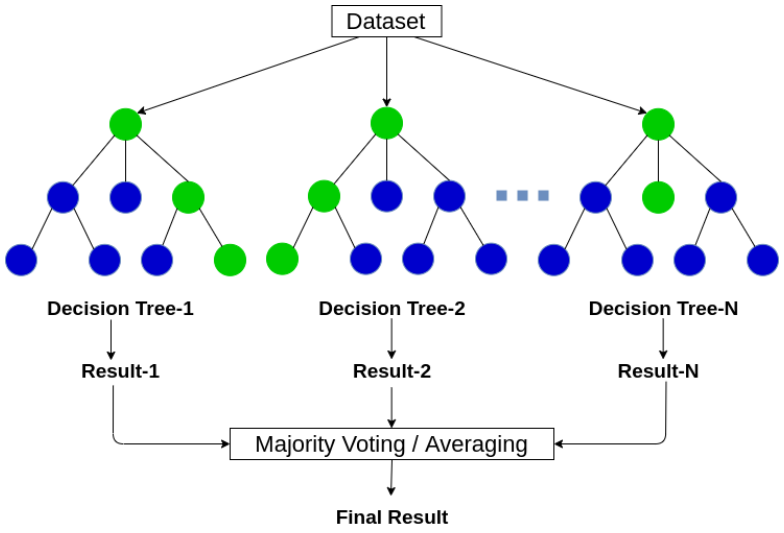


Figure (2.9) Random Forest Technique

2.7.4. Long-Short Term Memory

LSTM is an artificial Recurrent Neural Network (RNN) architecture used in the field of deep learning. It is an efficient algorithm to construct a sequential time series mode It's well known that RNN is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. RNNs can use their internal state (memory) to process variable-length sequences of inputs. A RNN can be thought of as multiple copies of the same network, each passing a message to a successor, they might be able to connect previous information to the present task. However, as that gap grows, RNNs become unable to learn to connect the information. The short-term memory problem of RNN is that short-term memory has a greater impact, but long-term memory has a small impact [44] **.** LSTM models are often applied in time series analysis. Within LSTM networks, cells form units, and units form layers. Each cell has input, output, and forget gate, Equations (2.8, 2.9, 2.10) show the form of the forward pass of the LSTM unit the *input gate* i_t ,

the *output gate* o_t , and the *forget gate* f_t . The forget gate decides what can be propagated from the previous memory units, the input gate which information must be accepted, the output gate generates the new long-term memory. Given the input sequence $x_t$and the number *h* of hidden units, the gates are defined as follows [45]:

input gate: $i_t = \sigma(x_t W_{xi} + H_{t-1} W_{hi} + b_i)$ (2.8)

forget gate: $f_t = \sigma(x_t W_{xf} + H_{t-1} W_{hf} + b_f)$ (2.9)

output gate: $o_t = \sigma(x_t W_{xo} + H_{t-1} W_{ho} + b_o)$ (2.10)

The heart of a LSTM network is it's cell or say cell state which provides a bit of memory to the LSTM so it can remember the past. Equations (2.11), (2.12), (2.13) show the cell state, candidate cell state and the final output:

intermediate cell state: $\tilde{c}_t = \tan h\ (x_t W_{xc} + H_{t-1} W_{hc} + b_c)$ (2.11)

cell state: $c_t = f_t\ o c_{t-1} * o\tilde{c}_t$ (2.12)

new state: $h_t = o_t\ o \tan(c_t)$ (2.13)

Where, o is the element-wise multiplication $W_{xi}\ W_{xf}\ W_{xo}\ \ W_{hi}\ W_{hf}\ W_{ho}$ are the weight parameters, and $b_i\ b_f\ \ b_o$ the bias parameters. The sigmoid σ and tangent functions *tanh* are the activation functions.[35]

2.7.5. Rule-Based Systems

Expert systems have been increasingly popular for commercial applications. A rule-based system is a special type of expert system. A rule-based system typically consists of a set of if-then rules, which can serve many purposes such as decision support or predictive decision making in real applications. One of the main challenges in this area is the design of such systems which could be based on both expert knowledge and data. Thus the design techniques can be divided into two categories: expert-based construction and data-based construction. The former follows a traditional engineering approach, while the latter follows a machine

learning approach. For both approaches, the design of rule-based systems could be used for practical tasks such as classification, regression, and association.

Any rule consists of two parts: the IF part, called the antecedent (premise or condition) and the THEN part called the consequent (conclusion or action). The basic syntax of a rule is: **IF <antecedence > THEN < consequent > .**

In general, a rule can have multiple antecedents joined by the keywords AND (conjunction), OR (disjunction) or a combination of both. However, it is a good habit to avoid mixing conjunctions and disjunctions in the same rule. The antecedent of a rule incorporates two parts: an object (linguistic object) and its value.

As soon as knowledge is provided by a human expert, we can input it into a computer. We expect the computer to

1- Act as an intelligent assistant in some specific domain of expertise or to solve a problem that would otherwise have to be solved by an expert.

2- Be able to integrate new knowledge and to show its knowledge in a form that is easy to read and understand, and to deal with simple sentences in a natural language rather than an artificial programming language.

3- Explain how it reaches a particular conclusion.

In other words, an expert system is a computer program capable of performing at the level of a human expert in a narrow problem area. The most popular expert systems are rule-based systems. A great number have been built and successfully applied in such areas as business and engineering, medicine and geology, power systems, and mining.[48]

2.7.5.1. Rule Coverage Measure

A rule set can consist of multiple rules RS = {R1, R2, ..., Rn} A rule R covers an instance Ij if the attributes of the instance satisfy the condition of the rule. The Coverage of a rule is the number of instances that satisfy the antecedent of a rule

Coverage of a rule:

- Count (instances with antecedent) / Count (training set) [49].

2.8. Evaluation Metrics

The essential part of any forecasting technique is model evaluation. After the model's creation, the basic concept of model accuracy is to compare the original target with the predicted one to calculate the predicted mistakes. The Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the mean square error (MSE), and R-Squared are the standard metrics used in regression analysis to compute the model performance. [46]

In this study, the regression model has been applied for prediction, so both the MAE and RMSE measures are used to evaluate the model accuracy.

2.8.1.1. Root Mean Square Error:

It is a commonly used measure of the model's differences between predicted and actual values. This metric gives an indicator of the error distribution. The RMSE can be expressed mathematically as in Equation (2.14) [47]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2} \qquad (2.14)$$

where n is the number of instances, P_i and A_i are the predicted and actual values of an instance i.

2.8.1.2. Mean Absolute Error:

MAE is a standard accuracy measure extracted by taking the average value of the absolute difference for the whole set of regression variables. The MAE can be explained in Equation (2.15) [47] :

$$MAE = \frac{1}{N} \sum_{i=1}^{n} |P_i - A_i| \quad (2.15)$$

where n is the number of instances, P_i and A_i are the predicted and actual values of an instance i.

2.9. Visualization

Data visualization is a useful and powerful mechanism to aid the user during both data preprocessing and data mining. Through the visualization of the original data, the user can browse to acquire a "feel" for the properties of that data. In particular, visualization may be used for outlier detection, Furthermore, a visual interface aids the user in selecting the appropriate data. Data visualization strategies are classified in respect of three aspects. Firstly, their focus, i.e. symbolic versus geometric; secondly their stimulus (2D versus 3D); and lastly, their display (static or dynamic), In addition, data in a data repository can be viewed as different levels of granularity or abstraction, or as different combinations of attributes or dimensions. The data can be presented in various visual formats, including box plots, scatter plots, 3D cubes, data distribution charts, curves, volume visualization, surfaces, graphs, or maps.[50]

2.9.1. A geographic information system (GIS) mapping

GIS combines layers of data about a location to create a sufficient understanding of that location. GIS is defined in a variety of ways. The United States Geological Survey (USGS) defines it as a computer system that analyzes and displays referenced data geographically. it makes use of data that is linked to a certain location. GIS mapping is the process of creating a map by entering data layers

into GIS software, and this is the most effective way to display geographical data. Thus, humans absorb information that is represented visually better than in numerical form [51]. There are many types of GIS maps like Choropleth maps, Heatmap, Proportional symbol maps, Dot density maps, and Animated time series maps. This thesis uses Choropleth maps to present our work visually**.**

2.9.1.1. Choropleth Maps

The choropleth Maps display geographical regions that are colors, shaded, or patterns in accordance to a data variable. The data variable represents itself by using a density of color in each geographical region of the map. either blending between two colors, progression of a single hue, transparent to opaque, dark to light, or a full-color spectrum. This type of map is useful to visualize the change of a variable over defined locations. [52]

Chapter Three
The Proposed Methodology

3.1 Introduction

The methodology presented in this chapter began with preprocessing the data with Sifting, Features Extraction, Aggregation, Handling missing values, Normalization, and Data Integration. Following that, re-weighting specific data in order to prepare it for the Forecasting and Evaluation steps. In addition to Knowledge Discovery about the relationship between disease spread and human mobility, with an interpretation of the correlation patterns found through that procedure, the final step is the Visualization step with a dynamic map to show the spread density of Covid-19 as well as the human movement during the same time period and to show how close the forecasting was to the actual values.

3.2 The Architecture of the Proposed Methodology

The proposed architecture consists of three main models (Knowledge Discovery Model, Forecasting Model, and Visualization Model). The Knowledge Discovery Model involves two stages: The first stage where the two datasets are preprocessed with Sifting, Extract Features, Aggregation, Handling missing values, integration of the two datasets, and Normalize them to be prepared for the next stage where the Correlation Coefficient between human mobility and the spread of Covid-19 new infections is explored through the Knowledge Discovery Model. On the other hand, the Forecasting Model involves three main stages (Re-weight the data, Forecasting, and, Evaluation). The re-weighting step was about assigning weights to the newly infected cases of Covid-19 in order to add importance to the latest months of the infection history which the forecasting process basically depends more on them, the second step represents the Forecasting using Random Forest Regressor to predict the newly infected cases of the disease as well as prediction of mobility patterns (Driving and Walking). The last step includes an Evaluation of the results to estimate the predicted error based on the test data, in addition, to comparing the Random

Forest prediction results with predictions from (ARIMA, LSTM, and LR) techniques.

The Visualization Model represents the final model, which depicts the Covid-19 spread and human movements over a period of nearly 13 months for the seven countries involved in the work on a dynamic map, as well as visualizes the predicted values for both covid-19 infections and human mobility to demonstrate the accuracy of the forecasted results. Figure (3.1) shows the methodology diagram in detail.

3.3 Preprocessing Stage

At this stage, the data should be in a form that suits our work, so, a set of operations are applied to the Covid-19 and human mobility datasets such as (Data sifting, Feature extraction, Data aggregation, Handling missing values, and Normalization) to prepare the whole dataset for the next step of finding the correlation between the disease spread and human mobility.

3.3.1. Preprocessing Steps of the Covid-19 Dataset

The epidemic dataset requires some preprocessing as shown in the following steps:

1- Sifting the Epidemic Dataset:

The sifting process for the epidemic dataset includes selecting a specific region of interest from the huge WHO dataset that consists of data for all countries around the world.

2- Feature Extraction of the Covid-19 dataset:

The epidemic dataset required the exclusion of cumulative new cases, death, and cumulative deaths. Thus, the study focused on the new infections of the disease acquired from the World Health Organization (WHO) official website.

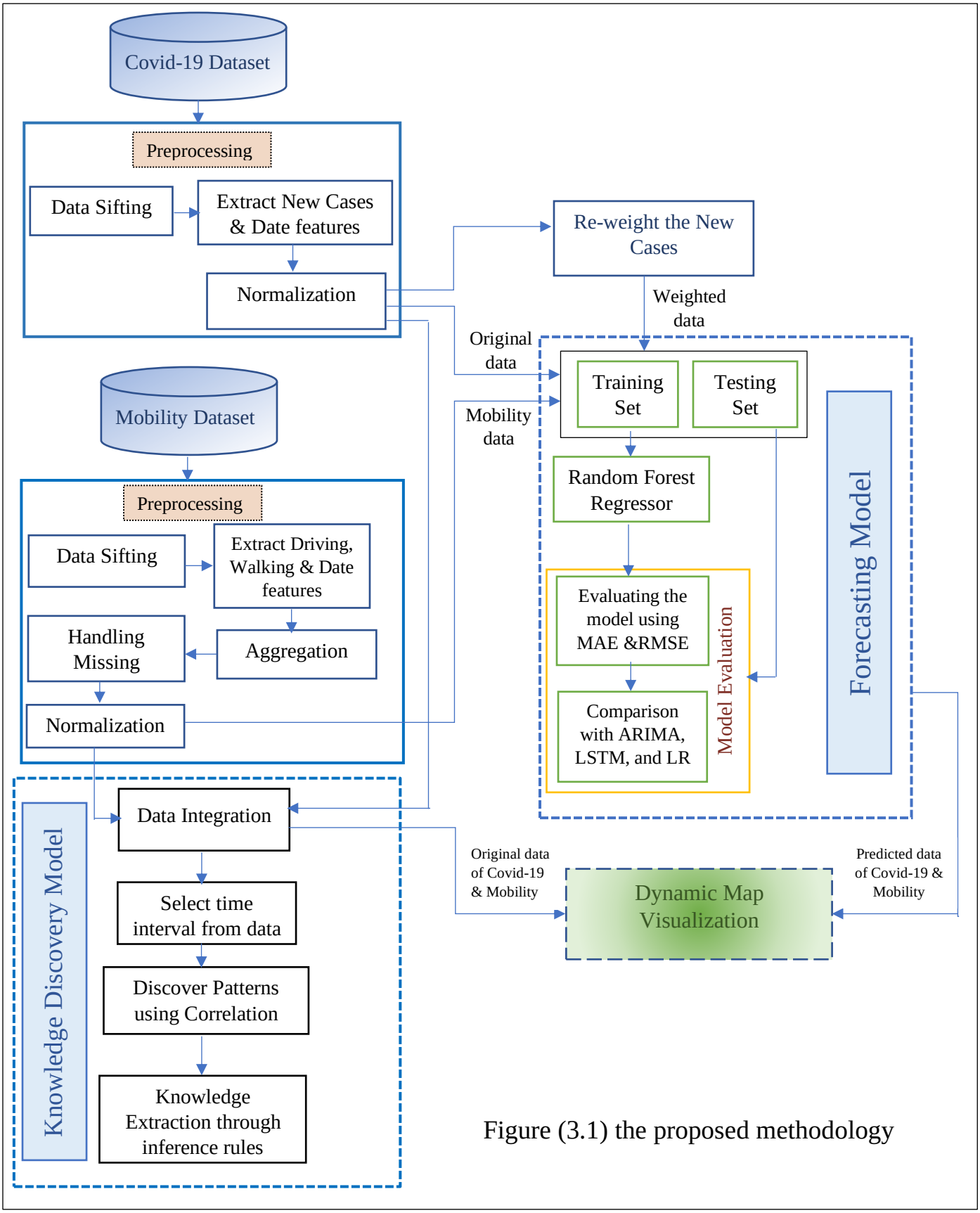


Figure (3.1) the proposed methodology

3- Normalization of Covid-19 Dataset:

Due to the large disparity in the values of the epidemic data over the countries and the huge diversity with the mobility dataset, it was necessary to normalize the data into a scale between 0 and 1 by using the Min-Max normalization formula to make the data proportioned to each other in order to deal with them together in the next stages. Algorithm (3.1) illustrates data normalization steps.

3.3.2. Preprocessing of Mobility Dataset

1. Sifting the Mobility Dataset

The mobility dataset sifting according to this work work includes choosing a certain focus area from the massive Apple mobility trends dataset, which contains data for most regions and countries around the world.

2. Feature Extraction of the Human Mobility Dataset

The mobility dataset requires the exclusion of transit. As a result, thus, concentrated primarily on driving and walking, which have been selected from Apple's mobility trends reports official website as mentioned before.

3. Aggregation of the Data

The aggregation procedure needed in this work was to aggregate mobility trend data for all regions within a specific country that was involved in the study to make it compatible with WHO data that contains countries' data. Algorithm (3.2) shows the aggregation steps.

4. Handling Missing Values of Mobility Dataset

The little bit gaps within Apple dataset have been processed where the mean method has been applied to calculate the missing values by taking the average of

two data around the missing one in order to have full data that explains the daily trends clearly as illustrated in Algorithm (3.3).

5. Normalization of the mobility dataset

As mentioned before in the Covid-10 dataset pre-processing step, the normalization needed to make both datasets included in this work proportioned to each other so, scaling the data in the range of 0 and 1 is the best way to achieve that purpose. Data normalization steps are illustrated in Algorithm (3.1).

Algorithm (3.1) Normalization

***Algorithm name**: Normalization*
***Input**: Array $X_{m,n}$ where m: the number of instances, n:the number of features.*
***Output**: Array $Z_{m,n}$ with normalized values.*
***Definition:** Min [j]and Max [j] two arrays that hold the min & max value for any feature*
Begin
***Step 1**: For j= 1 to n :*
***Step 2**: Min [j] ⟵ 0, Max [j] ⟵ 0 ; // Initialization*
***Step 3**: For i = 1 to m:*
***Step 4:** If (x_{ji} < Min [j]) then*
***Step 5:** Min [j] ⟵ x_{ji}*
***Step 6:** end if*
***Step 7:** Else (if x_{ji} > Max [j]) then*
***Step 8:** Max [j] ⟵ x_{ji}*
***Step 9:** end if*
***Step 10:** Update Max [j] and Min [j];*
***Step 11:** end for*
***Step 12:** For i = 1 to m*
***Step 13:** $Z_{ji} \longleftarrow \frac{x_{ji} - \min[j]}{\max[j] - \min[j]}$*
***Step 14:** end for*
***Step 15:** end for*
***Step 16**: Return Z_{mn}*
END

Algorithm (3.2) Aggregation Algorithm

***Algorithm name**: Aggregation*

***Input**: Array $X_{m,n}$ represents Apple Mobility Dataset where m: number of instances, n: number of features.*

***Output**: Array $M_{t,2}$ represents the aggregated features where t: number of instances,2: number of features(Driving & Walking).*

***Definition**:*

r: city , C: country

D_k : array that aggregate the Driving transportation_type for all cities belong to specific country on specific day (for each day),k:the number of days

W_s : array that aggregate the Walking transportation_type for all cities belong to specific country on specific day (for each day).s: the number of days

P: the first day from 417 days (where each day is a feature).

Begin:

Step 1: *For each $r \in C$ do*

Step 2: *$k \longleftarrow 0$, $s \longleftarrow 0$*

Step 3: *For j = P to n*

Step 4: *For i = 1 to m //all cities*

***Step 5**:* *If transportation_type = Driving then*

Step 6: *$D_k \longleftarrow D_k + X_{ij}$*

Step 7: *end if*

Step 8: *Else*

Step 9: *If transportation_type = Walking then*

Step 10: *$W_s \longleftarrow W_s + X_{ij}$*

Step 11: *end if*

Step 12: *end for*

Step 13: $D_k = \frac{D_k}{m}$, $W_s = \frac{W_s}{m}$

Step 14: k ++ , s++

Step 15: *end for*

Step 16: *end for*

Step 17 *For i = 1 to t*
Step 18: *For j=1 to 2*
Step 19: $M_{ij} \leftarrow D_{i,1} \cup W_{i,2}$
Step 20: *end for*
Step 21: *end for*
Step 22: *Return* $M_{t,2}$
End

Algorithm (3.3) Handling Missing Values

Algorithm name*: Handling Missing Values*
Input*: Array* $X_{m,n}$ *with some missing values, where m: the number of instances, n: the number of features.*
Output*: Array* $X_{m,n}$ *after process the missing values.*
Definition*: sum: summation counter*
Begin
Step 1: *sum* $\leftarrow$ *0*
Step 2: *For k = 1 to n*
Step 3: *For l = 1 to m*
Step 4: *If v is missing then* *// v any value in feature l*
Step 5*:* *sum* $\leftarrow v_{k-1,l} + v_{k+1,l}$
Step 6: *v =sum / 2*
Step 7: $X_{k,l} \leftarrow v$
Step 8: *end if*
Step 9: *end for*
Step 10: *end for*
Step 11: *return* $X_{m,n}$

End

3.4. Knowledge Discovery Model

The aim of this part of this work is to discover patterns of the relationship between human mobility and the spread of COVID-19 by examining the degree of correlation between them according to time series analysis. The following steps have been achieving that work: (For complete details, see Algorithm (3.5))

3.4.1. Integration of the Data:

As mentioned before the methodology input comes with two datasets. Since the data acquisition process was from multiple sources, the two datasets have been consolidated to create a uniform dataset that could be used as input for the next step in the Knowledge Discovery Model to find out the patterns of correlation between the combined features (see Figure (3.2)). this study carried out the integration process separately for each country involved in the work. Since the data integration process must be done with caution, both datasets for a specific country have been examined and their features were integrated according to the date feature. The integration steps are clarified in Algorithm (3.4).

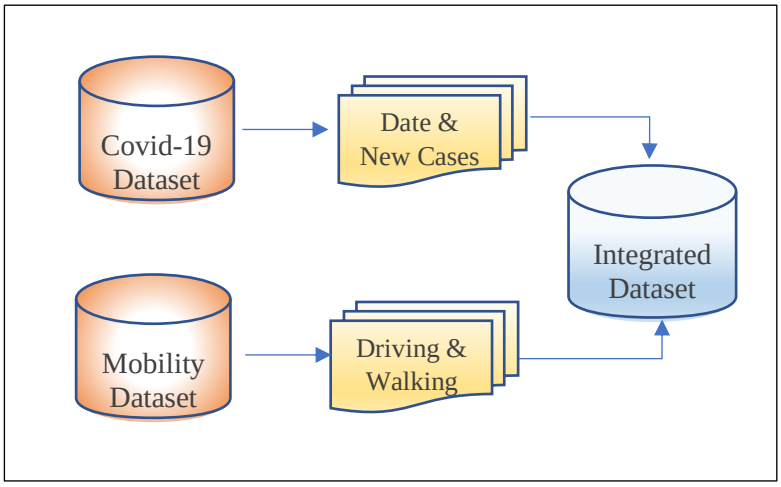


Figure (3.2) The Integration procedure

Algorithm (3.4) Integration

***Algorithm name**: Integration*

***Input**: Two Arrays $X_{m,n}$ & $Y_{m,l}$ that represent WHO & mobility datasets, where m: the number of instances, n: the number of WHO dataset features (date, new cases),l : the number of mobility dataset features (date, driving, walking).*
***Output**: Array $DS_{m,4}$ represents the integrated dataset with m instances and 4 features (date, new cases, driving, and walking)*
***Definition:** w:WHO features(1:date,2: new cases),z:mobiity features(2: driving, 3:walking)*
Begin

Step 1: *For i = 1 to m*
Step 2: *For j=1 to (n+(l-1))*
Step 3: $DS_{ij} \leftarrow X_{i,w} \cup Y_{i,z}$ *// w = 1,2, z=2,3*
Step 4: *end for*
Step 5: *end for*
Step 6: *Return* DS_{m4}
End

3.4.2. Interval's Selection and Correlation Patterns Discovery

The selection of parts of the dataset as the data to be used in the analysis of the relationship between the epidemic spread and human mobility has to be specific. This entails determining which part of the data must be used to obtain knowledge. The data from 11-March-2020 till the end of April – 2021 for each country have been partitioned into three periods, using mobility values as a partitioning criterion as follows:

1- The first period has been chosen after seeking low mobility values in proportion to the entire mobility dataset. For all countries, this type of data converges in the month of March 2020, the month of announcing the disease as a pandemic, where a lot of countries' policymakers and authorities around the world declare a lockdown

by enforcing restrictions on gatherings and limiting human movement in numerous and crowded locations by closing schools, workplaces, and most of the state's facilities, in addition to canceling public events and much more. China was the first country to implement this strategy by establishing a full lockdown on its residents, restricting their movement, and forcing them to stay at home. This approach was very effective and they were able to curb the spread of the virus. Thus, many countries have started to employ the same nonpharmaceutical intervention strategy, so, countries from all over the world have been chosen to implement kind of closure, including the United Arab Emirates from Asia and the Middle East, Southern Africa from Africa, The United Kingdom, Germany, and Denmark from Europe, and Brazil from Latin America. But the dates of these countries' lockdown weren't taken as the starting point of the study to determine the effect of human movements on the epidemic spread because there wasn't a clear citizen's commitment to the stay-at-home recommendation at the beginning, so instead, a period of time in which there was a fair commitment to closure through the same March month has been chosen to clarify the correlation clearly. The number of days that the searcher used in the process of correlation finding ranged between two weeks to three months according to the low mobility, and shifting between 1-3 days as an incubation period. The values of Pearson's correlation coefficient obtained were positive numbers above 0.5, which indicates a strong correlation between the new infection of the disease and human mobility as the relation between them is direct when get closed to one.

2- the second period was chosen based on the high mobility rate, which is the opposite of the closure that appears in the second half of the year 2020 when Apple mobility trends exceed the base 100 and sometimes more to clarify its correlation with infectious disease transmission. Many countries' political, economic, and social systems have deteriorated as a result of the closure, which

lasted many months. So, decision-makers in those countries moved to lift the lockdown completely or partially due to the aforementioned reasons, and not due to the decline of the epidemic or control it. This action comes with bad consequences where the daily infections increased for the study countries, the correlation was positive between viral transmission and human movement, in addition to the correlation coefficient that exceeds 0.5 for those countries. The number of days used to analyze and find the correlation ranged between one month and three months in the second half of the year 2020.

3- The third period of time was the first four months of the year 2021, where the mobility values oscillated between high and low. Here the recorded values of the correlation coefficient don't give any indication of the relationship pattern except it is a fluctuating relationship where for some countries there was a weak positive correlation and for other countries, there was a negative correlation or no correlation. Thus, the actual impact of the mobility on epidemic spread couldn't be clarified well in this period.

3.4.3. Inference Rules and Knowledge Extraction

After the analysis and the calculation of the correlation between the spread of the disease and human movements, patterns for that correlation have been reached in this study and explained by five rules. The rule of inference takes the linguistic premises in the form (If-Then) and delivers the output conclusion or knowledge. The five rules are illustrated as follows:

R1: **IF** correlation coefficient >= θ_1 **THEN** disease spread and human mobility are strongly related.

R2: **IF** correlation coefficient < θ_1 and => θ_2 **THEN** disease spread and human mobility are moderately related.

R3: **IF** correlation coefficient < θ_2 and > θ_3 **THEN** the disease spread and human mobility are barely related.

R4: **IF** correlation coefficient = θ_3 **THEN** the disease spread and human mobility unrelated.

R5: **IF** correlation coefficient < θ_3 **THEN** disease spread and human mobility negatively related

Where:

θ_1 = 0.6, θ_2 =*0.4*, θ_3 =*0*

Then the final decision is based on the coverage of each rule as in the following equations:

$$\textbf{Strong correlation} = \frac{\sum_{i=1}^{n} \boldsymbol{strongly\ related + moderately\ related}}{\boldsymbol{number\ of\ countries}}$$

$$\textbf{Weak correlation} = \frac{\sum_{i=1}^{n} \boldsymbol{bearly\ related}}{\boldsymbol{number\ of\ countries}}$$

$$\textbf{No correlation} = \frac{\sum_{i=1}^{n} \boldsymbol{unrelated}}{\boldsymbol{number\ of\ countries}}$$

$$\textbf{Negative correlation} = \frac{\sum_{i=1}^{n} \boldsymbol{negativly\ related}}{\boldsymbol{number\ of\ countries}}$$

$$\textbf{Fluctuating correlation} = \frac{\sum_{i=1}^{n} \boldsymbol{unrelated+related+negativly\ related\ +moderately\ related+strongly\ related}}{\boldsymbol{number\ of\ countries}}$$

Algorithm (3.5) Knowledge discovery model

***Algorithm Name**: Knowledge discovery*

***Input**: Array X_{mn} represent mobility, where m: the number of instances. n: the number of features.*

Array Y_m represent Covid-19 new cases, where m: the number of instances.

***Output**: Relationship patterns.*

***Definition:** Array r_5 represents the five rules, S1,S2,S3: counters, C_{ln} : correlation coefficient array between (Covid19-new cases and driving ,*

Covid19 new cases and walking) for 6 countries, where l: the number of instances (countries) , n: the number of features.

Begin

//calculate Pearsons correlation coefficient between Covid -19 new cases and mobility trends

Step1*: For l= 1 to 6*

Step2: *For i = 1 to n*

Step3: *For j = 1 to m*

Step4: $S1 = S1 + (x_{ij} - \bar{x})(y_j - \bar{y})$ *// $\bar{x}$ mean of x, $\bar{y}$ mean of y*

Step5*:* $S2 = S2 + (x_{ij} - \bar{x})^2$

Step6: $S3 = S3 + (y_j - \bar{y})^2$

Step7: *End For*

Step8: $C_{li} = \frac{S1}{\sqrt{S2\sqrt{S3}}}$ *// the Pearson correlation formula*

Step9*: End For*

Step10: *End For*

//Inference rules and knowledge extraction

Step 11: $\theta_1 = 0.6, \theta_2 = 0.4\ \theta_3 = 0$ *//correlation thresholds*

Step 12: r[5] ⟵ {0}

Step 13*: For l=1 to 6* *// for the six countries*

Step14: *if* $C_{l1} >= \theta_1$ and $C_{l2} >= \theta_1$ then

Step15**:** $(r_1++);$

Step16**:** *else if* $\theta_1 > C_{l1} >= \theta_2$ *and* $\theta_1 > C_{l2} >= \theta_2$ then

Step17**:** $(r_2++);$

Step18: *else if* $\theta_2 > C_{l1} > 0$ and $\theta_2 > C_{l2} > 0$ then

Step19: $(r_3++);$

Step20: *else if* C_{l1}= 0 and C_{l2} = 0 then

Step21: *(r_4++);*

Step 22: *else if* $C_{l1} < 0$ and $C_{l2} < 0$ then

Step 23: *(r_5++);*

Step 24: *end if*

Step 25: *end for*

Step 26: *end for*

//calculate the coverage for each rule:

Step 27: *Strong ⟵ (r1+r2) / total number of countries;*

Step 28: *Weak ⟵ (r3) / total number of countries;*

Step 29: *unrelated⟵ (r4) / total number of countries;*

Step 30: *Fluctuating ⟵ (r1+r2+r3+r4+r5) / total number of countries;*

Step 31: *If (Strong > Weak & unrelated & Fluctuating) then*
Step 32: *Relation:="strong";*
Step 33: *else if(Weak > Strong & unrelated & Fluctuating) then*
Step 34: *Relation:="weak";*
Step 35: *else if(unrelated > Strong & Weak &Fluctuating) then*
Step 36: *Relation:="unrelated";*
Step 37: *else*
Step 38: *Relation:=" Fluctuating ";*
Step 39: *end if*
End.

3.5. Forecasting Stage:

This stage started by experimenting four different models (ARIMA, LSTM, LR, and Random Forest regressor). The best one among them has been chosen based on the applied evaluation criteria, The RandomForest regressor provides close forecasts to reality. Figure (3.4) and Algorithm (3.6) depict the entire procedure of the chosen model, and the steps below go through the work in depth.

1- Random Forest Regressor for DSC Time Series Forecasting:

The random forest regressor is an ensemble algorithm based on a decision tree, where the features in the forest's trees are drawn randomly. At the beginning the model have ben supplying with the raw data of the outbreak's new daily cases for the seven countries involved in the study from March 11, 2020, to April 30, 2021. The data was divided into 80% for training and 20% for testing, and the parameters initialization of the model were as follows:

- The number of trees in the forest = 200
- The function to measure the quality of a split is “squared_error” which is equal to variance reduction as splitting criterion.
- The minimum number of samples required to split an internal node=2.
- The minimum number of samples required to be at a leaf node =1.
- Controlling the randomness of the bootstrapping of the samples used when building trees.

After getting a good result, an attempt to enhance them has been introduced, which was adding weights to the New Cases feature of Covid-19 infections to add more importance to the recent months which is the feature infections depend primarily on them, So, after dividing the data into two parts (Training and Testing), we went through another process to handle the training data in a way that would

allow us to make very close predictions to the truth. Thus, the New Cases feature of the training dataset has been split into two subsets, the values of the New Cases of the disease for the first subset represent (200 days out of 333) being reduced by 80 percent by multiplicating the values of New Cases of Covid-19 with the factor 0.2, the factor value have been chosen as the best factor that the study reached based on trial and error, whereas the rest of the data (the second subset) was increased by double by multiplicate the values of New Cases of Covid-19 with the factor 2, the factor value have been chosen as the best factor that the study reached based on trial and error . The idea behind that is to give a greater emphasis for the chronology of time series. In this way, the New Cases feature has been re-weighted in a way that shows the importance of the past few months of the infection's history in forecasting Covid-19 New Cases for 83 day with RandomForest Regressor. This strategy decreases errors and provides accurate forecasts even when employs on the other three forecasting models.

2- Random Forest Regressor for DSM Time Series Forecasting:

The Random Forest regressor was fitted to the Mobility Data to predict the Driving and Walking as a human transmission mechanism by applying Algorithm (3.6).

3- Evaluation

This stage involves evaluating the forecasting results, which was done using MAE and RMSE to compute the error rate between the actual and predicted values for both Covid-19 new infections and Human Mobility. Also, a comparison between the evaluation measures for (ARIMA model, LSTM, LR, and the Random Forest regressor) has been introduced , the study also, presented the actual and predicted

values for the Covid-19 New Cases and the Mobility trends of all the involved countries.

Algorithm (3.6): Random Forest regressor for time series

Algorithm name: *Random Forest regressor for time series*

Input: *the datasets of Covid-19 raw data and the weighted data as well as the mobility dataset.*

Output*: the ensemble of trees {T_b}1^B with less MAE & RMSE as well as the predicted values of the input data*

Begin

Step1: *Until the minimum number of node size (n min) is reached do*

Step 2*: B bootstrap samples are taken from the training data each sample has size n, the same size of other samples.*

Step 3: *For b=1 to B:*

Step 4: *Grow a random forest tree* T_b *by Repeat recursively on each resulting node the following steps until a stopping criterion is met*

step 5: *a. from p predictors select m predictors randomly.*

Step 6: *b. Using the variance criterion, choose the best split among the previously chosen variables.*

Step 7: *c. Cut in accordance with the chosen split to make a prediction at new point x*

Step 8: *end for*

Step 9: *For a new observation (x) apply the regression function*

$$\hat{f}_{rf^B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\ (x).$$

Step 10: *DSCP ← predicted values of Covid-19 new cases.*

Step 11: *DSCWP ← predicted values of weighted Covid-19 new cases.*

Step 12: *DSMP ← predicted values of mobility.*

Step 13: *Compute MAE and RMSE for all testing variables using equations (2.14) and (2.15).*

End

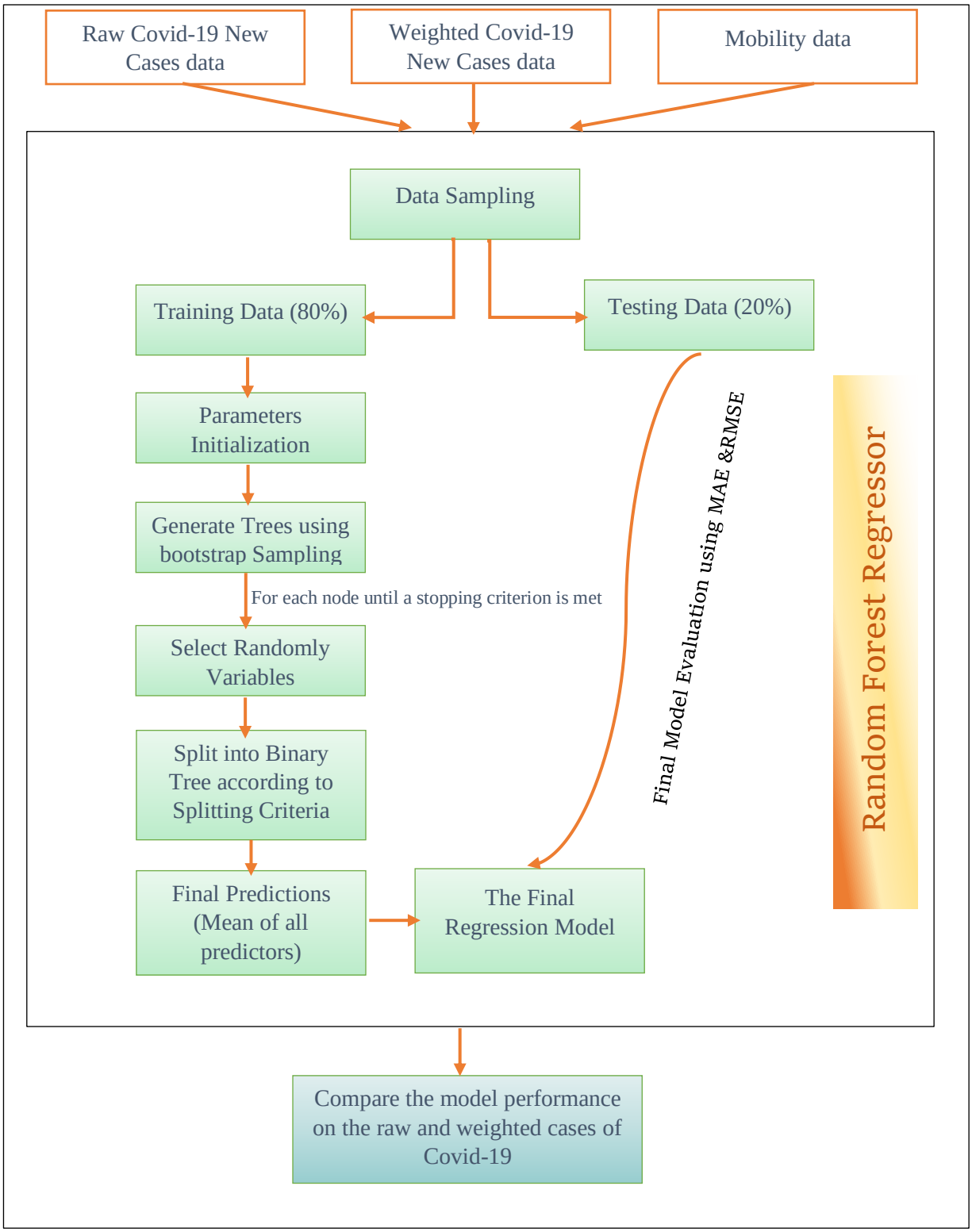


Figure (3.3) Random Forest Forecasting

3.6. Visualization

The choropleth map used in this study, which is a type of thematic map that depicts the spatial pattern of a specific subject in a specific geographic area, is a simple way to see how a variable change over time or to show the degree of variability within a region. So, how the disease spreads over time and how mobility affects that action has been clarified with this technique, especially at the beginning of the announcement of the disease as a pandemic, when some policymakers and authorities issued recommendations to stay at home and restrict movement, as well as to visualized predicted Covid-19 infections and the predicted mobility to show the closeness to the truth on a dynamic worldwide map. The steps of the visualization procedure are applied as follows: (for more details see Algorithm (3.7)).

1- Visualize the whole Dataset of Covid-19 and Mobility:

a- Integrate the datasets of Covid-19 and mobility for all the six countries together.

b. Add a new feature to represent the country name and (iso_alpha_3) feature to represent the three-letter country code that states the official code for each country. The iso_alpha_3 is an important addition that allows the map to precisely allocate the country position.

c. Visualize the data of Covid-19 and Mobility for all the involved countries on a dynamic worldwide map according to the chronology of the new infections.

2- Visualize the Predicted Data of Covid-19 and Mobility:

In this part, the above steps have been applied to the predicted dataset of the covid-19 and human mobility.

Algorithm (3.7) Visualization

Algorithm name: Visualization

Input: *Normalized DSC and DSM resulted from algorithm (3.1) as well as the DSCP and DSMP resulted from algorithm (3.6).*

Output: *dynamic worldwide map that visualize all the data.*

Definition:

UD ← Universal Dataset

PD ← Predicted Dataset

Begin

//Visualize the whole dataset of Covid-19 and mobility:

Step 1: *UD ← aggregate all countries normalize dataset for DSC and DSM.*

Step 2: *add country name feature and country code (iso_alpha_3) feature to the UD.*

Step 3: *visualize the UD on worldwide dynamic map.*

//Visualize the predicted data of Covid-19 and mobility:

Step 4: *PD ← aggregate all countries data for DSCP and DSMP.*

Step 5: *add country name feature and country code (iso_alpha_3) feature to the PD.*

Step 6: *visualize the PD on worldwide dynamic map.*

End

Chapter Four

The Experimental Results

4.1 Introduction

This chapter presents the detailed results and discussion of forecasting both future COVID-19 infections and human movements. It also clarifies the comparative forecasting results on the original and modified datasets that fitted to four prediction techniques for seven countries used in this study, in addition to discovering the relationship between disease spread and human mobility for those countries, Finally, all of the results were visually represented on a dynamic map.

4.2 System Requirement

1. Hardware: Processor Intel® Core ™ i5, RAM 4 GB,
2. Operation System: Windows 10, 64-bit.
3. Programing Language: Python (3.8)
4. Supported Platform: Visual Studio 2015, and SQL Server 2014 Management Studio.

4.3 Description of Dataset

Table (4.1) shows a summarization of the data used in this study, all the participating countries, the source of the data, and the starting and ending dates of the study period are presented.

4.3.1. Collection of Covid-19 Dataset

The dataset was downloaded as "comma-separated values" (CSV file format) from WHO official website (WHO, 2020-2021) which is a completely open-access database that provides daily data on newly confirmed, death, cases for all countries. Because the start date of Coronavirus disease varies from country to country, the date when the World Health Organization announced the disease as a worldwide epidemic has been considered as the start date for this work, thus the data used for

Covid-19 new cases from March 11, 2020, till the end of March 30, 2021, as shown in Figure (4.1) for a sample of data.

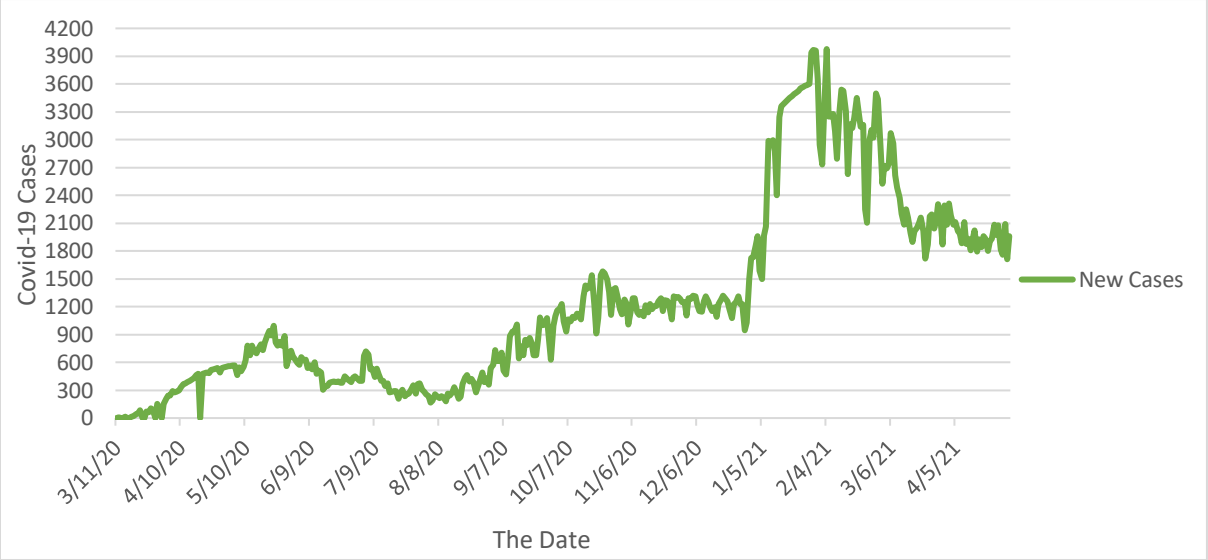


Figure (4.1) Covid-19 new infections timeline for the UAE

4.3.2.Collection of Mobility Dataset:

The Mobility data that the study used is provided by Apple Company (Apple, 2020-2021) which consists of daily human movement trends of walking, driving, and public transportation for many countries. These reports reflect route requests on Apple Maps. Figure (4.2) and (4.3) shows the mobility timeline for the UAE.

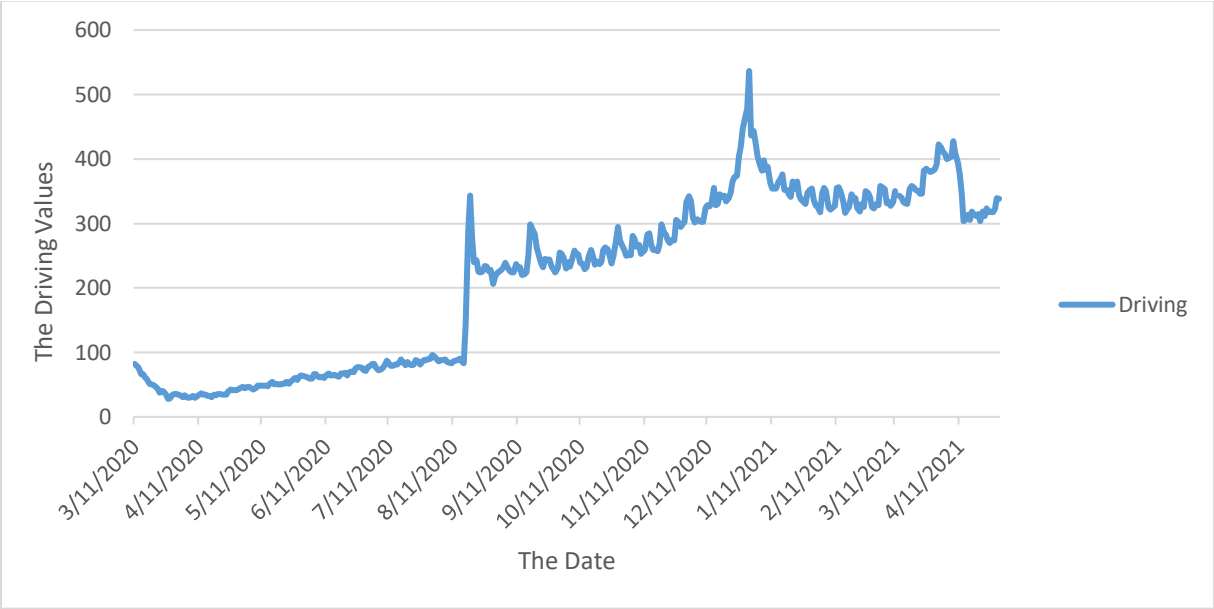


Figure (4.2) Driving mobility timeline for the UAE

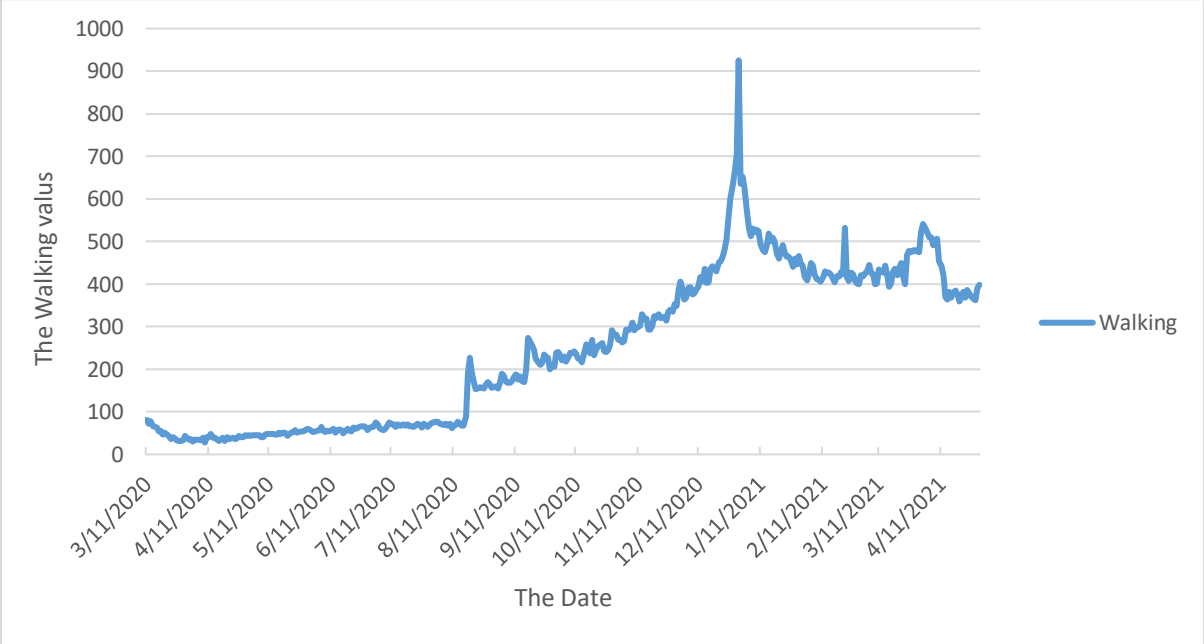


Figure (4.3) Walking mobility timeline for the UAE

Table (4.1) Summarization of dataset

The participated countries	Daily time period	Involved datasets
United Arab Emarat	From 11/3/2020 To 30/4/2021	Covid-19 Dataset from WHO & Human mobility dataset from Apple Company
South Africa		
United Kingdom		
Germany		
Denmark		
Brazil		
China	From 11/3/2020 To 30/4/2021	Covid-19 Dataset from WHO

4.4. Steps of Data Preprocessing

This study carried out the Pre-processing steps separately for each country involved in the study for both Covid-19 and Mobility datasets. The whole pre-processing has introduced in the steps that follow.

1- Covid-19 Data Preprocessing:

The first step of preprocessing the pandemic dataset revolves around sifting the data to ensure that it is compliant with the study requirements in the process of preparing the data for the work. The sifting process of the huge WHO Covid-19 dataset is about eliciting the involved countries among all the countries around the world that WHO lists. After that, the features that help in the completion of work have been extracted that is the daily Date feature and the New Cases feature, which

represent the daily reported new infections of the disease for those countries while excluding the remaining features of the WHO Covid-19 reports. The followed step

was the Normalization process which aims to organize the data in a consistent manner within the range of 0 and 1 due to the vast difference in new disease infections between the research countries. All the steps are illustrated in Figur (4.4) for a sample of data.

2- Mobility Preprocessing

The process of sifting the mobility dataset includes choosing a specific focus area from the massive Apple mobility trends dataset that have mobility information for most regions and countries around the globe. Then extracting the features needed, which is the Date, Driving, and Walking features while excluding the transit of all the regions for a specific country. After that, the aggregate process is needed in order to get the full organized country data. Also, handling the gaps in Driving and Walking is necessary to get the whole structured country data. The last step is to normalize the data in the range of 0 and 1. Figure (4.5) shows those preprocessing steps for a sample of data.

Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
3/11/2020	DE	Germany	EURO	157	1296	0	2
3/12/2020	DE	Germany	EURO	271	1567	1	3
3/13/2020	DE	Germany	EURO	802	2369	2	5
3/14/2020	DE	Germany	EURO	693	3062	0	5
3/15/2020	DE	Germany	EURO	733	3795	3	8
3/16/2020	DE	Germany	EURO	1043	4838	4	12
3/17/2020	DE	Germany	EURO	1174	6012	1	13
3/18/2020	DE	Germany	EURO	1144	7156	0	13
3/19/2020	DE	Germany	EURO	1042	8198	0	13
3/20/2020	DE	Germany	EURO	5940	14138	30	43
3/21/2020	DE	Germany	EURO	4049	18187	2	45
3/22/2020	DE	Germany	EURO	3276	21463	22	67
3/23/2020	DE	Germany	EURO	3311	24774	27	94
3/24/2020	DE	Germany	EURO	4438	29212	32	126
3/25/2020	DE	Germany	EURO	2342	31554	23	149
3/26/2020	DE	Germany	EURO	4954	36508	49	198
3/27/2020	DE	Germany	EURO	5780	42288	55	253
3/28/2020	DE	Germany	EURO	6294	48582	72	325
3/29/2020	DE	Germany	EURO	3965	52547	64	389
3/30/2020	DE	Germany	EURO	4751	57298	66	455
3/31/2020	DE	Germany	EURO	4615	61913	128	583
4/1/2020	DE	Germany	EURO	5453	67366	149	732
4/2/2020	DE	Germany	EURO	6156	73522	140	872
4/3/2020	DE	Germany	EURO	6174	79696	145	1017

Date	New Cases
3/11/2020	157
3/12/2020	271
3/13/2020	802
3/14/2020	693
3/15/2020	733
3/16/2020	1043
3/17/2020	1174
3/18/2020	1144
3/19/2020	1042
3/20/2020	5940
3/21/2020	4049
3/22/2020	3276
3/23/2020	3311
3/24/2020	4438
3/25/2020	2342
3/26/2020	4954
3/27/2020	5780
3/28/2020	6294
3/29/2020	3965
3/30/2020	4751
3/31/2020	4615
4/1/2020	5453
4/2/2020	6156
4/3/2020	6174

Normalization

Date	New Cases
3/11/2020	0.0046481
3/12/2020	0.0080232
3/13/2020	0.023744
3/14/2020	0.0205169
3/15/2020	0.0217012
3/16/2020	0.030879
3/17/2020	0.0347574
3/18/2020	0.0338692
3/19/2020	0.0308494
3/20/2020	0.1758593
3/21/2020	0.1198745
3/22/2020	0.0969891
3/23/2020	0.0980253
3/24/2020	0.1313912
3/25/2020	0.0693371
3/26/2020	0.1466679
3/27/2020	0.1711224
3/28/2020	0.1863398
3/29/2020	0.1173876
3/30/2020	0.1406578
3/31/2020	0.1366314
4/1/2020	0.1614412
4/2/2020	0.1822542
4/3/2020	0.1827871

Figure (4.4) Sample of preprocessing WHO dataset for Germany

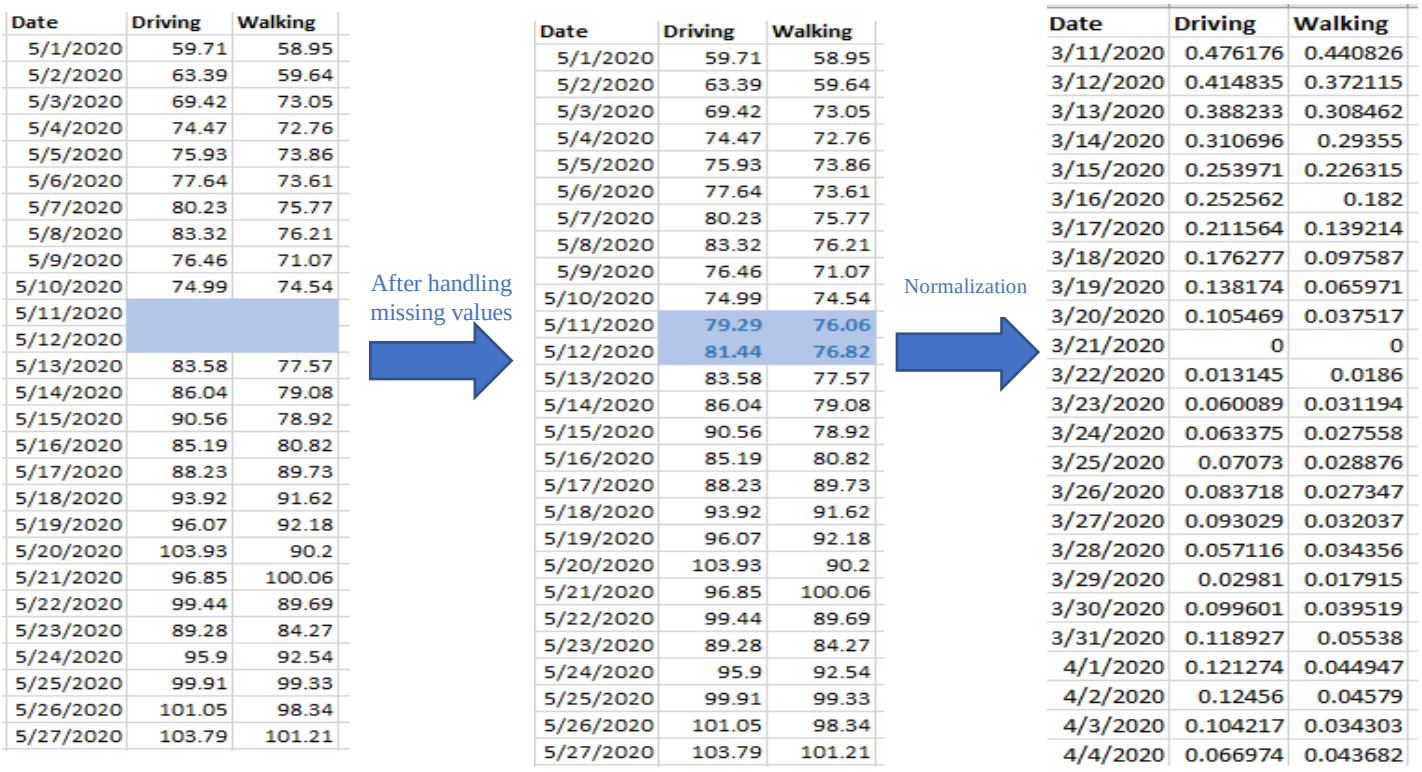
geo_type	region	transportation_type	sub-region	country	1/13/2020	1/14/2020	1/15/2020	1/16/2020	1/17/2020	1/18/2020	1/19/2020	1/20/2020	1/21/2020	1/22/2020	1/23/2020	1/24/2020
city	Aachen	driving	North Rhine	Germany	100	100.73	102.86	102.65	109.39	109.62	98.21	102.74	103.85	102.01	101.4	107.57
city	Aachen	walking	North Rhine	Germany	100	98.53	97.68	100.06	107.12	123.64	86.13	108.47	96.5	93.88	92.18	100.74
city	Augsbur	driving	Bavaria	Germany	100	98.58	97.43	103.45	110.4	120.59	101.34	96.07	95.09	103.05	97.82	113.52
city	Augsbur	walking	Bavaria	Germany	100	108.39	106.35	111.69	120.69	152.46	80.37	89.45	85.78	96.25	95.93	119.71
city	Berlin	driving		Germany	100	105.36	107.04	109.71	116.27	110.4	98.89	101.88	103.34	100.95	102.77	110.28
city	Berlin	transit		Germany	100	104.84	104.32	109.63	119.85	121.96	100.37	98.47	101.49	100.53	105.14	115.72
city	Berlin	walking		Germany	100	107.17	111.6	116.41	134.55	149.26	99.5	95.27	101.05	99.09	106.89	127.18
city	Bochum	driving	North Rhine	Germany	100	100.72	105.4	100.43	111.27	101.16	91.71	97.6	101.83	102.99	99.66	104.59
city	Bochum	transit	North Rhine	Germany	100	107.13	109.66	102.93	105.33	102.96	99	106.09	110.76	103.2	101.96	111.42
city	Bochum	walking	North Rhine	Germany	100	103.34	106.03	109.23	113.88	121.21	83.67	95.97	102.26	95.71	103.03	120.61
city	Bremen	driving	Bremen (sta	Germany	100	101.19	99.53	105.02	135.2	99.12	85.96	94.18	104.68	95.74	108.12	106.06
city	Bremen	walking	Bremen (sta	Germany	100	88.14	90.21	97.26	124.81	125.43	80.12	86.28	88.07	88.97	97.78	105.28
city	Cologne	driving	North Rhine	Germany	100	94.81	95.04	97.11	107.21	104.64	83.92	92.87	97.35	93.95	94.22	99.96
city	Cologne	transit	North Rhine	Germany	100	99.91	96.18	97.26	111.85	113.46	88.93	93.66	93.66	88.2	90.16	97.05
city	Cologne	walking	North Rhine	Germany	100	105.98	106.25	115.44	134.22	178.12	84.19	92.34	98.18	95.55	97.19	116.57
city	Dresden	driving	Saxony	Germany	100	98.17	103.5	112.26	114.78	111.79	99.2	98.91	102.76	105.84	104.54	110.74
city	Dresden	walking	Saxony	Germany	100	104.39	105.89	114.34	128.38	161.43	97.07	102.93	108.9	106.12	104.9	128.05
city	Dusseld	driving	North Rhine	Germany	100	104.58	103	103.91	114.26	105.07	95	106.28	114.81	113.61	110.98	106.21
city	Dusseld	transit	North Rhine	Germany	100	102.97	104.19	104.58	126.43	132.63	113.22	111.03	118.55	113.57	110.06	123.78
city	Dusseld	walking	North Rhine	Germany	100	107.29	107.73	112.65	139.82	167.68	96.46	111.2	118.71	114.13	120.01	133.96
city	Frankfur	driving	Hesse	Germany	100	101.83	105.38	106.95	106.42	103.71	93.43	100.06	106.47	105.02	104.42	109.65
city	Frankfur	walking	Hesse	Germany	100	105.87	111.85	112.71	122.96	139.95	85.42	104.31	109.05	104.32	109.74	119.2
city	Hambur	driving		Germany	100	101.73	102.87	106.65	112.44	105.27	99.15	100.28	104.6	105.72	104.64	110.4
city	Hambur	transit		Germany	100	98.86	99.74	102.18	115.69	123.12	103.5	103.82	101.58	105.36	105.38	119.21

Feature extraction & Aggregation (Driving)

region	transportation_type	3/11/2020	3/12/2020	3/13/2020	3/14/2020	3/15/2020	3/16/2020	3/17/2020
Aachen	driving	111.7	90.88	87.88	77.8	66.45	70.51	64.58
Augsbur	driving	89.84	84.53	87.34	78.85	67.69	65.6	59.43
Berlin	driving	94.53	89.6	83.87	75.05	64.53	70.44	64.33
Bochum	driving	94.63	89.72	83.33	75.3	64.25	66.98	63.95
Bremen	driving	94.05	87.51	84.8	72.39	60.14	68.88	66.14
Cologne	driving	85.55	87.54	76.84	70.51	54.64	57.29	51.16
Dresden	driving	101.01	91.37	91.98	77.81	72.69	74.58	72.32
Dusseldo	driving	93.09	84.61	77.37	66.68	55.87	61.38	57.18
Frankfurt	driving	103	87.23	81.8	72.78	64.57	66.64	59.27
Hamburg	driving	100.68	91.76	86.08	79.03	66.51	70.89	62.26
Hannove	driving	104.23	92.02	86.48	70.98	60.71	71.84	64.94
Karlsruhe	driving	103.22	90.78	87.18	67.82	61.96	70.17	60.04
Leipzig	driving	93.1	84.4	82.85	75.06	64.28	64.8	60.21
Mannhei	driving	97.31	90.78	83.99	77.35	67.75	69.11	65.04
Munich	driving	97.21	91.02	85.52	71.75	72.41	66.98	58.47
MÃ¶nche	driving	98.91	86.85	76.9	68.79	60.73	66.02	60.77
MÃ¼nste	driving	91.78	85.38	86.26	77.44	63	61.27	57.65
Nurembe	driving	93.5	88.74	84.43	73.3	63.93	66.57	66.36
Stuttgart	driving	94.16	86.28	79.05	62.89	57.41	61.52	52.27
Wiesbad	driving	80.52	76.09	68.07	59.23	52.19	53.08	48.08
Baden-W	driving	98.55	89.67	87.16	76.33	74.11	68.28	62.66
Bavaria	driving	97.35	90.7	91.88	79.7	77.62	67.12	61.34
Branden	driving	96.07	92.04	88.34	86.33	83.92	77.16	76.47
Bremen (	driving	94.43	88.37	85.11	73.33	62.25	69.74	67.15
Hesse	driving	105.31	91.03	88.41	78.75	72.27	70.01	63.17
Lower Sa	driving	101.18	92.48	94.2	82.9	77.16	73.9	67.96
Mecklenl	driving	104.89	98.2	104.3	99.97	93.09	78.96	72.39
North Rh	driving	94.28	88.98	83.24	75.34	66.7	65.79	62.02
Rhinelan	driving	91.47	87.74	82.2	76.83	76.29	66.11	61.15
Saarlanc	driving	99.39	89.48	85.02	78.66	75.39	73.44	67.01
Saxony	driving	96.43	90.53	89.13	77.99	74.94	72.48	71.64
Saxony-	driving	95.51	91.68	95.04	78.91	79.32	78.9	70.71
Schleswi	driving	104.38	94.02	93.53	87.73	77.72	72.46	65.83
Thuringia	driving	98.81	91.35	88.28	77.31	81.95	69.28	66.2
		97.06	89.22	85.82	75.91	68.66	68.48	63.24

Data Structuring

Figure (4.5) Continued



Date	Driving	Walking
5/1/2020	59.71	58.95
5/2/2020	63.39	59.64
5/3/2020	69.42	73.05
5/4/2020	74.47	72.76
5/5/2020	75.93	73.86
5/6/2020	77.64	73.61
5/7/2020	80.23	75.77
5/8/2020	83.32	76.21
5/9/2020	76.46	71.07
5/10/2020	74.99	74.54
5/11/2020		
5/12/2020		
5/13/2020	83.58	77.57
5/14/2020	86.04	79.08
5/15/2020	90.56	78.92
5/16/2020	85.19	80.82
5/17/2020	88.23	89.73
5/18/2020	93.92	91.62
5/19/2020	96.07	92.18
5/20/2020	103.93	90.2
5/21/2020	96.85	100.06
5/22/2020	99.44	89.69
5/23/2020	89.28	84.27
5/24/2020	95.9	92.54
5/25/2020	99.91	99.33
5/26/2020	101.05	98.34
5/27/2020	103.79	101.21

Date	Driving	Walking
5/1/2020	59.71	58.95
5/2/2020	63.39	59.64
5/3/2020	69.42	73.05
5/4/2020	74.47	72.76
5/5/2020	75.93	73.86
5/6/2020	77.64	73.61
5/7/2020	80.23	75.77
5/8/2020	83.32	76.21
5/9/2020	76.46	71.07
5/10/2020	74.99	74.54
5/11/2020	79.29	76.06
5/12/2020	81.44	76.82
5/13/2020	83.58	77.57
5/14/2020	86.04	79.08
5/15/2020	90.56	78.92
5/16/2020	85.19	80.82
5/17/2020	88.23	89.73
5/18/2020	93.92	91.62
5/19/2020	96.07	92.18
5/20/2020	103.93	90.2
5/21/2020	96.85	100.06
5/22/2020	99.44	89.69
5/23/2020	89.28	84.27
5/24/2020	95.9	92.54
5/25/2020	99.91	99.33
5/26/2020	101.05	98.34
5/27/2020	103.79	101.21

Date	Driving	Walking
3/11/2020	0.476176	0.440826
3/12/2020	0.414835	0.372115
3/13/2020	0.388233	0.308462
3/14/2020	0.310696	0.29355
3/15/2020	0.253971	0.226315
3/16/2020	0.252562	0.182
3/17/2020	0.211564	0.139214
3/18/2020	0.176277	0.097587
3/19/2020	0.138174	0.065971
3/20/2020	0.105469	0.037517
3/21/2020	0	0
3/22/2020	0.013145	0.0186
3/23/2020	0.060089	0.031194
3/24/2020	0.063375	0.027558
3/25/2020	0.07073	0.028876
3/26/2020	0.083718	0.027347
3/27/2020	0.093029	0.032037
3/28/2020	0.057116	0.034356
3/29/2020	0.02981	0.017915
3/30/2020	0.099601	0.039519
3/31/2020	0.118927	0.05538
4/1/2020	0.121274	0.044947
4/2/2020	0.12456	0.04579
4/3/2020	0.104217	0.034303
4/4/2020	0.066974	0.043682

Figure (4.5) Sample of the procedure of preprocessing mobility dataset (Germany)

4.5. Results of Knowledge Discovery Model

The model started with the integration of Mobility and WHO datasets. Each of the six countries has gone through the integration process separately. Figure (4.6) shows the integration procedure for a sample of the data. After that, the trend of the data has been analyzed for both the Coronavirus spread and human mobility statistically and attempted to select two periods of the year 2020 where human movements vary between low values (below 50%) and high values that exceed the base of the Apple trend (100) and one period of the year 2021 in order to demonstrate how human mobility affects viral transmission. After the selection of the data as intervals according to time periods, Pearson's Correlation Coefficient has been calculated to clarify the relationship between mobility and the epidemic spread where each directional relationship links two independent variables labeled with a

value in the range of -1 and 1, and because the virus's incubation period (the time between initial infection and the onset of symptoms) is usually between 1 and 3 days, thus, between 1- 3 shifting days have been chosen for the new infections of the study countries.

The above steps were implemented for all the participating countries. Table (4.2) describes the Correlation Coefficient of the uniformed dataset for the first period for those countries in detail where the start and end dates of this period represent the period when the mobility recorded low values according to Apple data. The minimum and maximum numbers of mobility were listed to clarify the low trend of mobility for each country in proportion to the whole data for a specific country. Thus, the Correlation Coefficient registers positive numbers between 0.60 and 0.80 for Driving and between 0.55 and 0.85 for Walking in all countries. In addition to Figures from (4.7) to (4.12), which depict the trend of Driving, Walking, and New Cases of COVID -19 for the same countries during the same period.

Date	New Cases
3/11/2020	0
3/12/2020	0.0014501
3/13/2020	0.0031706
3/14/2020	0.0041046
3/15/2020	0.00408
3/16/2020	0.0026545
3/17/2020	0.0036253
3/18/2020	0.0056899
3/19/2020	0.0076562
3/20/2020	0.0105073
3/21/2020	0.0110726
3/22/2020	0.0137148
3/23/2020	0.0128177
3/24/2020	0.0150912
3/25/2020	0.026766
3/26/2020	0.0273927
3/27/2020	0.0312515
3/28/2020	0.0361181
3/29/2020	0.037347
3/30/2020	0.0327754
3/31/2020	0.0332915

Date	Driving	Walking
3/11/2020	0.74991	0.576173
3/12/2020	0.654826	0.491129
3/13/2020	0.734816	0.636151
3/14/2020	0.73581	0.735842
3/15/2020	0.44107	0.320727
3/16/2020	0.461678	0.342004
3/17/2020	0.366323	0.274007
3/18/2020	0.312003	0.213524
3/19/2020	0.302061	0.206768
3/20/2020	0.324657	0.24768
3/21/2020	0.222252	0.169708
3/22/2020	0.144975	0.10108
3/23/2020	0.182845	0.13814
3/24/2020	0.092733	0.087632
3/25/2020	0.086135	0.095145
3/26/2020	0.08261	0.094829
3/27/2020	0.103489	0.114843
3/28/2020	0.057936	0.081066
3/29/2020	0.01889	0.037881
3/30/2020	0.064443	0.072795
3/31/2020	0.07791	0.095776

Integration process

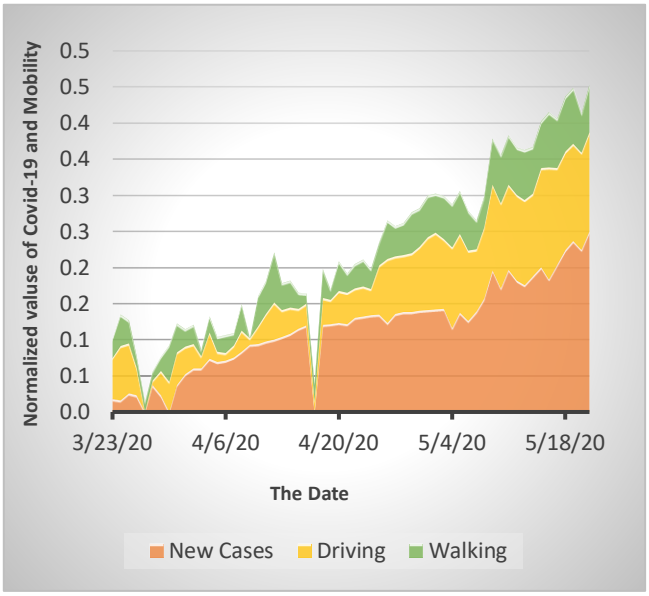
Date	New Case	Driving	Walking
3/11/2020	0	0.74991	0.576173
3/12/2020	0.00145	0.654826	0.491129
3/13/2020	0.003171	0.734816	0.636151
3/14/2020	0.004105	0.73581	0.735842
3/15/2020	0.00408	0.44107	0.320727
3/16/2020	0.002654	0.461678	0.342004
3/17/2020	0.003625	0.366323	0.274007
3/18/2020	0.00569	0.312003	0.213524
3/19/2020	0.007656	0.302061	0.206768
3/20/2020	0.010507	0.324657	0.24768
3/21/2020	0.011073	0.222252	0.169708
3/22/2020	0.013715	0.144975	0.10108
3/23/2020	0.012818	0.182845	0.13814
3/24/2020	0.015091	0.092733	0.087632
3/25/2020	0.026766	0.086135	0.095145
3/26/2020	0.027393	0.08261	0.094829
3/27/2020	0.031252	0.103489	0.114843
3/28/2020	0.036118	0.057936	0.081066
3/29/2020	0.037347	0.01889	0.037881
3/30/2020	0.032775	0.064443	0.072795
3/31/2020	0.033292	0.07791	0.095776

Figure (4.6) Sample of the Integration Process

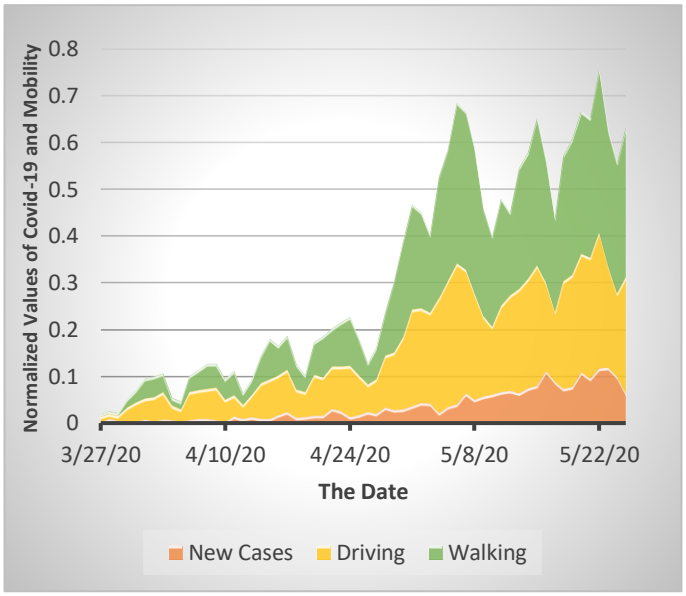
Table (4.2) Person's correlation coefficient of mobility with COVID-19 new cases according to **Low mobility Rate**

Countries	Start Date	End Date	The number of shifting days	Minimum number of mobility		Maximum number of mobility		Correlation between New Cases& Driving	Correlation between New Cases& Walking
				Driving	Walking	Driving	Walking		
ARE	3/23/2020	5/21/2020	2 Days	27.92	27.87	54.14	50.86	0.790	0.713
ZAF	3/27/2020	5/25/2020	2 Days	11.32	12.92	59.17	57.78	0.801	0.850
GBR	3/25/2020	4/23/2020	2 days	25.2	31.47	38.13	51.95	0.60	0.58
DEU	3/20/2020	4/2/2020	1 Day	36.2	37.19	52.12	47.7	0.67	0.55
DNK	3/15/2020	4/13/2020	No Shift	56.52	51.34	91.86	94.20	0.650	0.546
BRA	3/20/2020	4/18/2020	2 Days	30.62	20.65	68.34	47.81	0.653	0.659

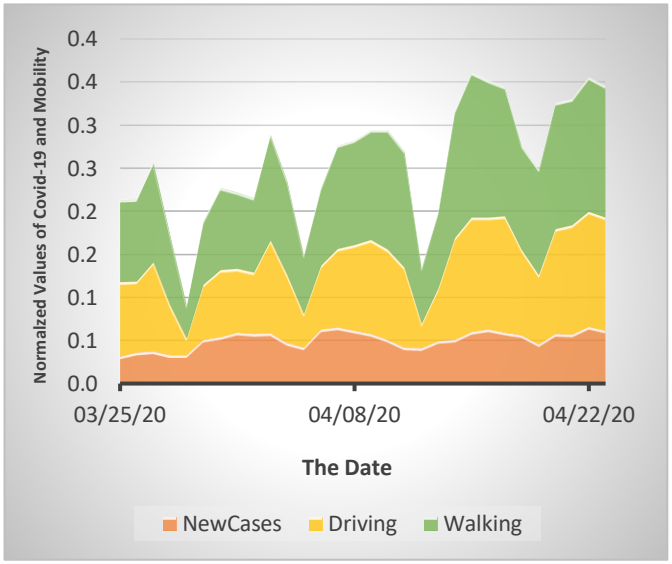
Table (4.3) listed the start and end dates for the second period for each country, where Apple registers high rates of mobility reaching a maximum of 343.57 % for driving and 292.07 % for Walking in UAE. The Correlation Coefficient values were positive numbers ranging between 0.05 and 0.8 for Driving and Walking in all countries. See also Figures (4.13) to (4.18). Table (4.4) describes how the infections increased when Apple recorded high mobility rates. All those experimental results indicate a strong positive relationship between viral transmission and mobility. Where whenever there is an increasingly obvious movement the spread of the disease increases and vice versa.



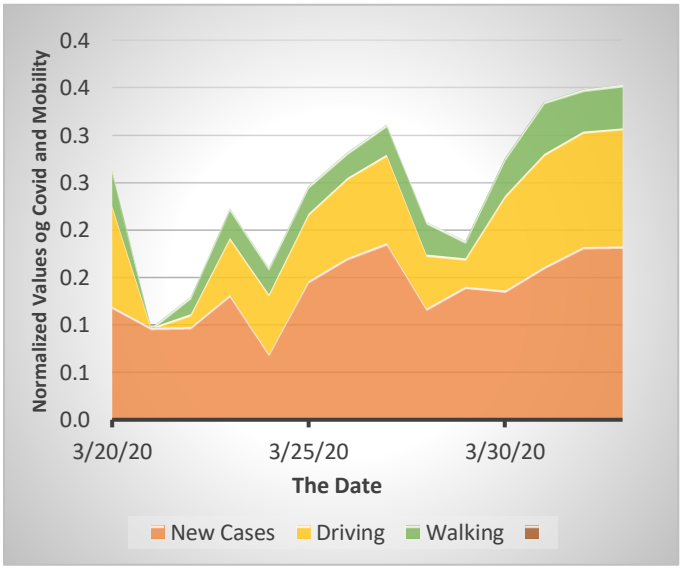
Figure(4.7) the correlation between Covid-19 New Cases and mobility through low mobility rate of (United Arab Emarat)



Figure(4.8) the correlation between Covid-19 New Cases and mobility through low mobility rate of (South Africa)



Figure(4.9) the correlation between Covid-19 New Cases and mobility through low mobility rate of (United Kingdom)



Figure(4.10) the correlation between Covid-19 New Cases and mobility through low mobility rate of (Germany)

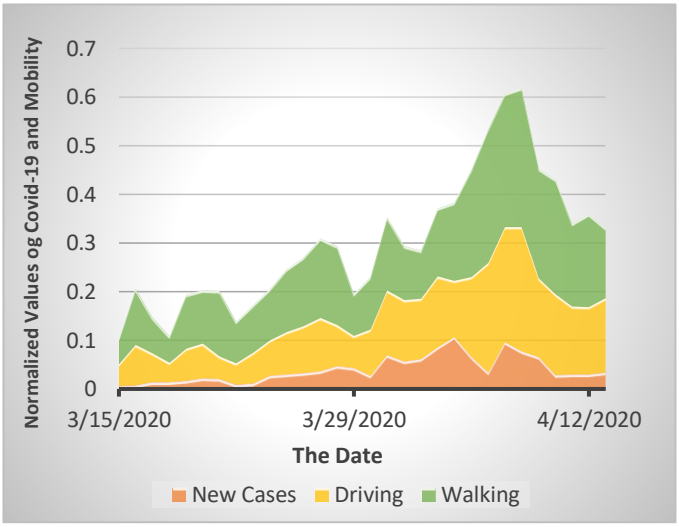


Figure (4.11) the correlation between Covid-19 New Cases and mobility through low mobility rate of (Denmark)

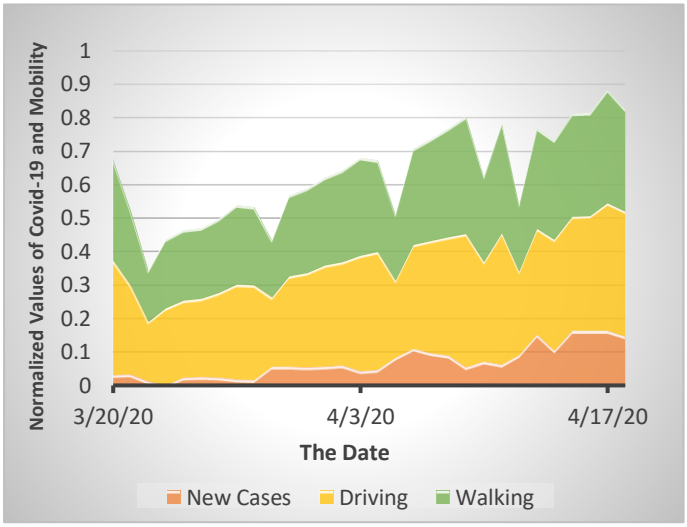


Figure (4.12) the correlation between Covid-19 New Cases and mobility through low mobility rate of (Brazil)

Table (4.3) Person's correlation coefficient of mobility with COVID-19 new cases according to **high mobility Rate**

Countries	Start Date	End Date	The number of shifting days	Minimum number of mobility		Maximum number of mobility		Correlation between New Cases& Driving	Correlation between New Cases& Walking
				Driving	Walking	Driving	Walking		
ARE	8/1/2020	10/29/2020	2 Days	83.53	61.74	343.57	292.07	0.632	0.832
ZAF	10/20/2020	12/18/2020	No shift	87.71	68.59	170.26	135.1	0.593	0.500
GBR	7/1/2020	8/29/2020	1 Day	83.53	61.74	121.53	160.66	0.800	0.765
DEU	11/15/2020	12/14/2020	2 Days	66.27	99.73	94.02	128.89	0.500	0.510
DNK	6/25/2020	7/24/2020	2 Days	159.79	138.13	213.60	224.59	0.576	0.711
BRA	7/01/2020	8/19/2020	2 Days	71.07	44.46	135.21	99.30	0.541	0.578

Table (4.4) number of infections through the Second period of the study

countries	Start Date	Number of New Cases recorded	End Date	Number of New Cases recorded
ARE	8/1/2020	400	10/29/2020	1400
ZAF	10/20/2020	1461	12/18/2020	9126
GBR	7/1/2020	729	8/29/2020	1549
DEU	11/15/2020	10824	12/14/2020	14432
DNK	6/25/2020	47	7/24/2020	51
BRA	7/01/2020	24052	8/19/2020	47784

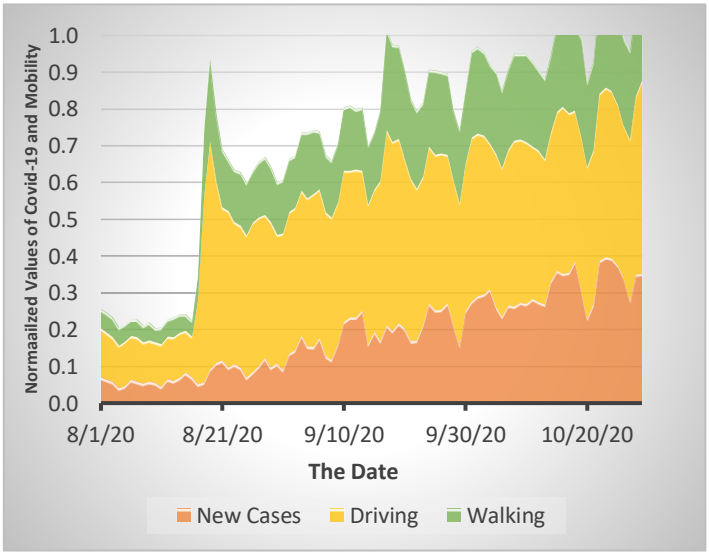


Figure (4.13) the correlation between Covid-19 New Cases and mobility through high mobility rate of (United Arab Emarat)

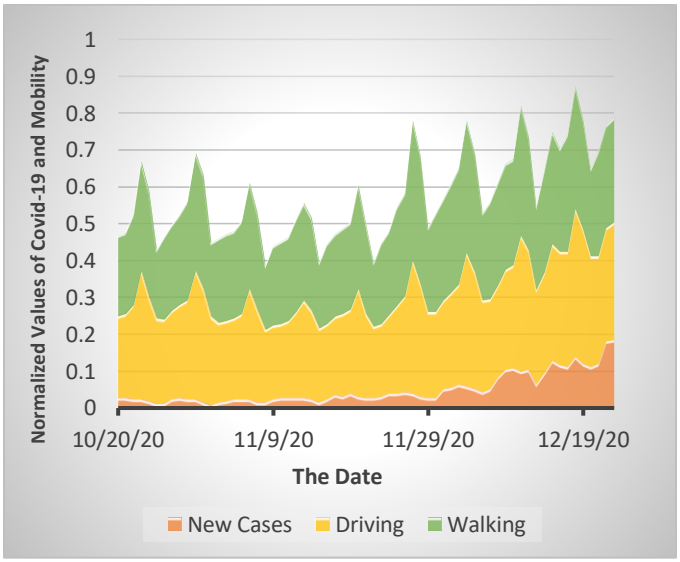


Figure (4.14) the correlation between Covid-19 New Cases and mobility through high mobility rate of (South Africa)

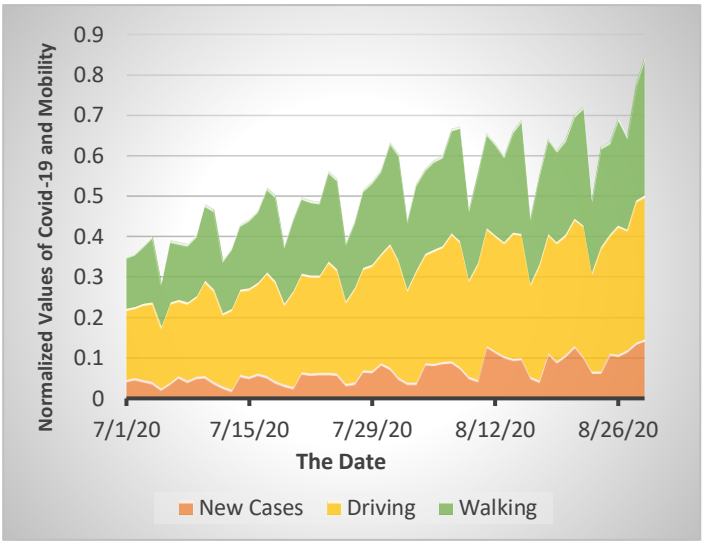


Figure (4.15) the correlation between Covid-19 New Cases and mobility through high mobility rate of (United Kingdom)

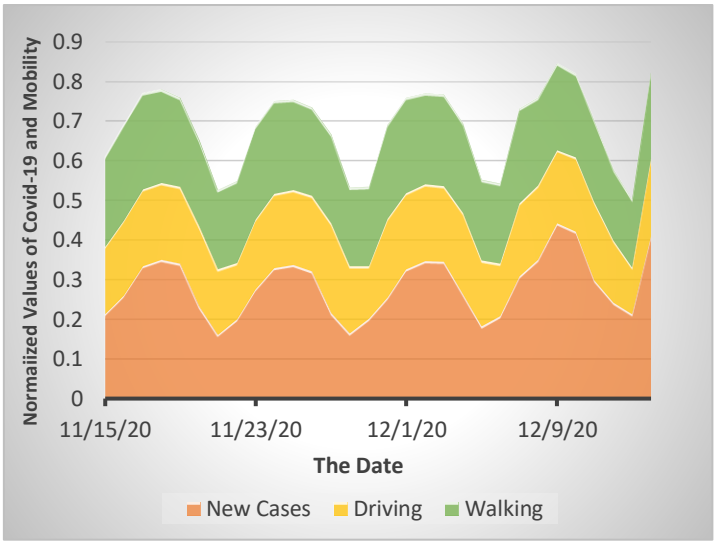


Figure (4.16) the correlation between Covid-19 New Cases and mobility through high mobility rate of (Germany)

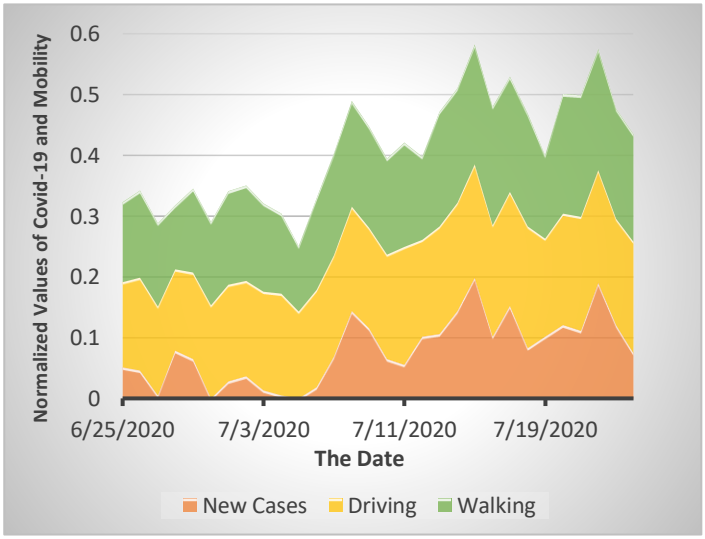


Figure (4.17) the correlation between Covid-19 New Cases and mobility through high mobility rate of (Denmark)

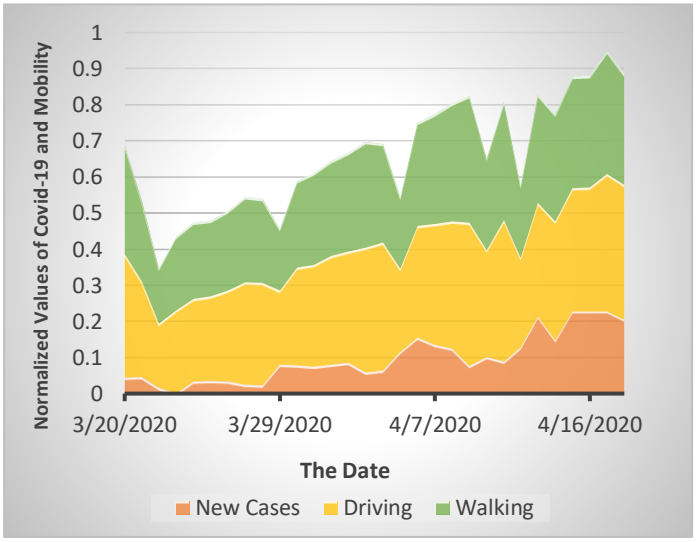


Figure (4.18) the correlation between Covid-19 New Cases and mobility through high mobility rate of (Brazil)

Table (4.5) demonstrates the Correlation Coefficient for the third period which is for the year 2021. You can see positive and negative values for the Correlation Coefficient, So, the relationship between mobility and viral transmission is not stable due to two key factors:

1- Many people gain immunity after healing from the infection.

2- The appearance of the vaccine and many people have been immunized.

From all the above experimental results and conclusions, the knowledge about the relationship patterns is gained and explained by five rules that are mentioned in chapter three, where the coverage of each rule is illustrated in Tables (4.6), (4,7), and (4,8), where in the first and second table can confirm that the relationship pattern is strong correlation whereas the third table points to another pattern of the relationship which is fluctuating correlation.

In the end, the findings reveal an explicit relationship between the movement of humans and coronavirus spread. So, as a result, as the mobility increases, the infections do so, and vice versa, and that is improved throughout this research.

Table (4.5) The Correlation Coefficient for all countries from January 2021 to the end of April 2021

Countries	Correlation Coefficient between New Cases & Driving	Correlation Coefficient between New Cases & Walking
ARE	-0.19	0.04
ZAF	-0.32	-0.55
GBR	-0.66	-0.61
DEU	0.40	0.34
DNK	-0.37	-0.34
BRA	0.04	0.01

Table (4.6) rules coverage for the first period of finding correlation

	Strong correlation		Week correlation	No correlation	Negative correlation
rules	**R1**	**R2**	**R3**	**R4**	**R5**
coverage	**0.67**	**0.33**	**0**	**0**	**0**

Table (4.7) rules coverage for the second period of finding correlation

	Strong correlation		Week correlation	No correlation	Negative correlation
rules	**R1**	**R2**	**R3**	**R4**	**R5**
coverage	**0.5**	**0.5**	**0**	**0**	**0**

Table (4.8) rules coverage for the third period of finding correlation

	Strong correlation		Week correlation	No correlation	Negative correlation
rules	**R1**	**R2**	**R3**	**R4**	**R5**
coverage	**0**	**0**	**0.17**	**0.17**	**0.66**

4.6. Results of Forecasting Covid-19 New Cases

As the researcher mentioned before, the Random Forest regressor introduced the best prediction on the Covid-19 cases that fitted to the four models after dividing the data into 80% (333 days) for training data and 20% (83 days) for testing data for all the study countries. After that, the training data of Covid-19 New Cases are re-weighted to improve the results. Table (4.9) illustrates the Re-weighting procedure for a sample of the data Table (4.10) shows a comparison between the adopted model and the other models according to the evaluation criteria used used in the study; the table also clarifies how the forecasting results improved after re-

weighting the data. in addition to tables (4.11) and (4.12) show a compression between the Actual and Random Forest Predictions values for the embedded countries according to weighted Data and figures from (4.19) to (4.25) which demonstrate the difference between actual and predicted values in forecasting Covid-19 Cases with Random Forest Regressor for all countries.

Table (4.9) Sample of the process of re-weighting the data of the United Arab Emirates country

The Date	Original New Cases	The Wight	Weighted New Cases
3/11/2020	**0**	**0.2**	**0**
3/12/2020	**11**	**0.2**	**2.2**
3/13/2020	**0**	**0.2**	**0**
. . . .	. . . .	. . . .	. . . .
9/24/2020	**1083**	**0.2**	**217**
9/25/2020	**1002**	**0.2**	**200**
9/26/2020	**1008**	**0.2**	**202**
9/27/2020	**1078**	**2**	**2156**
9/28/2020	**851**	**2**	**1702**
9/29/2020	**626**	**2**	**1252**
. . . .	. . . .	. . . .	. . . .
2/4/2021	**3977**	**2**	**7954**
2/5/2021	**3249**	**2**	**6498**
2/6/2021	**3251**	**2**	**6502**

Table (4.10) Comparison between the Evaluation Measures for (Covid-19 New Cases and the Re-Weighted New Cases) for the forecasting models

Countries	Evaluation Measures	Forecasting Models for New Cases				Forecasting Models for Re-Weighted New Cases			
		ARIMA	LSTM	LR	RF	ARIMA	LSTM	LR	RF
United Arab Emarat	RMSE	0.269	0.067	0.038	**0.026**	0.660	0.058	0.035	**0.022**
	MAE	0.237	0.050	0.025	**0.019**	0.646	0.030	0.016	**0.011**
China	RMSE	0.069	0.031	0.110	**0.012**	0.159	0.031	0.046	**0.011**
	MAE	0.065	0.023	0.048	**0.008**	0.157	0.020	0.024	**0.007**
South Africa	RMSE	0.055	0.015	0.040	**0.005**	0.099	0.010	0.027	**0.004**
	MAE	0.053	0.012	0.025	**0.003**	0.098	0.006	0.012	**0.002**
United Kingdom	RMSE	0.174	0.016	0.033	**0.009**	0.204	0.015	0.032	**0.006**
	MAE	0.168	0.010	0.016	**0.005**	0.203	0.006	0.014	**0.003**
Germany	RMSE	0.430	0.154	0.088	**0.042**	0.212	0.069	0.044	**0.019**
	MAE	0.355	0.108	0.042	**0.027**	0.174	0.057	0.026	**0.014**
Denmark	RMSE	0.055	0.039	0.034	**0.013**	0.037	0.021	0.026	**0.007**
	MAE	0.042	0.024	0.018	**0.009**	0.032	0.015	0.012	**0.005**
Brazil	RMSE	0.331	0.209	0.105	**0.062**	0.215	0.121	0.067	**0.034**
	MAE	0.296	0.156	0.078	**0.047**	0.177	0.087	0.043	**0.026**

Table (4.11) Comparison between the Actual and Random Forest Predictions values for the embedded countries in (**Asia and the Middle East, Africa and Latin America**) according to **Weighted Data**

The Date	ARE		CHN		ZAF		BRA	
	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
2/7/2021	3276	4595	31	57	3184	4849	50872	61345
2/8/2021	3093	3503	44	53	2435	2537	50630	46273
2/9/2021	2798	3011	47	49	1376	2034	26845	42217
2/10/2021	3310	3223	45	46	1742	2198	23439	32029
2/11/2021	3539	3374	21	30	3159	2832	51486	61103
2/12/2021	3525	3445	34	34	2488	2470	59602	62213
2/13/2021	3307	3329	33	36	2781	2656	54742	54300
2/14/2021	2631	2795	19	26	2382	2363	51546	50947
2/15/2021	3167	2932	21	23	1744	1990	44299	44565
2/16/2021	3123	3075	25	25	1102	1481	24759	28513
2/17/2021	3236	3148	15	20	1210	1452	32197	32326
2/18/2021	3452	3292	28	25	2320	2038	55271	54209
2/19/2021	3294	3318	20	24	2327	2267	56766	58809
2/20/2021	3140	3107	22	21	1911	1978	51879	52214
2/21/2021	3158	3005	23	22	1690	1774	51050	49091

Table (4.12) Comparison between the Actual and Random Forest Predictions values for the embedded countries in (**Europe)** according to **Weighted Data**

The Date	GBR		DEU		DNK	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
2/7/2021	15838	22204	8616	12820	403	593
2/8/2021	11946	15129	4535	6777	425	526
2/9/2021	11355	12906	3379	4877	384	528
2/10/2021	15682	14388	8072	8148	435	491
2/11/2021	13792	13881	10237	11357	455	470
2/12/2021	13476	13315	9860	10013	445	419
2/13/2021	12732	13392	8354	8506	366	355
2/14/2021	12145	12002	6114	6192	313	344
2/15/2021	8751	9928	4426	5164	331	353
2/16/2021	8752	9961	3856	5264	338	378
2/17/2021	14241	12032	7556	7531	397	484
2/18/2021	12467	13149	10207	9832	513	519
2/19/2021	11555	11554	9113	9493	497	523
2/20/2021	11391	11379	9164	9121	548	495
2/21/2021	10376	10809	7676	7805	457	453

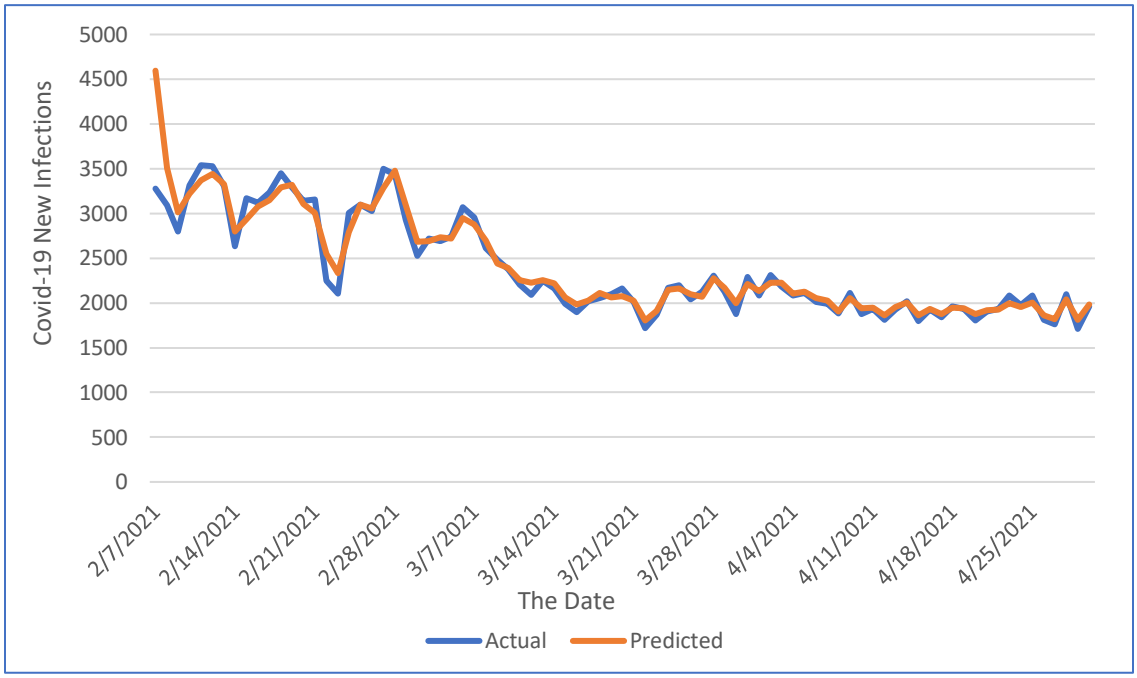


Figure (4.19) The actual and predicted value of **Covid-19 new infection for (United Arab Emarat)**

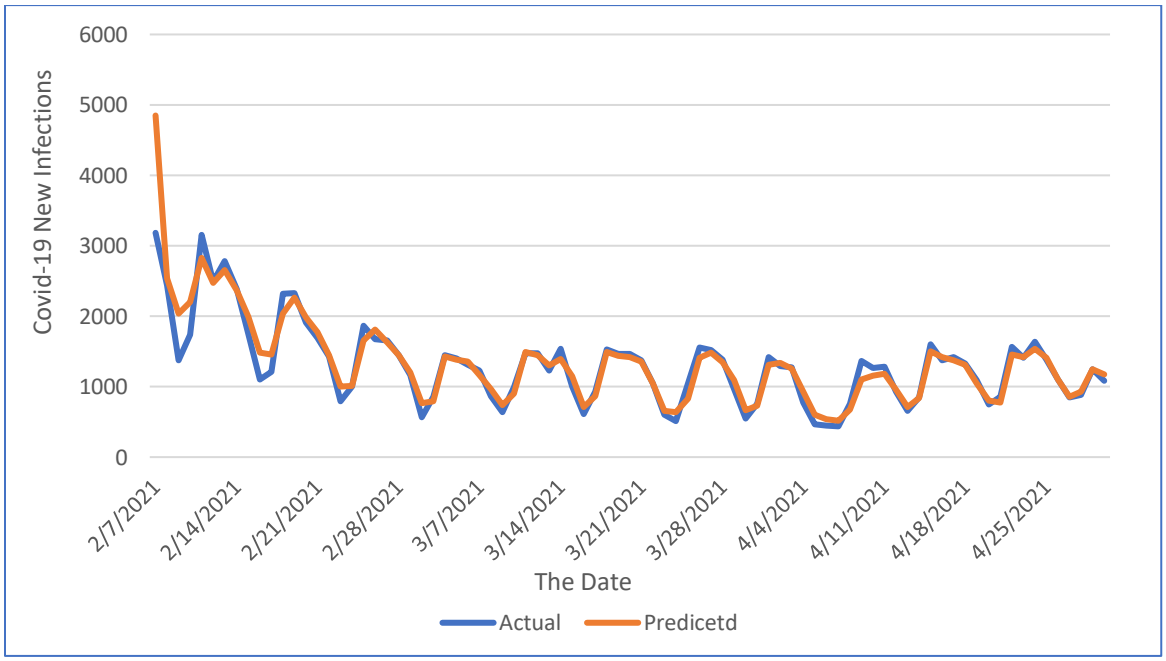


Figure (4.20) The actual and predicted value of **Covid-19 new infection for (South Africa)**

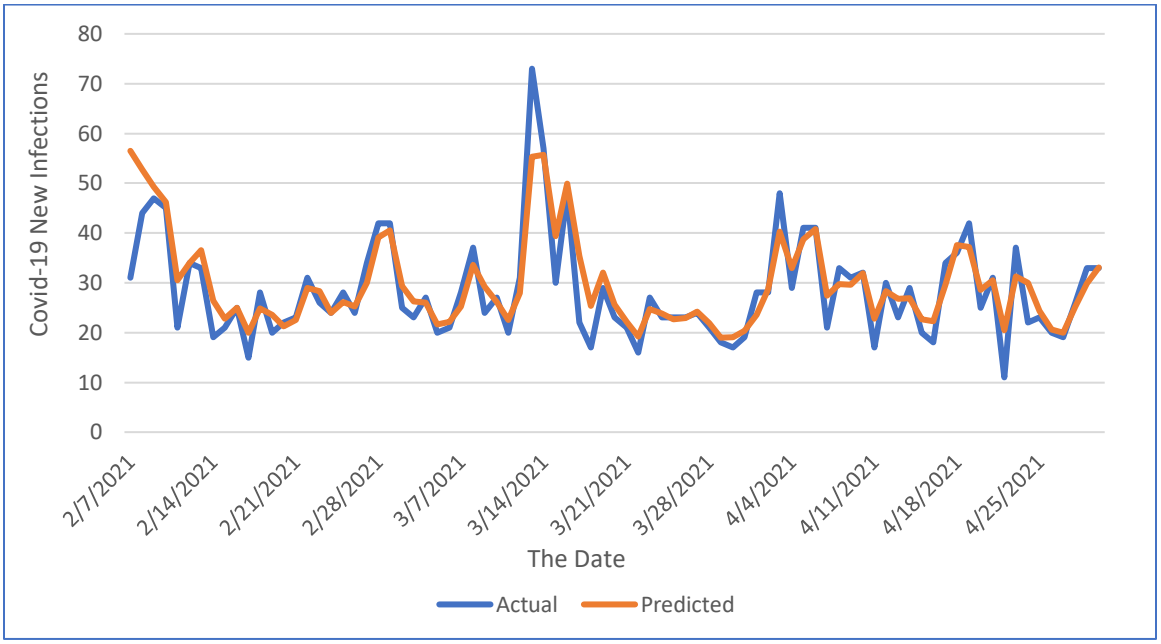


Figure (4.21) The actual and predicted values of **Covid-19 new infection for (China)**

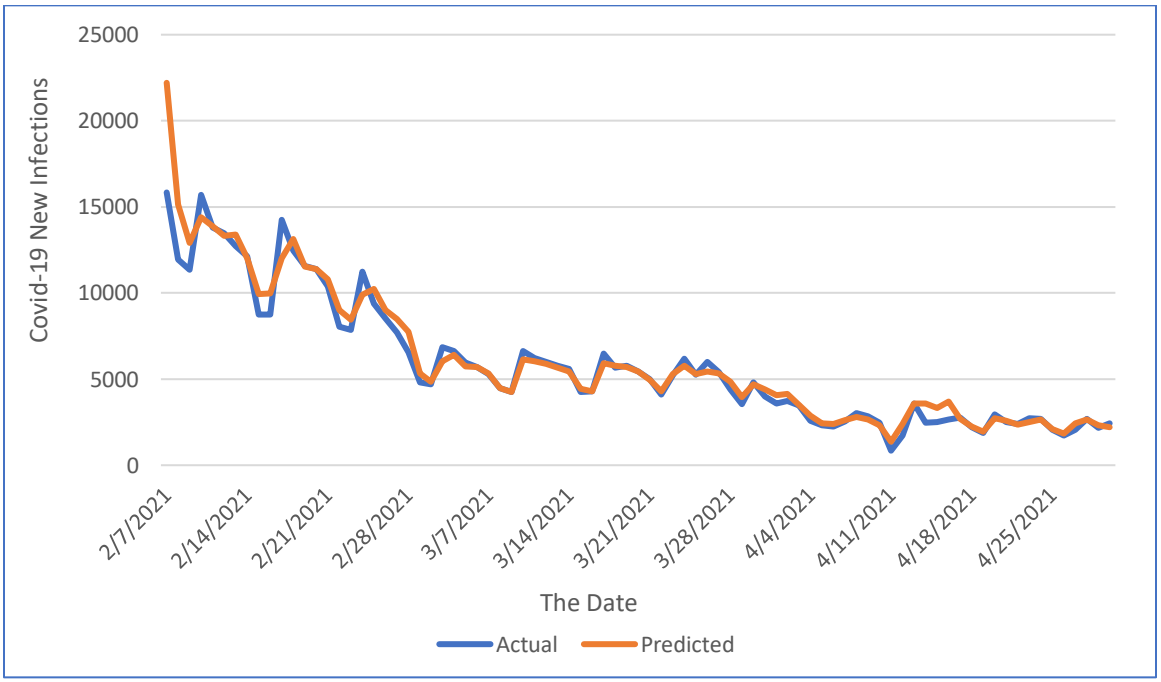


Figure (4.22) The actual and predicted values of **Covid-19 new infection for (United Kingdom)**

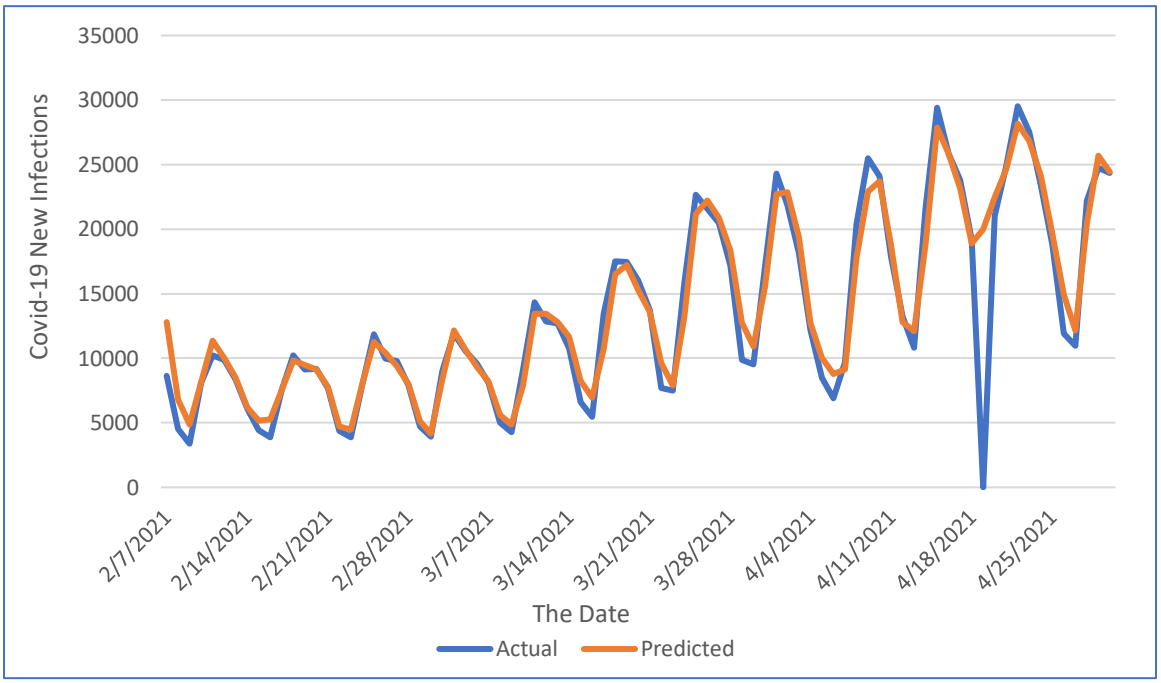


Figure (4.23) The actual and predicted values of **Covid-19 new infection for (Germany)**

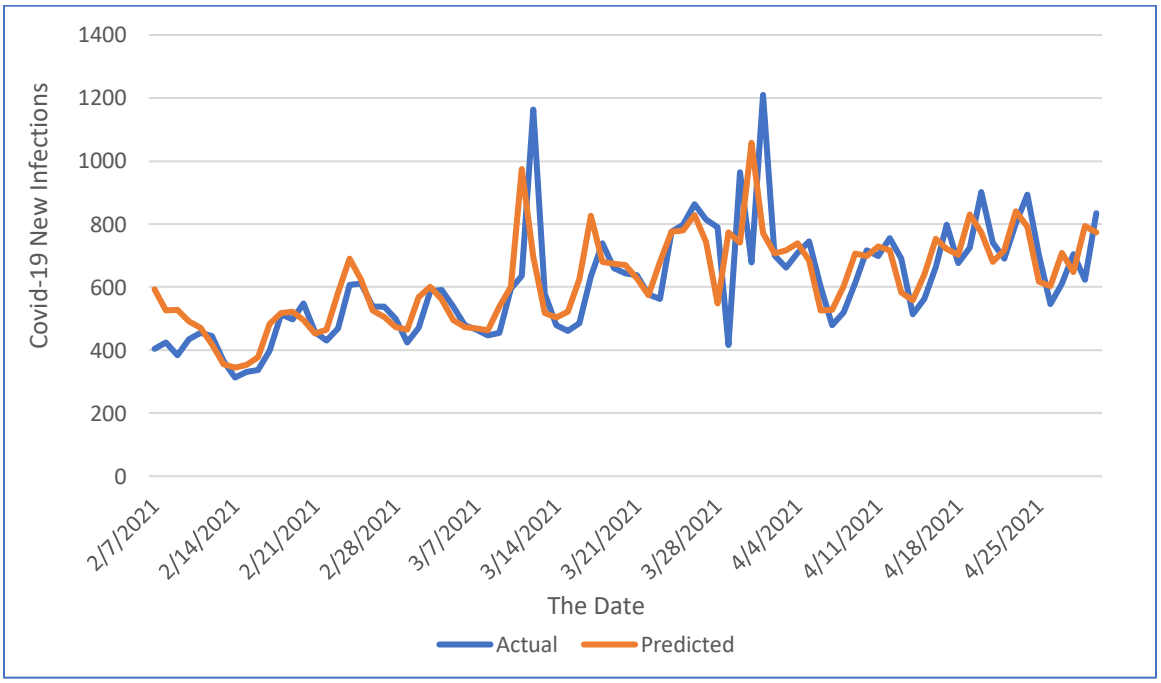


Figure (4.24) The actual and predicted values of **Covid-19 new infection for (Denmark)**

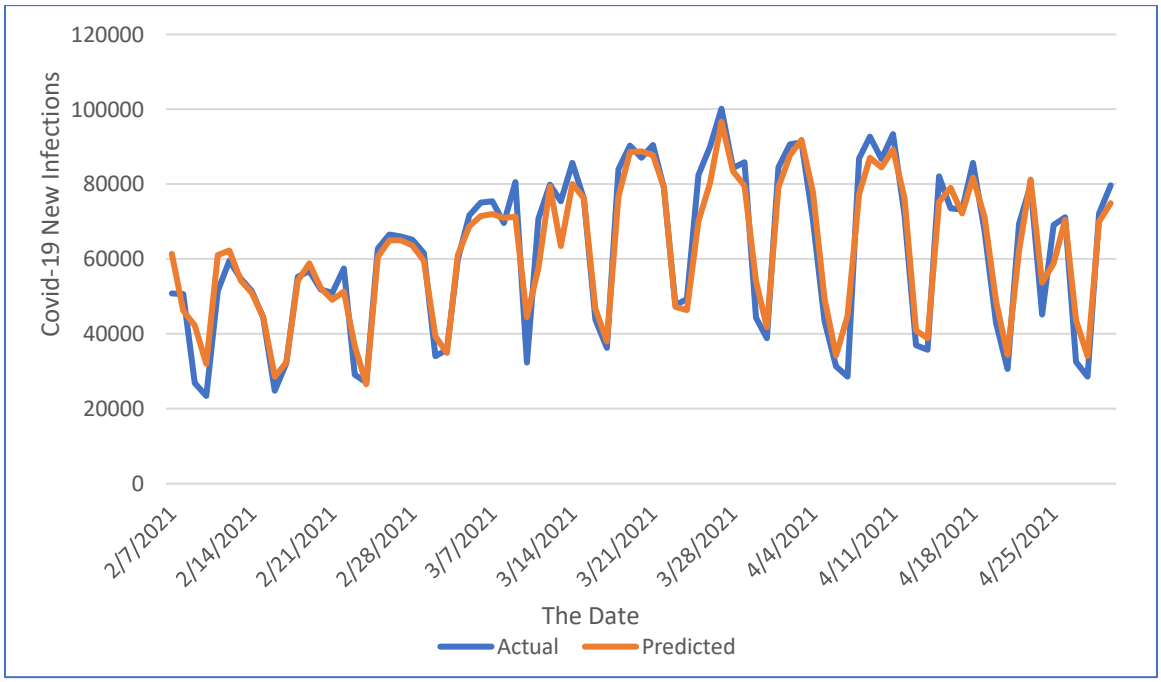


Figure (4.25) The actual and predicted values of **Covid-19 new infection for (Brazil)**

4.7. Results of Forecasting Mobility of Driving and Walking

The mobility data were fitted to the model that was mentioned earlier, and it yields predictions close enough to reality, as shown in Table (4.13) where the random forest regressor outperformed the other forecasting models. Table (4.14) (4.15) (4.16) (4.17) shows the actual and predicted values of driving and walking. Lastly, Figures (4.26) (4. 27) (4.28) (4.29) (4.30) (4.31) (4.32) (4.33) (4.34) (4.35) (4.36) (4.37) reveal that how close the predictions were to the true values of mobility for all the countries included, with the exception of China, which had no data in the Apple dataset.

Table (4.13) Comparison between the Evaluation Measures for (Mobility Trends) for the four forecasting models.

Countries	Evaluation Measures	Forecasting Models for Driving				Forecasting Models for Walking			
		ARIMA	LSTM	LR	RF	ARIMA	LSTM	LR	RF
United Arab Emarat	RMSE	0.061	0.026	0.021	**0.009**	0.053	0.028	0.015	**0.011**
	MAE	0.050	0.018	0.011	**0.007**	0.045	0.019	0.011	**0.007**
South Africa	RMSE	0.119	0.100	0.063	**0.026**	0.148	0.105	0.081	**0.035**
	MAE	0.080	0.071	0.046	**0.017**	0.120	0.079	0.061	**0.025**
United Kingdom	RMSE	0.163	0.078	0.062	**0.022**	0.147	0.085	0.072	**0.024**
	MAE	0.132	0.060	0.046	**0.015**	0.114	0.062	0.052	**0.017**
Germany	RMSE	0.129	0.043	0.053	**0.018**	0.158	0.042	0.033	**0.016**
	MAE	0.116	0.032	0.037	**0.013**	0.149	0.031	0.023	**0.011**
Denmark	RMSE	0.130	0.041	0.051	**0.016**	0.139	0.073	0.077	**0.025**
	MAE	0.113	0.027	0.032	**0.012**	0.117	0.053	0.053	**0.018**
Brazil	RMSE	0.156	0.140	0.107	**0.033**	0.337	0.166	0.122	**0.037**
	MAE	0.131	0.88	0.077	**0.020**	0.296	0.120	0.090	**0.026**

Table (4.14) Comparison between the Actual and Random Forest Predictions values for the embedded countries in (**Asia and the Middle East, Africa and Latin America**) according to **Driving**

The Date	ARE		ZAF		BRA	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
2/7/2021	325.8	331.34	87.93	88.88	91.12	96.87
2/8/2021	321.88	322.83	92.01	90.74	108.84	107.51
2/9/2021	324.38	324.88	94.19	93.98	112.74	110.65
2/10/2021	327.15	328.62	96.44	98	116.27	115.37
2/11/2021	355.33	351.02	103.06	104.09	119.55	120.18
2/12/2021	355.83	352.68	124.74	121.85	143.43	140.4
2/13/2021	348.79	348.99	107.08	108.17	146.36	146.04
2/14/2021	333.89	333.32	98.79	97.12	97.47	97.2
2/15/2021	316.48	322.79	89.73	92.08	103.15	104.85
2/16/2021	321.53	321.8	91.64	93.54	104.31	107.69
2/17/2021	325.07	326.22	94.68	95.66	108.55	109.89
2/18/2021	344.98	340.49	98.92	99.76	115.83	114.91
2/19/2021	338.12	340.7	116.62	110.98	129.36	125.23
2/20/2021	338.97	336.48	98.17	109.29	135.33	131.92
2/21/2021	323.25	327.78	86.77	87.76	92.42	105.75

Table (4.15) Comparison between the Actual and Random Forest Predictions values for the embedded countries in (**Europe**) according to **Driving** Data

The Date	GBR		DEU		DNK	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
2/7/2021	51	54.8	65.86	68.86	89.37	90.29
2/8/2021	61.28	62.7	73.68	71.99	88.53	90.33
2/9/2021	61.57	61.7	73.47	73.1	90.26	91.89
2/10/2021	63.39	63.36	71.58	71.45	91.69	92.88
2/11/2021	63.59	62.76	72.5	71.51	93.91	93.99
2/12/2021	71.89	70.41	76.37	74.6	94.57	97.35
2/13/2021	66.55	66.42	70.82	71.7	100.4	99.43
2/14/2021	56.11	58.43	74.38	73.48	97.33	98.8
2/15/2021	67.16	65.7	75.07	74.73	99.54	99.1
2/16/2021	67.38	67.9	77.55	75.84	97.92	100.23
2/17/2021	69.64	69.11	77.6	75.71	99.68	103.42
2/18/2021	69.42	68.82	79.3	76.5	105.06	105.26
2/19/2021	75.26	72.27	83.23	81.48	105.4	103.48
2/20/2021	70.37	72.63	81.2	81.13	107.04	100.8
2/21/2021	61.03	64.42	85.68	84.45	96.96	96.21

Table (4.16) Comparison between the Actual and Random Forest Predictions values for the embedded countries in (**Asia and the Middle East, Africa and Latin America**) according to **Walking** Data

The Date	ARE		ZAF		BRA	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
2/7/2021	420.95	424.31	72.54	74.73	69.35	73.25
2/8/2021	411.25	412.93	89.92	87.29	108.54	105.82
2/9/2021	410.5	412.76	92.12	92.21	113.91	112.76
2/10/2021	405.14	411.03	93.28	94.82	115.69	114.79
2/11/2021	416.61	418.22	97.52	96.93	112.73	113.57
2/12/2021	430.49	425.88	110.5	108.77	123.28	121.09
2/13/2021	426.74	427.53	99.58	105.46	119.89	118.26
2/14/2021	426.8	426.65	82.59	84.57	82.87	83.65
2/15/2021	422.08	425.79	87.58	86.71	95.01	99.65
2/16/2021	413.93	409.46	90.33	91.49	90.55	95.29
2/17/2021	404.29	404.89	95.35	94.54	104	103.71
2/18/2021	418.54	418.46	92.6	95.15	113.49	111.13
2/19/2021	418.01	419.72	102.03	90.52	120.05	118.92
2/20/2021	428.4	426.52	96.72	98.48	113.61	113.88
2/21/2021	423.74	425.61	75.03	74.93	76.57	81.13

Table (4.17) Comparison between the Actual and Random Forest Predictions values for the embedded countries in (**Europe**) according to **Walking** Data

The Date	GBR		DEU		DNK	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
2/7/2021	56.59	63.7	120.42	114.54	104.57	94.43
2/8/2021	71.5	73.26	129.01	126.07	88.91	92.01
2/9/2021	72.08	72.66	127.83	125.82	92.16	94.3
2/10/2021	74.74	73.56	127.55	125.56	94.45	98.84
2/11/2021	74.09	72.96	122.15	122.93	98.69	98.1
2/12/2021	85.8	80.03	118.71	119.35	97.49	105.87
2/13/2021	80.71	81.44	113.68	115.77	110.89	114.25
2/14/2021	63.11	65.59	119.99	117.78	114.71	112.13
2/15/2021	79.59	77.54	130.42	126.95	112.16	111.18
2/16/2021	79.27	78.86	129.15	129.92	103.94	112.12
2/17/2021	81.44	79.47	126.83	126.53	108.19	111.44
2/18/2021	80.87	79.09	125.37	125.18	111.41	112.07
2/19/2021	87.26	82.54	124.56	125.17	112.3	115.64
2/20/2021	85.12	84.28	122.76	124.66	120.4	121.33
2/21/2021	68.4	74.91	135.19	128.87	116.56	115.51

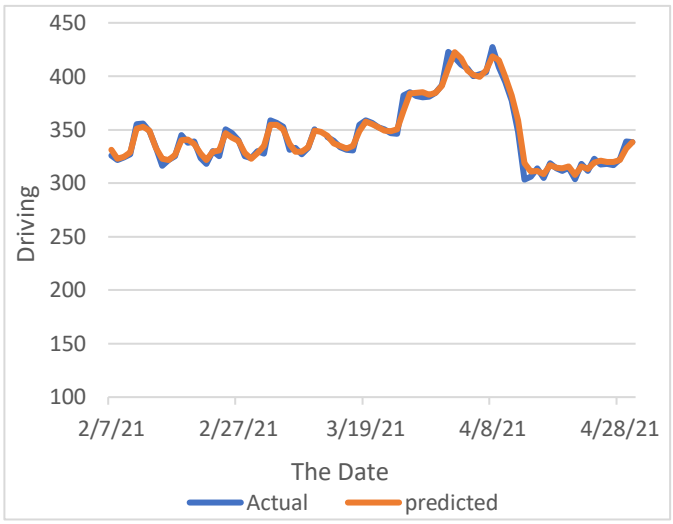


Figure (4.26) Driving forecasting (United Arab Emarat)

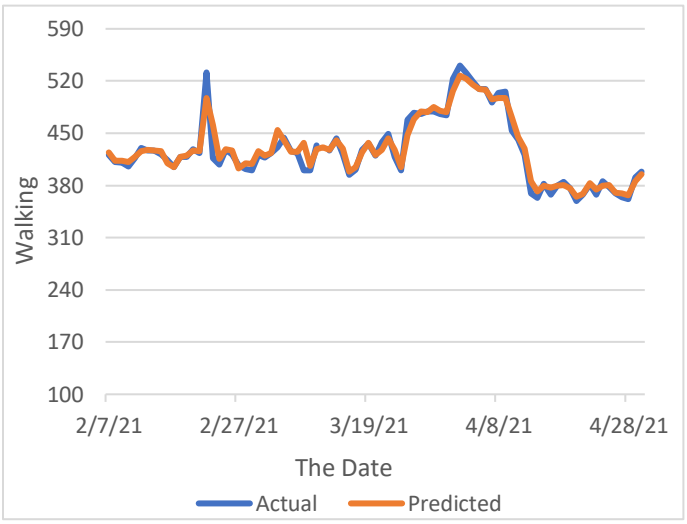


Figure (4.27) Walking forecasting (United Arab Emarat)

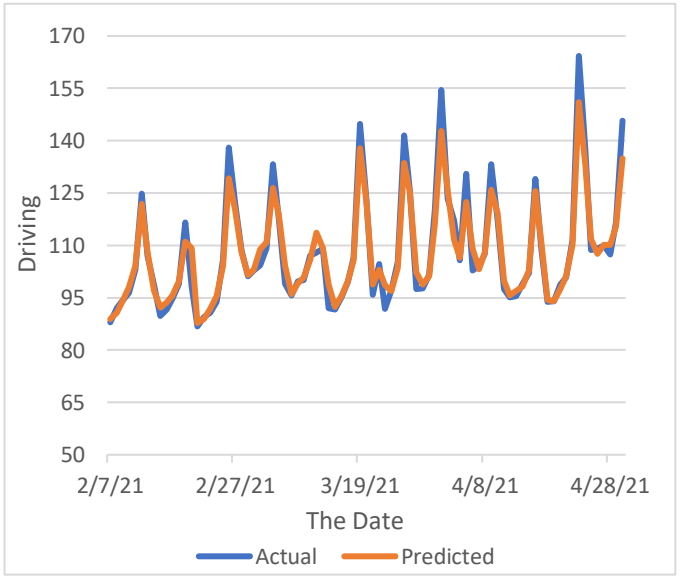


Figure (4.28) Driving forecasting (South Africa)

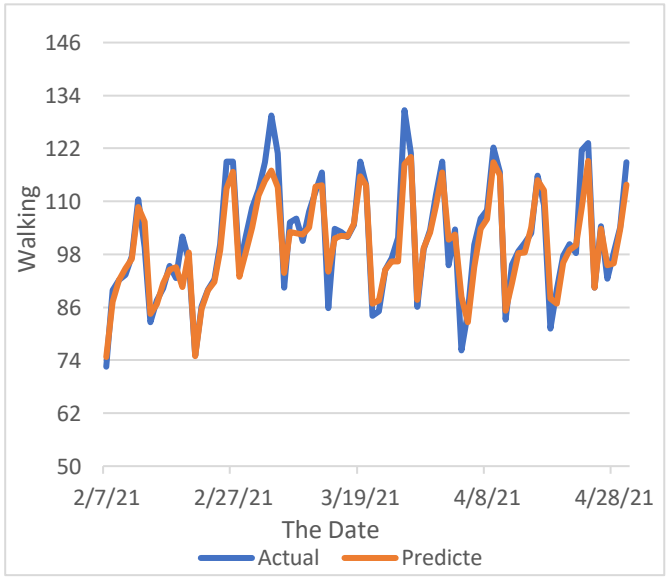


Figure (4.29) Walking forecasting (South Africa)

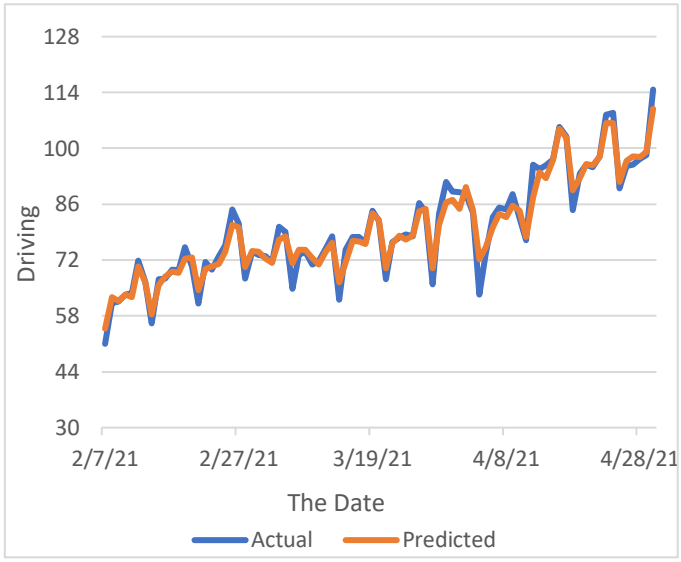


Figure (4.30) Driving forecasting (United Kingdom)

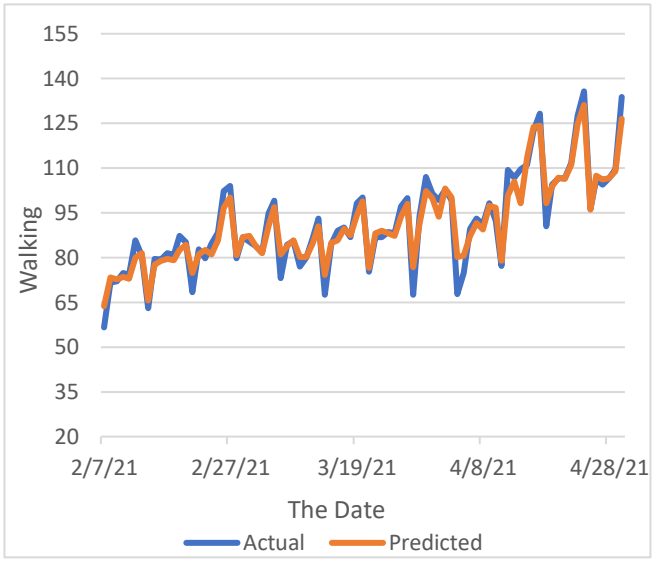


Figure (4.31) Walking forecasting (United Kingdom)

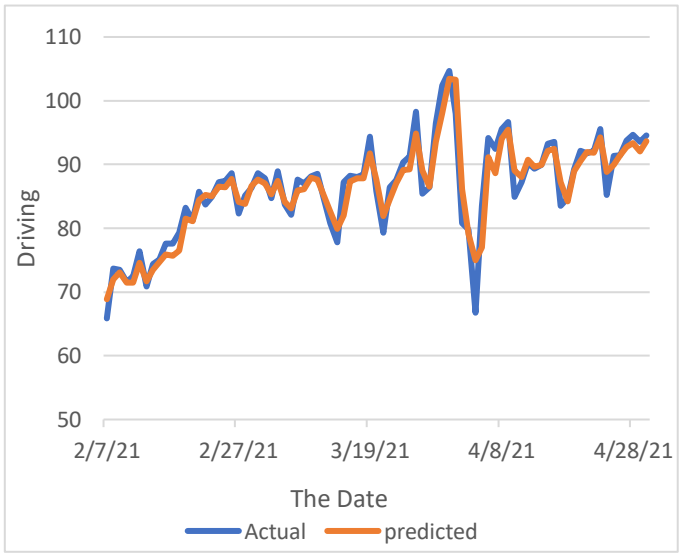


Figure (4.32) Driving forecasting (Germany)

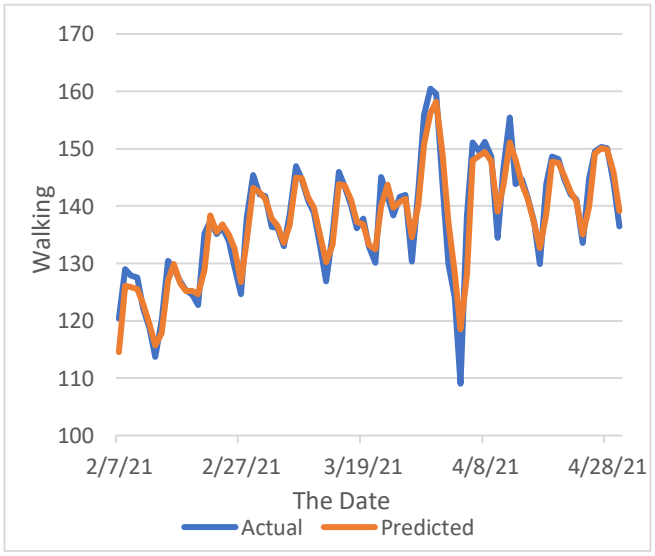


Figure (4.33) Walking forecasting (Germany)

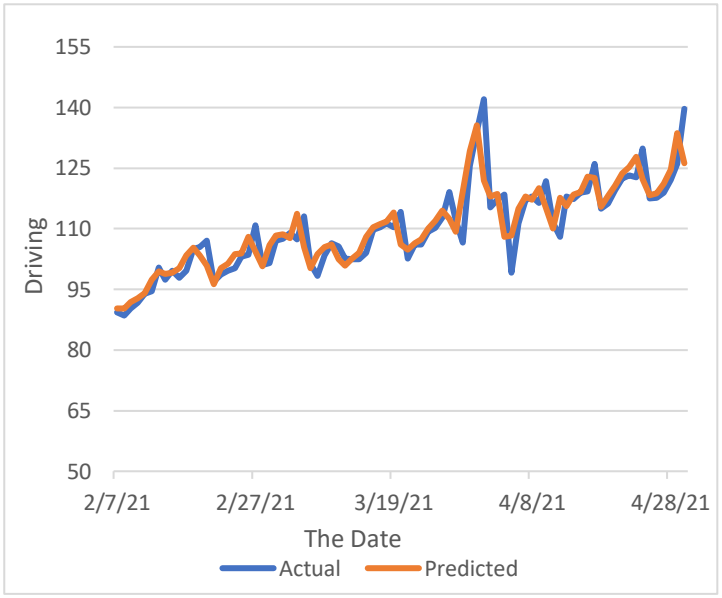


Figure (4.34) Driving forecasting (Denmark)

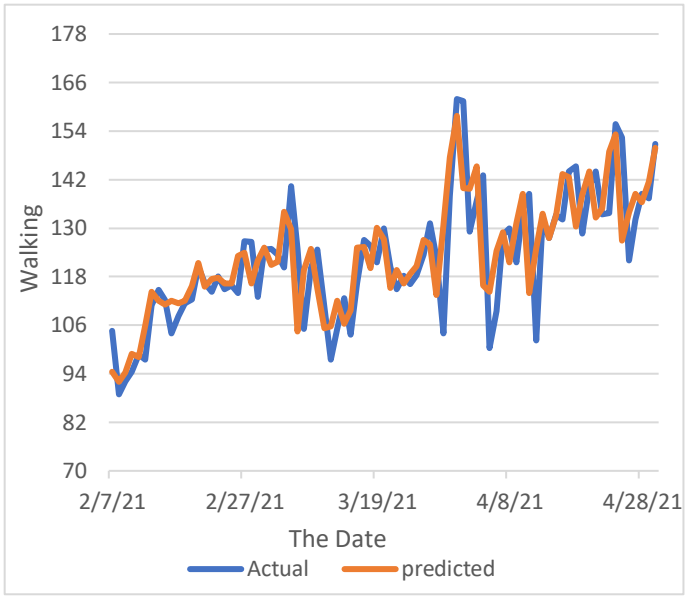


Figure (4.35) Walking forecasting (Denmark)

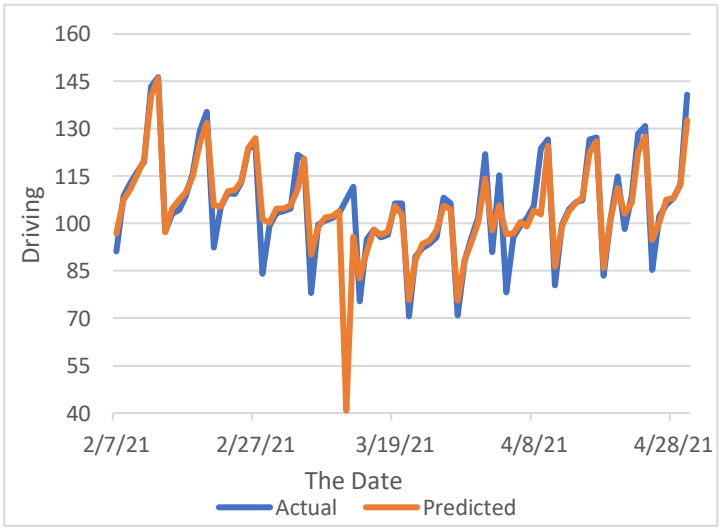


Figure (4.36) Driving forecasting (Brazil)

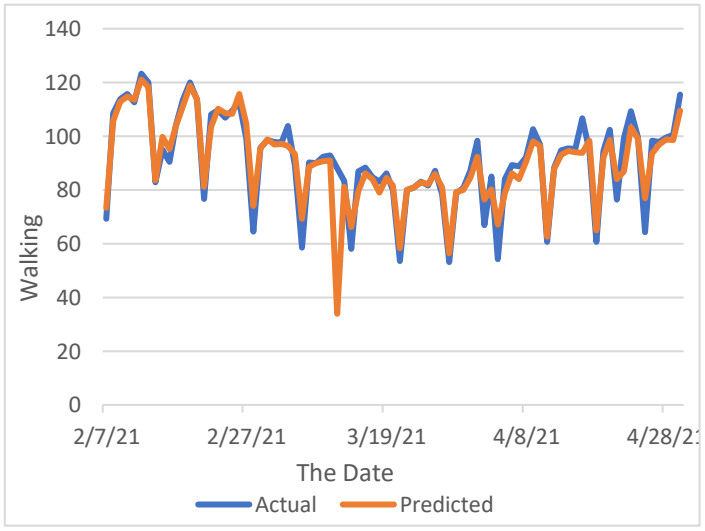


Figure (4.37) Walking forecasting (Brazil)

4.8. Evaluating the Forecasting Model

RMSE and MAE measurements were used to evaluate the Random Forest forecasting model, the result was compared to other three regression predictors (Autoregressive integrated moving average, long short-term memory and, linear regression) that were evaluated using the same metrics, All the models are tested according to the supplied test data and compared in order to check the difference. Figures from (4.38) to (4.45) show the bar charts of RMSE and MAE for Covid-19 data as well as for the two types of mobility which are driving and walking for all the study countries.

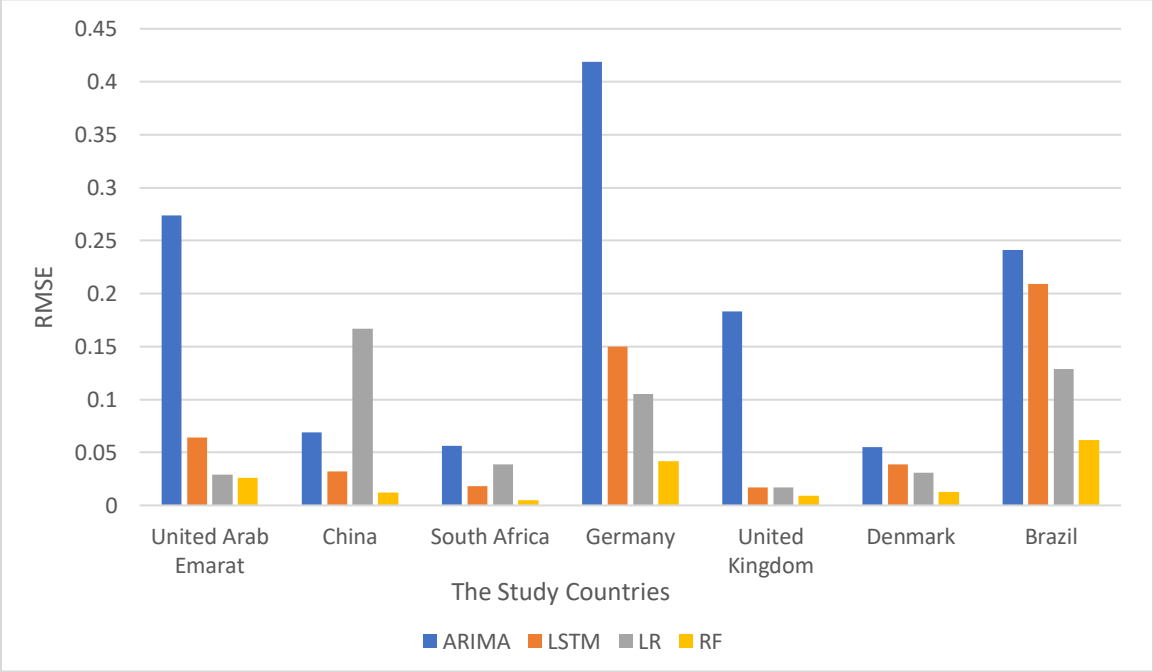


Figure (4.38) RMSE for prediction of Covid-19 original data

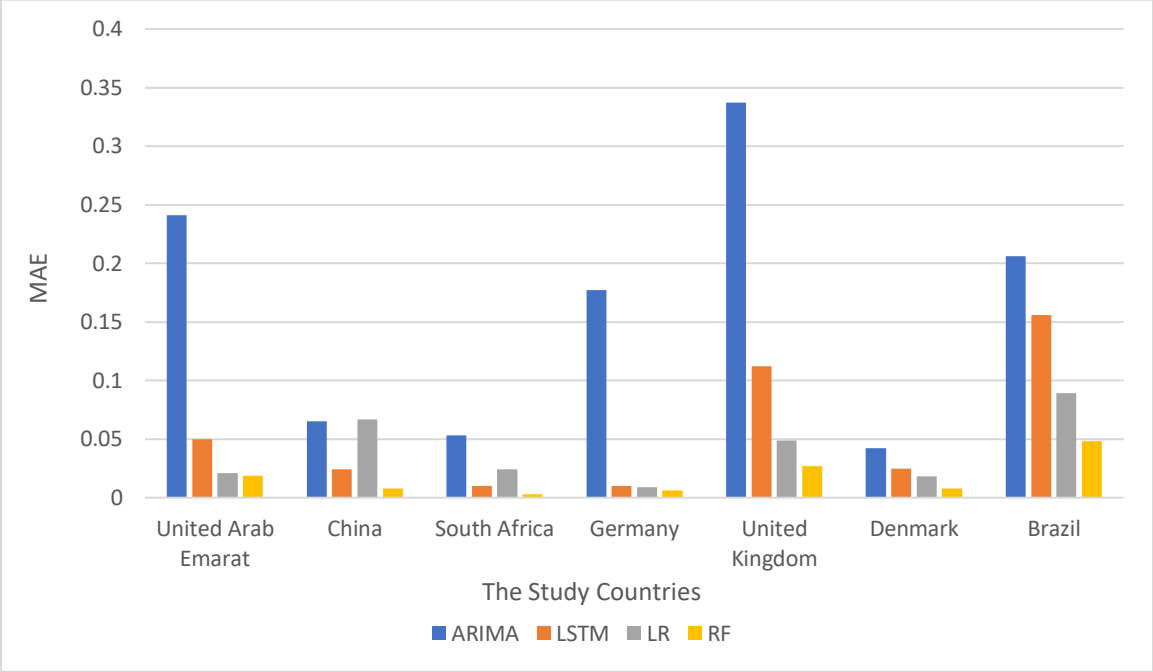


Figure (4.39) MAE for prediction of Covid-19 original data

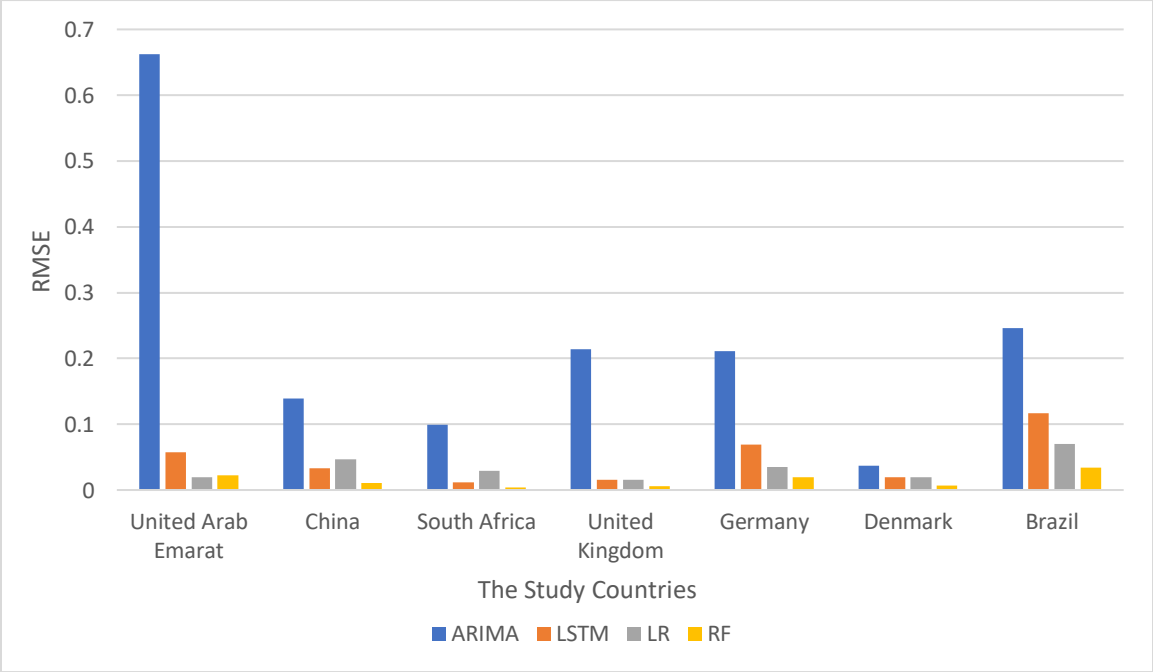


Figure (4.40) RMSE for prediction of Covid-19 weighted data

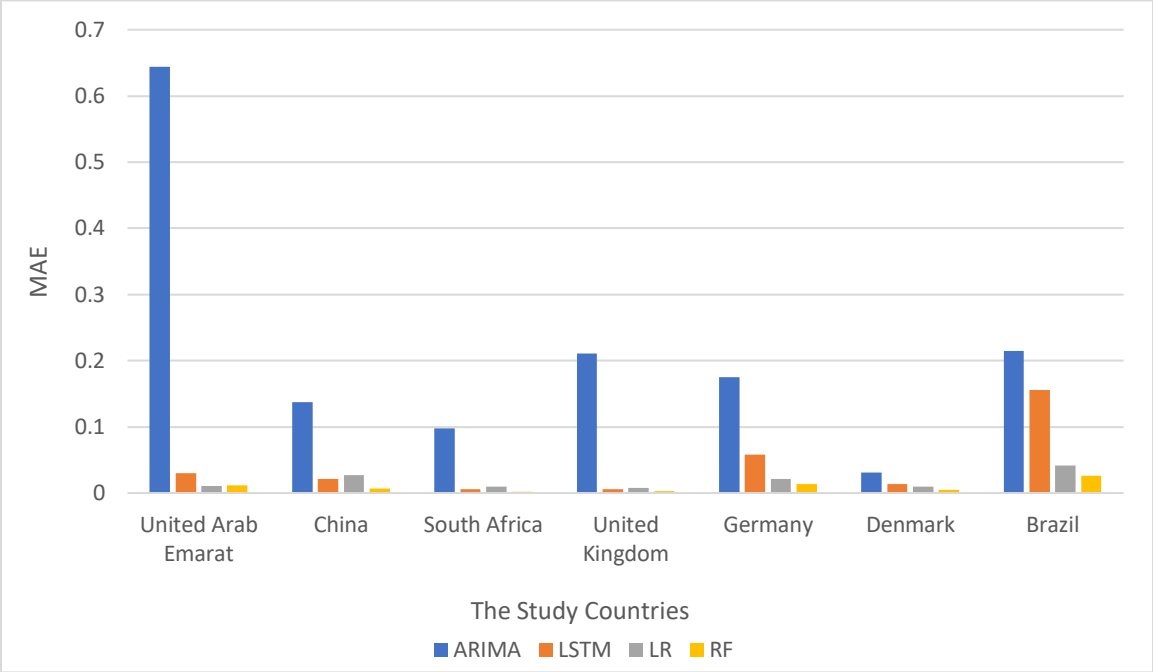


Figure (4.41) MAE for prediction of Covid-19 weighted data

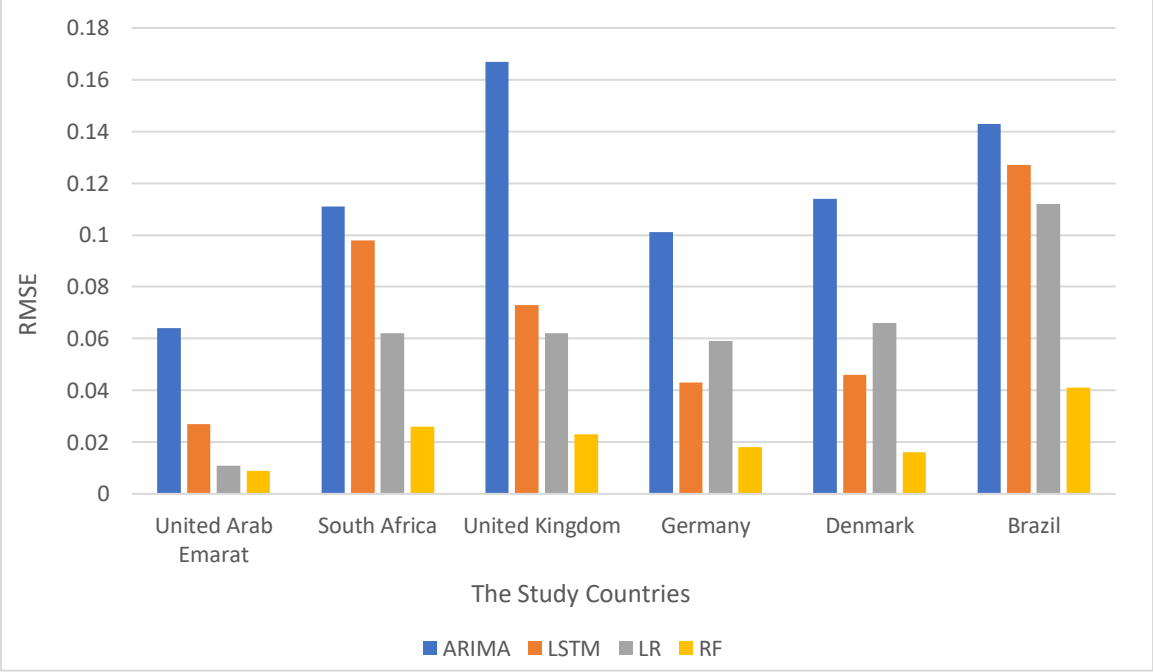


Figure (4. 42) RMSE for prediction of driving

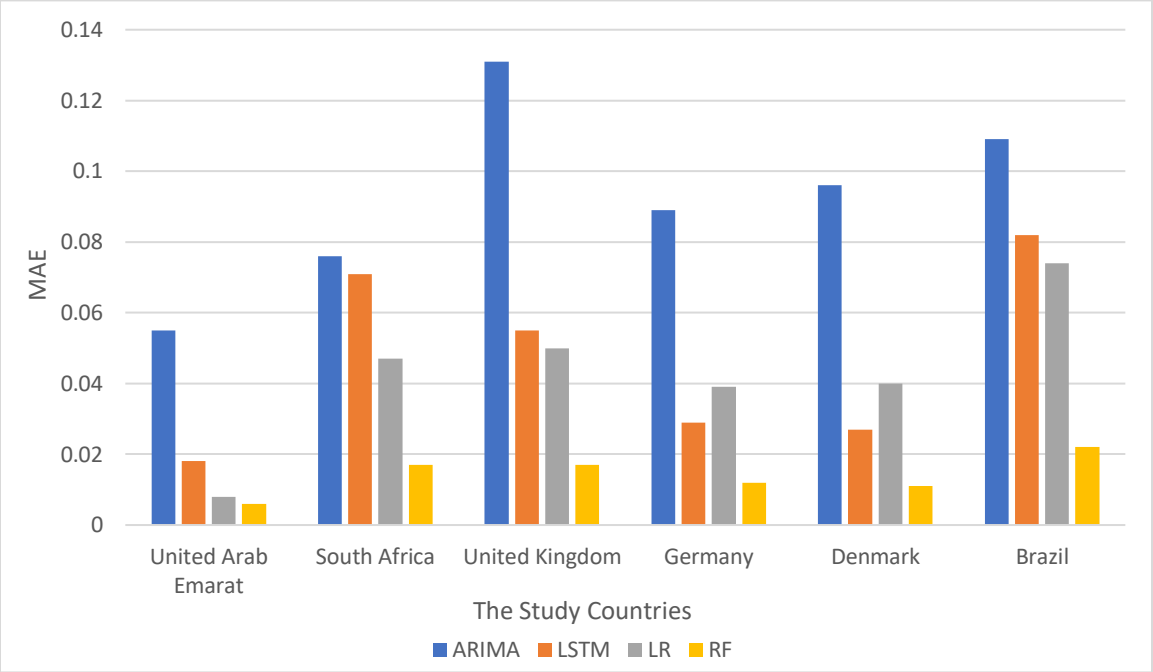


Figure (4.43) MAE for prediction of driving

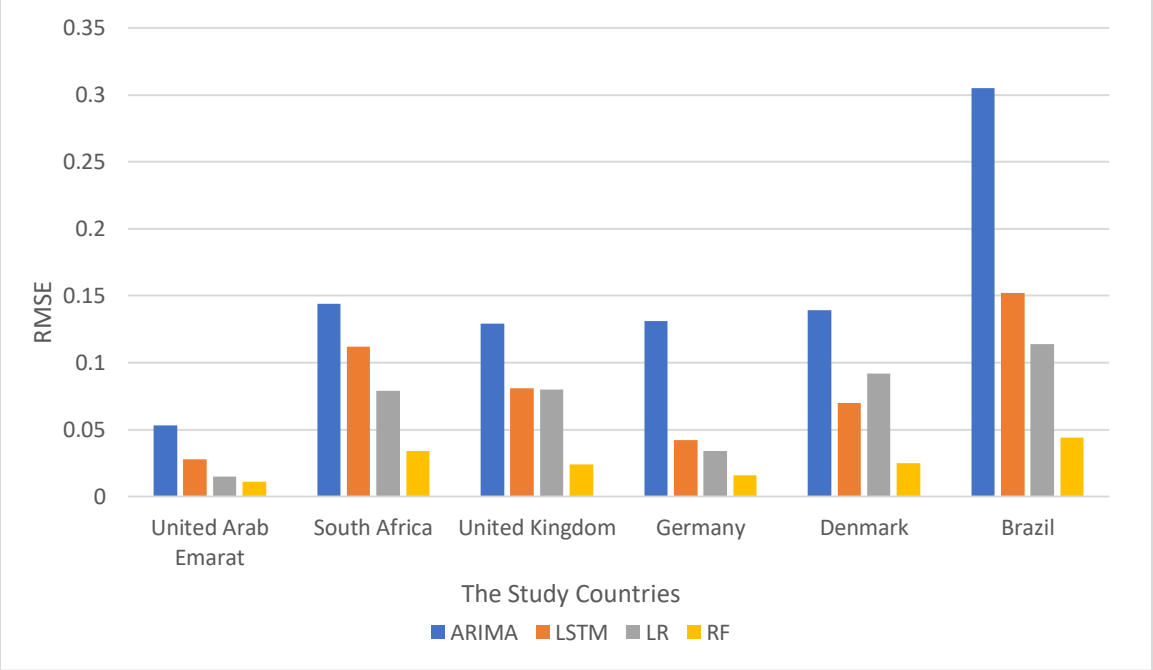


Figure (4.44) RMSE for prediction of walking

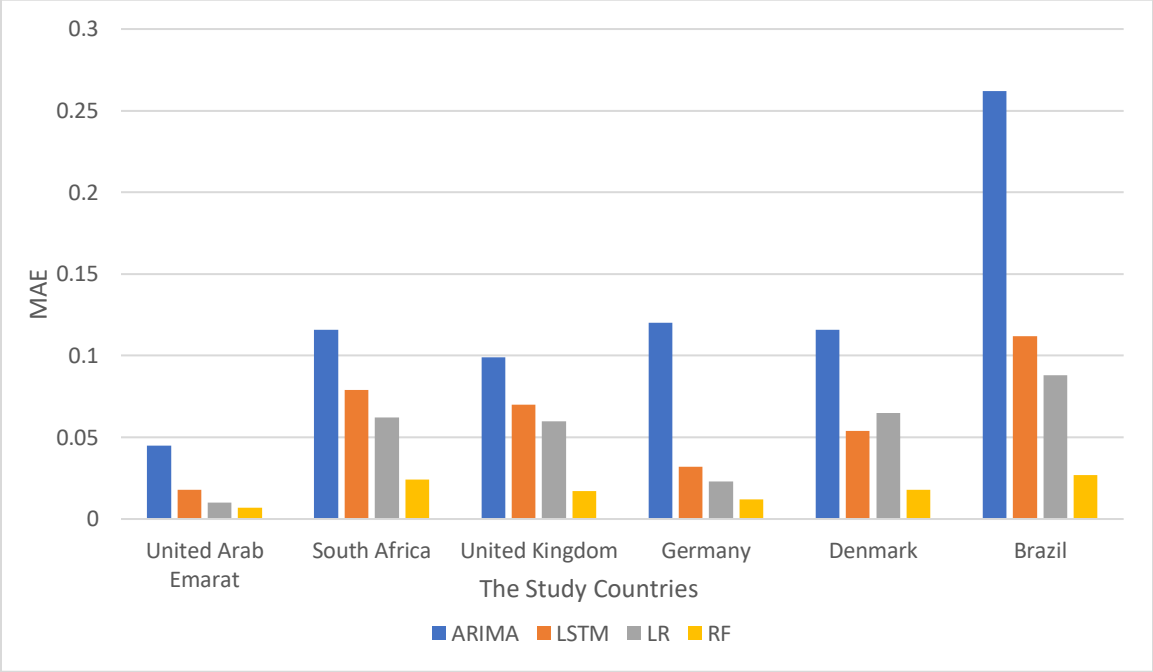
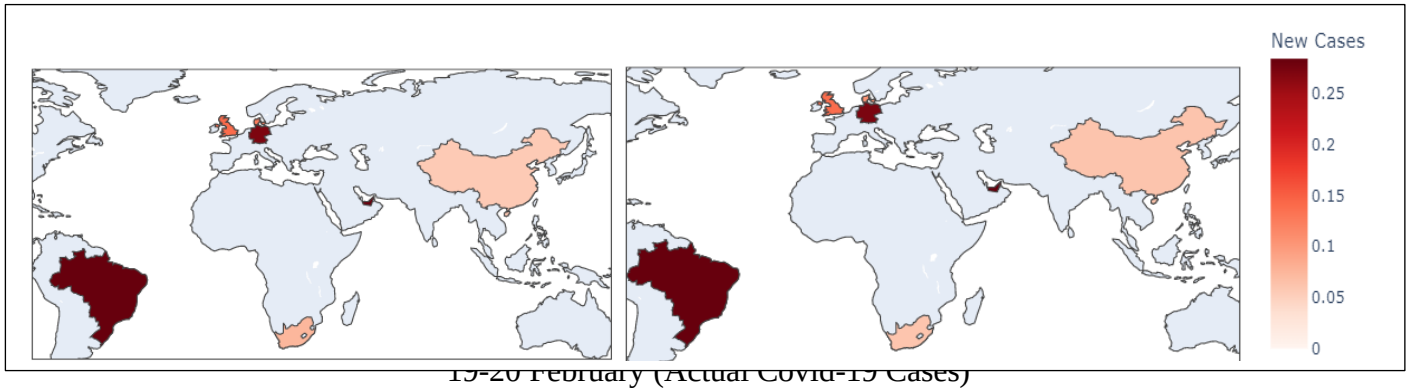


Figure (4.45) MAE for prediction of walking

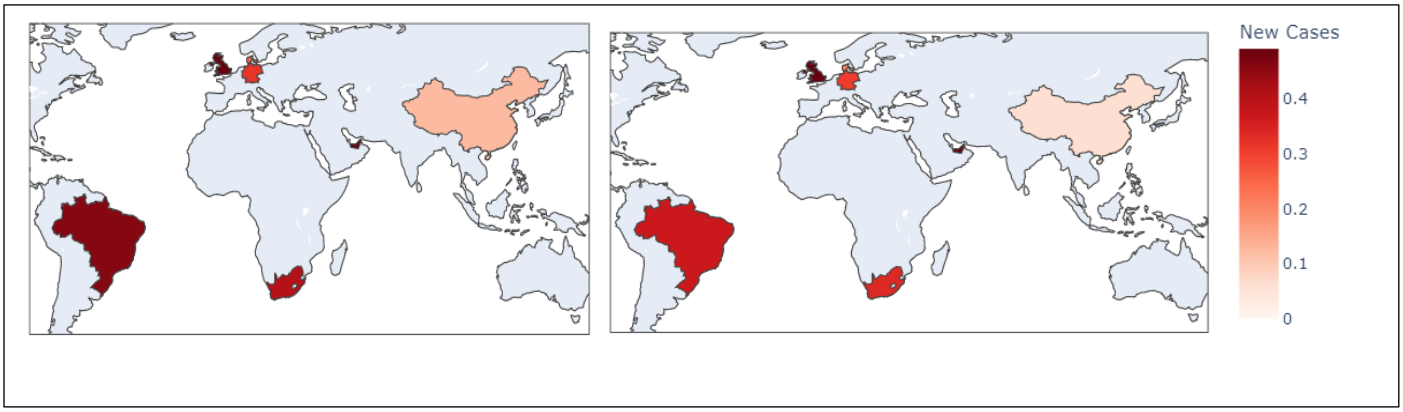
4.9. Visualization Phase

The goal of the dynamic choropleth map in this study was to create a map for surveillance of the spread of coronavirus disease on the ground as well as human movements by means of driving and walking according to time series**,** which allows us to compare the number of cases by region over time. This makes the visualization more insightful and compelling. This stage has employed the results from the previous stage (prediction) for COVID-19 cases, driving, and walking) in addition to visualize the original data. Two types of maps were created at this stage. The first map comes with a drop list to choose from (Covid-19 cases, driving, or walking), so you can easily find the correlation between the mobility and COVID-19 spread, especially with the periods of time that were analyzed earlier in this chapter. The second map's purpose was to visualize the predicted values. It also comes with a drop-down list to choose what you want to visualize (Covid-19 new infections, driving or walking). You can simply find out how close the prediction was to the

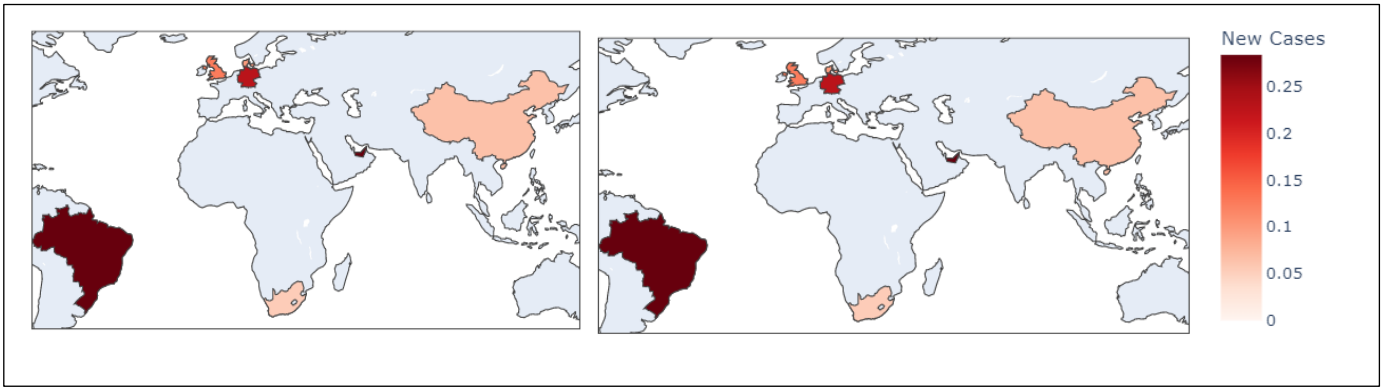
fact by comparing the two maps with the same dates. A world-wide map was introduced to visualize the seven countries involved in the study clearly. See Figures from (4.46) to (4.50).



19-20 February (Actual Covid-19 Cases)

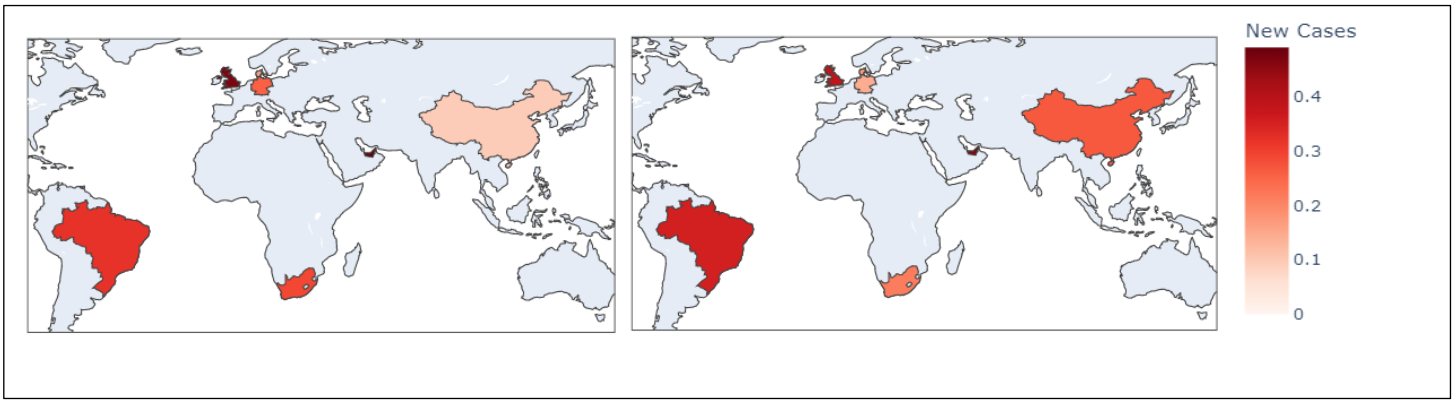


19-20 February (Predicted Covid-19 Cases)



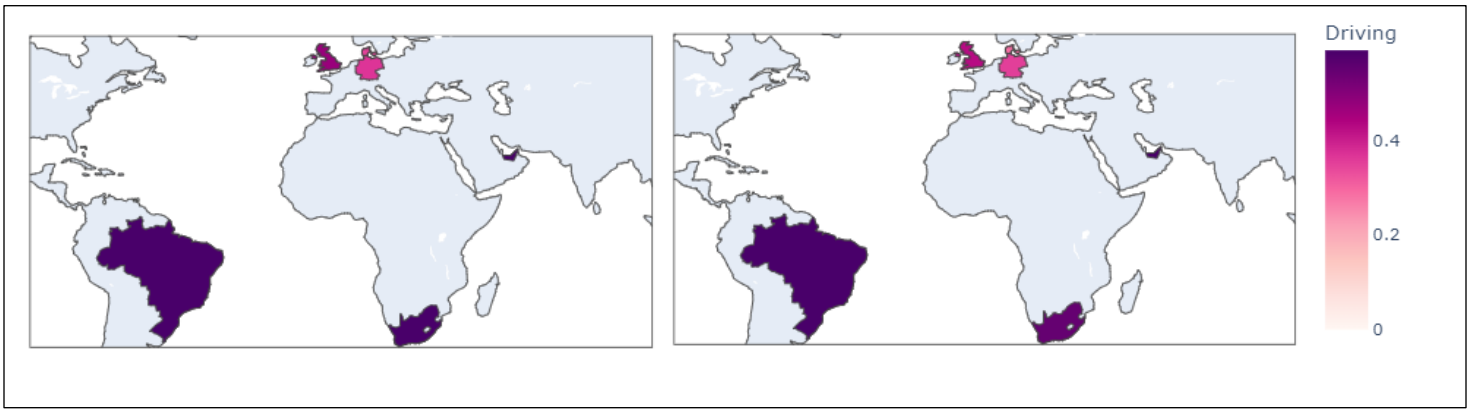
21-22 February (Actual Covid-19 Cases)

Figure (4.46) Continued

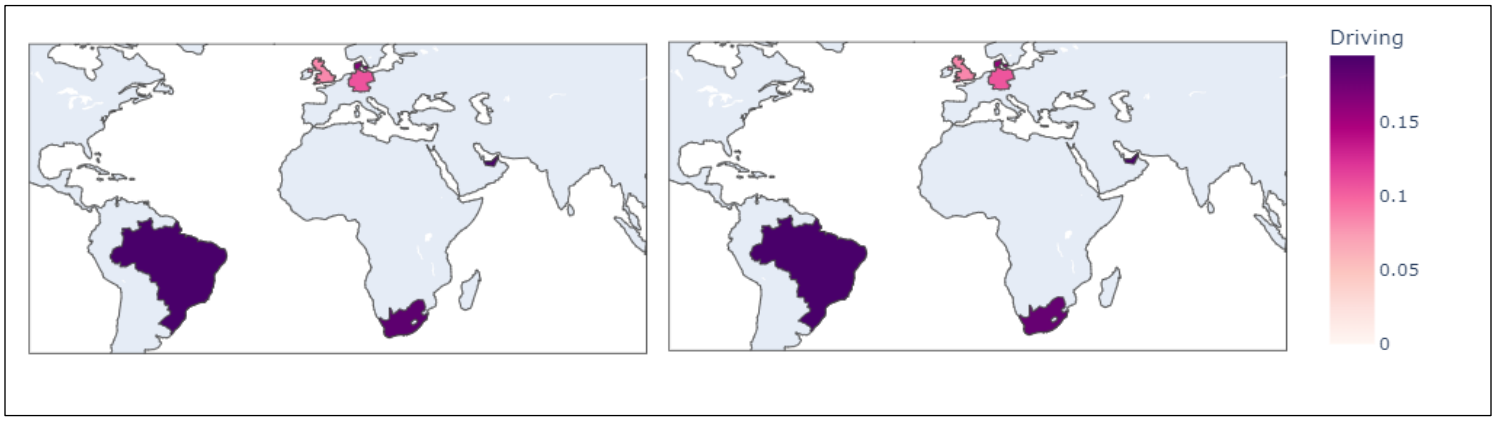


21-22 February (Predicted Covid-19 Cases)

Figure (4.46) Sample of Covid-19 actual and predicted new cases

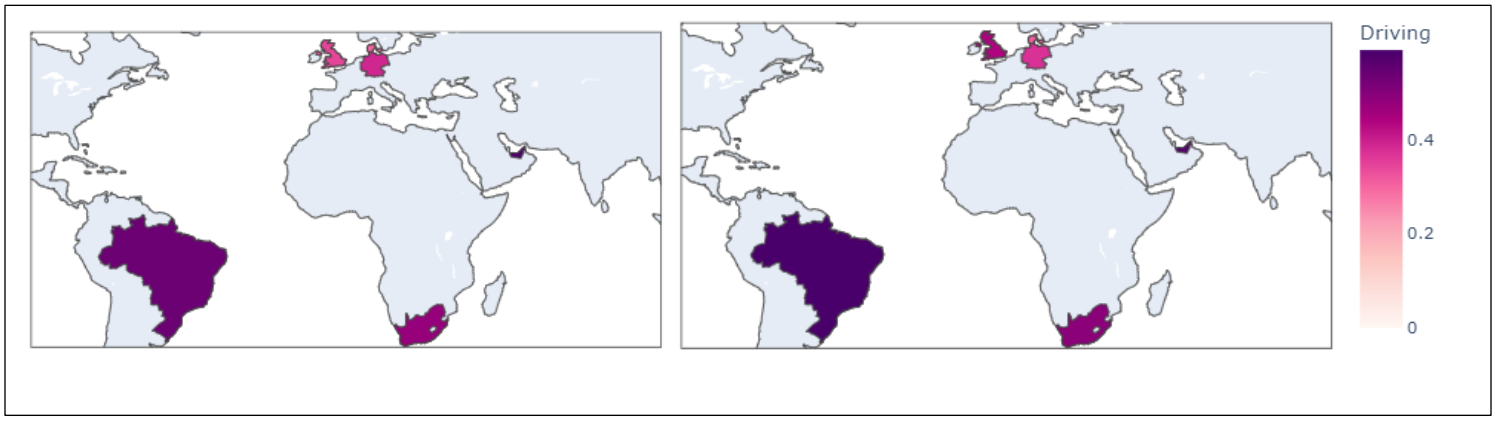


19-20 February (Actual Driving)



19-20 February (Predicted Driving)

Figure (4.47) Continued



21-22 February (Actual Driving)

21-22 February (Predicted Driving)

Figure (4.47) Sample of (Actual and predicted mobility 'Driving')

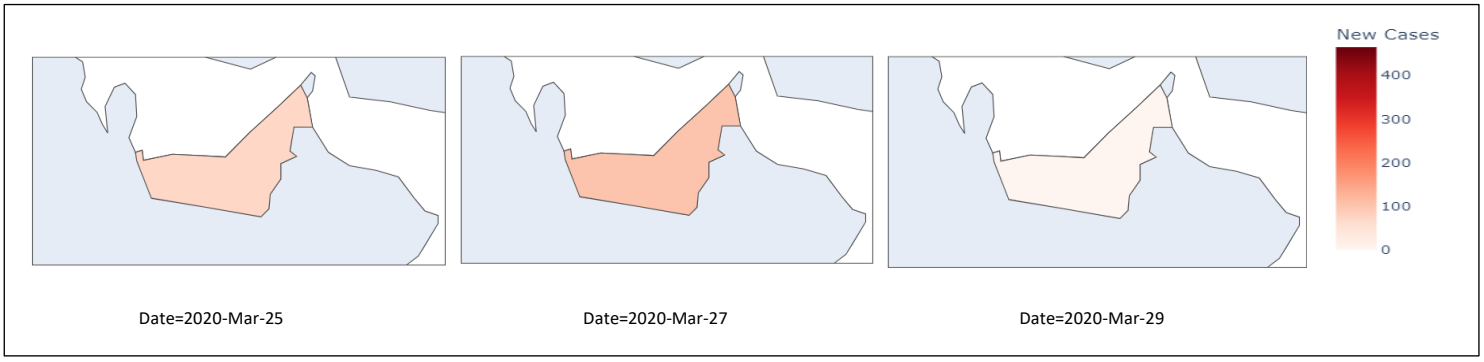
19-20 February (Actual Walking)

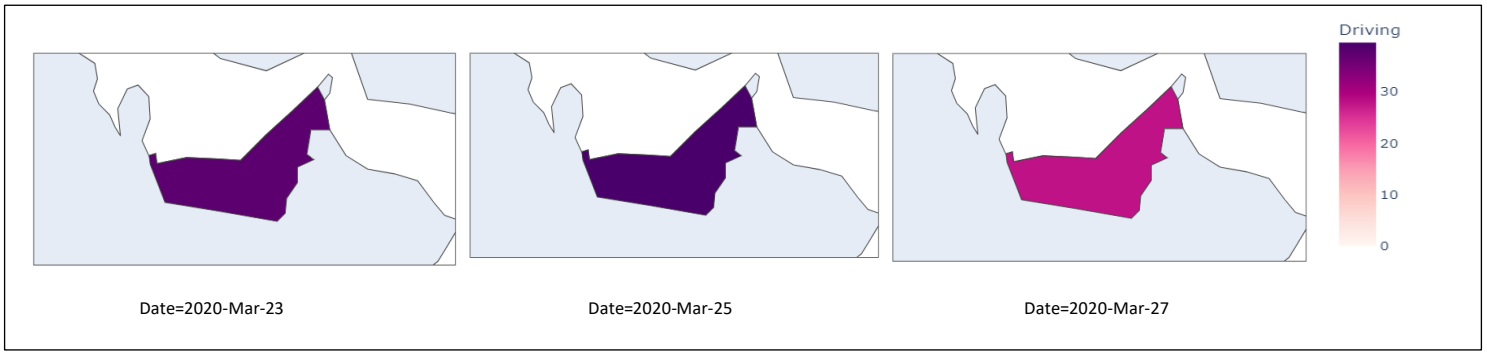
Figure (4.48) Continued

21-22 February 2021 (Actual Walking)

21-22 February 2021 (Predicted Walking)

Figure (4.48) Sample of (Actual and predicted mobility 'Walking')





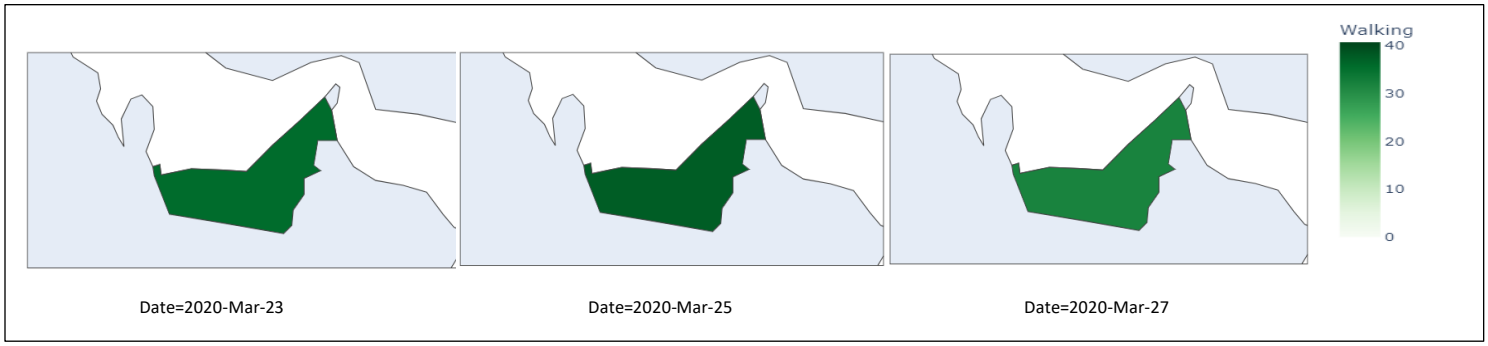


Figure (4.49) Continued

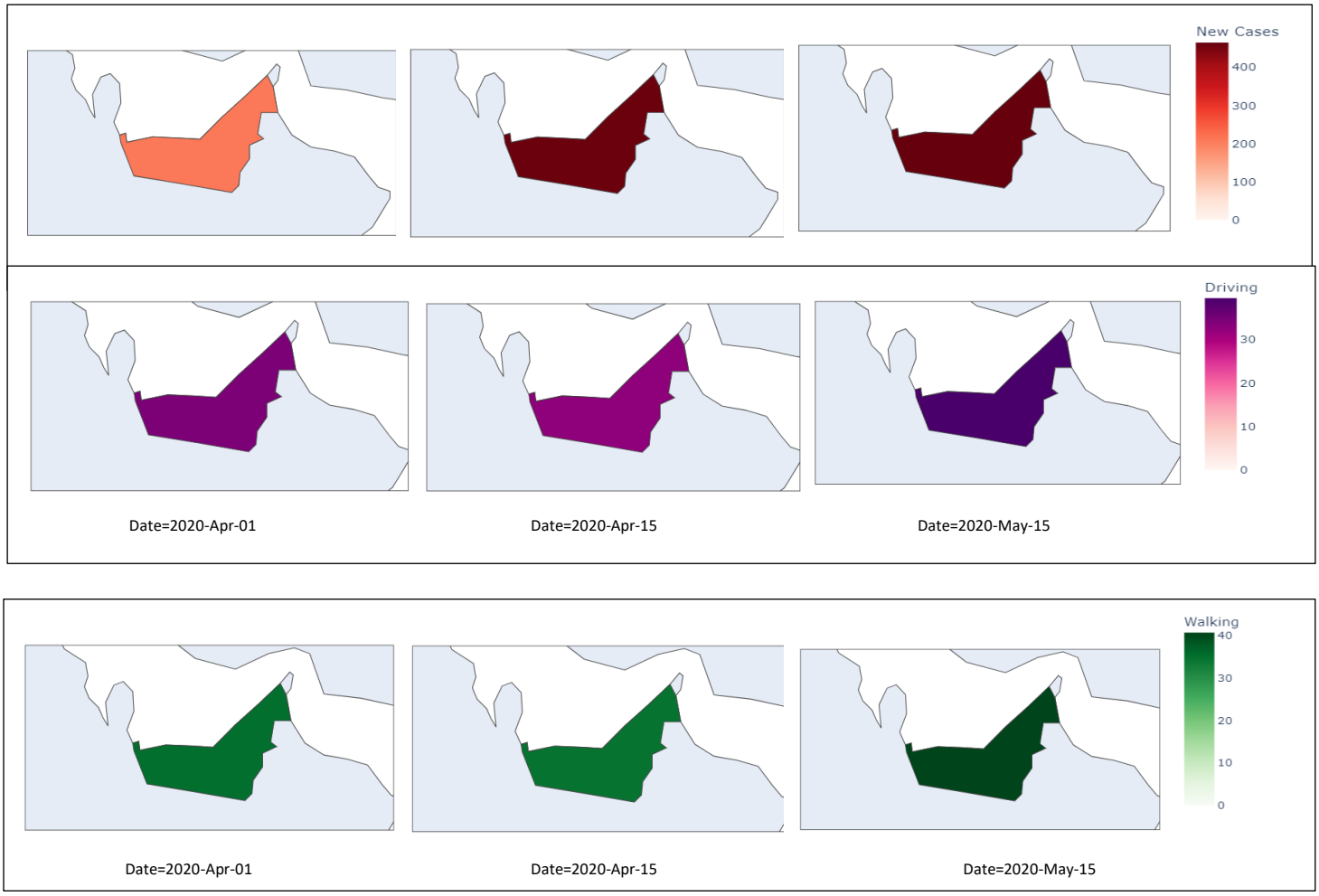
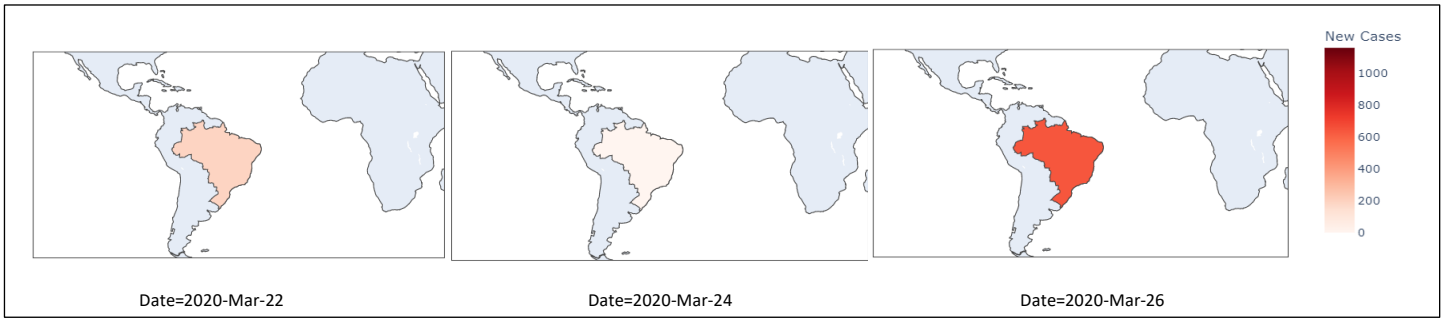
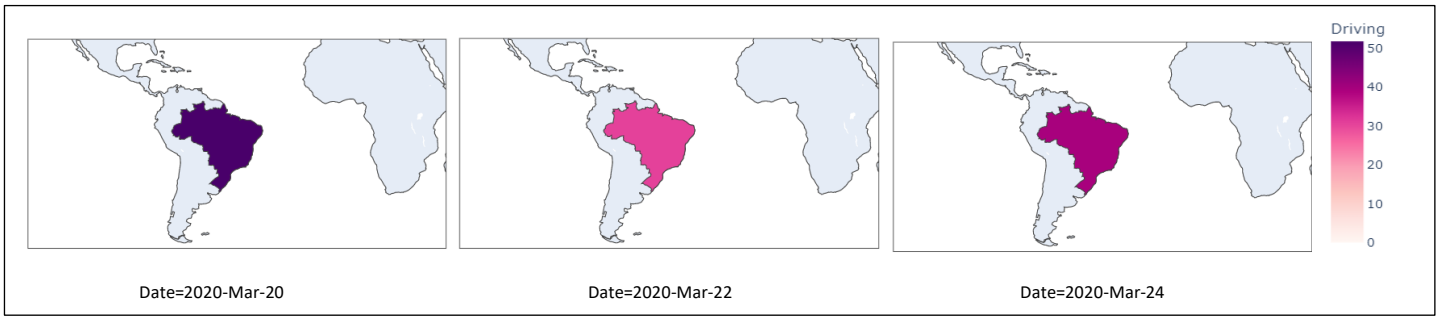


Figure (4.49) sample of United Arab Emirates data





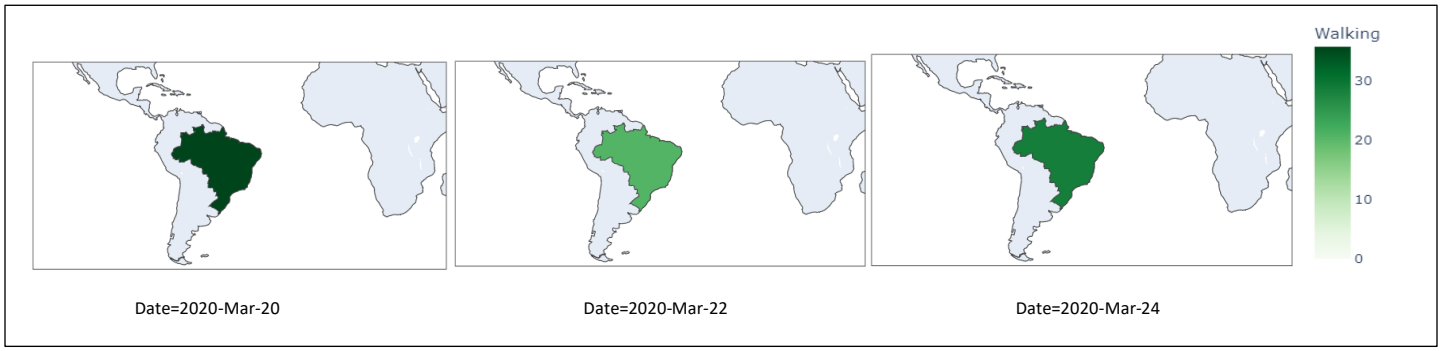


Figure (4.50) Continued

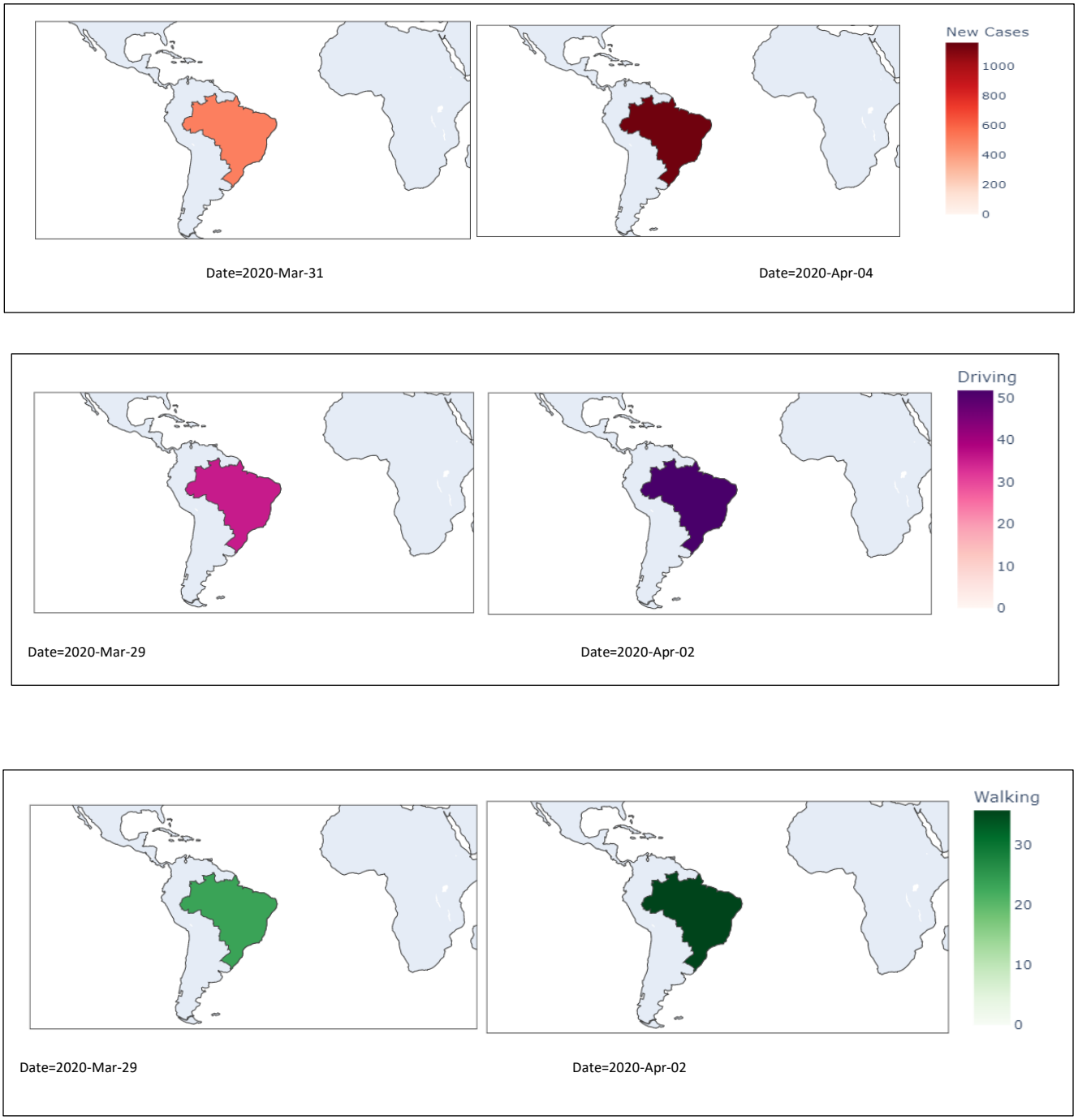


Figure (4.50) Sample of Brazil Data

Chapter Five
Conclusions and Future Work

5.1. Conclusion

Through the design, implementation, and discussion of the results of the system, the following concluded essential remarks:

1- According to the aggregation process of mobility data for each country involved in this work, the aggregated mobility dataset provided accurate information about the daily human mobility that influenced the steps of relationship modeling between Covid-19 transmission and human mobility, as well as the forecasting system. This step affected the work positively towards the right direction.

2- According to the experimental results, the knowledge discovery model of the proposed methodology has effectively modeled the relationship patterns between new infections of the disease spread and human mobility trends throughout the lockdown period for the study countries as well as during the lockdown lifting.

3- Our forecasting system has proved the efficiency of daily COVID-19 new cases and mobility trends of walking and driving prediction for almost three months.

4- The Random Forest regressor demonstrated its ability to present better results for daily predictions with less error than the ARIMA, LR, and LSTM models in terms of MAE and RMSE.

5- According to the visualization stage, two types of maps are created:

- The disease map of the whole integrated dataset provides a useful and different view on the daily spread of the disease, and the human movements after mapping it on a geographical worldwide map. By creating a map of the disease spread, it is possible to focus on human mobility as a cause that leads to the epidemic spread.
- Predictions map for COVID-19 new cases and human mobility that can reflect the closeness of the predictions to the truth thus it serves as a different view of evaluation.

5.2 Future Work

The suggestions for future works are as follows:

1- Applying the proposed system on other infectious disease datasets to model their spread relationship to human mobility.

2- Engaged other information such as population, the individual's gender, and their age as well as other factors that influence the disease transmission like taking into consideration, hand washing, mask use, and the rate of people who were vaccinated.

3- Studying one of the models that have been compared to the Random Forest regressor model and trying to minimize the error of prediction, as the errors of these models are not very bad.

References

References

References

[1] X. Yang, K. Stewart, L. Tang, Z. Xie, and Q. Li, "A review of GPS trajectories classification based on transportation mode," *Sensors (Switzerland)*, vol. 18, no. 11. MDPI AG, Nov. 02, 2018.

[2] R. Huerta and L. S. Tsimring, "Contact tracing and epidemics control in social networks," *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 66, no. 5, p. 4, Nov. 2002.

[3] E. Frias-Martinez, N. Oliver, A. Rubio, and V. Frias-Martinez, "Human Mobility in Advanced and Developing Economies: A Comparative Analysis. Human Mobility in Advanced and Developing Economies: A Comparative Analysis," 2010.

[4] K. Siła-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demšar, and A. S. Fotheringham, "Analysis of human mobility patterns from GPS trajectories and contextual information," *Int. J. Geogr. Inf. Sci.*, vol. 30, no. 5, pp. 881–906, May 2016.

[5] M. Wu, S. Han, M. Sun, and D. Han, "How the distance between regional and human mobility behavior affect the epidemic spreading," *Phys. A Stat. Mech. its Appl.*, vol. 492, pp. 1823–1830, Feb. 2018.

[6] L. Li, "Transmission dynamics of Ebola virus disease with human mobility in Sierra Leone," *Chaos, Solitons and Fractals*, vol. 104, pp. 575–579, Nov. 2017.

[7] R. M. Anderson *et al.*, "Epidemiology, transmission dynamics and control of SARS: The 2002-2003 epidemic," in *Philosophical Transactions of the Royal Society B: Biological Sciences*, Jul. 2004, vol. 359, no. 1447, pp. 1091–1105.

[8] H. S. Badr, H. Du, M. Marshall, E. Dong, M. M. Squire, and L. M. Gardner, "Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study," *Lancet Infect. Dis.*, Nov. 2020.

References

[9] M. Yuksel, Y. Aydede, and F. Begolli, "Dynamics of Social Mobility during the COVID-19 Pandemic in Canada," 2020.

[10] N. Askitas, K. Tatsiramos, and B. Verheyden, "Lockdown Strategies, Mobility Patterns and COVID-19," 2020.

[11] M. Yousaf, S. Zahir, M. Riaz, S. M. Hussain, and K. Shah, "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan," *Chaos, Solitons and Fractals*, vol. 138, Sep. 2020.

[12] D. S. Domingos, J. F. L. de Oliveira, and P. S. G. de Mattos Neto, "An intelligent hybridization of ARIMA with machine learning models for time series forecasting," *Knowledge-Based Syst.*, vol. 175, pp. 72–86, Jul. 2019.

[13] E. S. Al-Shamery and H. A. Al -Gashamy, "Enhanced Evolutionary Sequential Minimal Optimization Model for Inflation Prediction," 2018.

[14] X. Duan and X. Zhang, "ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data," *Data Br.*, vol. 31, Aug. 2020.

[15] S. Ghosal, S. Sengupta, M. Majumder, and B. Sinha, "Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 4, pp. 311–315, Jul. 2020.

[16] A. M. Almeshal, A. I. Almazrouee, M. R. Alenizi, and S. N. Alhajeri, "Forecasting the spread of COVID-19 in kuwait using compartmental and logistic regression models," *Appl. Sci.*, vol. 10, no. 10, 2020.

[17] G. L. Watson *et al.*, "Fusing a Bayesian case velocity model with random forest for predicting COVID-19 in the U.S.," *medRxiv*, p. 2020.05.15.20102608, Jan. 2020.

[18] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," *Big*

References

Data Min. Anal., vol. 4, no. 2, pp. 116–123, 2021.

[19] C. B. Aditya Satrio, W. Darmawan, B. U. Nadia, and N. Hanafiah, "Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET," in *Procedia Computer Science*, 2021, vol. 179, pp. 524–532.

[20] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons and Fractals*, vol. 135, 2020.

[21] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 1467–1474, 2020.

[22] Y. Zhou, R. Xu, D. Hu, Y. Yue, Q. Li, and J. Xia, "Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data," *Lancet Digit. Heal.*, vol. 2, no. 8, pp. e417–e424, Aug. 2020.

[23] A. Cartenì, L. Di Francesco, and M. Martino, "How mobility habits influenced the spread of the COVID-19 pandemic: Results from the Italian case study," *Sci. Total Environ.*, vol. 741, Nov. 2020.

[24] J. Saha, B. Barman, and P. Chouhan, "Lockdown for COVID-19 and its impact on community mobility in India: An analysis of the COVID-19 Community Mobility Reports, 2020," *Child. Youth Serv. Rev.*, vol. 116, Sep. 2020.

[25] A. Tobías, "Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up," *Sci. Total Environ.*, vol. 725, Jul. 2020.

[26] L. I. Oztig and O. E. Askin, "Human mobility and coronavirus disease 2019 (COVID-19): a negative binomial regression analysis," *Public Health*, vol.

185, pp. 364–367, Aug. 2020.

[27] S. A. Alasadi and W. S. Bhaya, "Review of Data Preprocessing Techniques.pdf," *Journal of Engineering and Applie Sciences*, vol. 12, no. 16. pp. 4102–4107, 2017.

[28] D. P. in D. M. I. S. R. L. 2015S Garcia, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining. Intelligent Systems Reference Library. 2015*, vol. 10. 2015.

[29] R. L. Hale, *Introduction to Data Mining*, vol. 19, no. 1. 2018.

[30] H. K. Obayes, "Drug Utilization Using Multi-objective Deep Neural Network And Spatial Visualization," 2020.

[31] C. Nwagu, O. Kenneth, and I. Chinecherem, "Knowledge Discovery in Databases (KDD): An Overview," vol. 15, no. 12, pp. 13–16, 2017.

[32] O. Esezi Isaac and A. Eric Chikweru, "Test for Significance of Pearson's Correlation Coefficient (r)," *Int. J. Innov. Math. Stat. Energy Policies*, vol. 1, no. 1, pp. 11–23, 2018.

[33] M. Baak, R. Koopman, H. Snoek, and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics," *Comput. Stat. Data Anal.*, vol. 152, p. 107043, 2020.

[34] B. V Chowdary and Y. Radhika, "A survey on applications of data mining techniques," *Int. J. Appl*, vol. 13, no. 7, pp. 5384–5392, 2018, [Online]. Available: https://www.ripublication.com/ijaer18/ijaerv13n7_112.pdf.

[35] N. Six, N. Herbaut, and C. Salinesi, "Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: A focused review," *Blockchain Res. Appl.*, p. 100061, 2022.

[36] N. P. Dharani, P. Bojja, and P. Raja Kumari, "Evaluation of performance of an LR and SVR models to predict COVID-19 pandemic," *Mater. Today Proc*, 2021.

References

[37] H. Pratyaksa, A. E. Permanasari, S. Fauziati, and I. Fitriana, "ARIMA implementation to predict the amount of antiseptic medicine usage in veterinary hospital," *Proc. 2016 1st Int. Conf. Biomed. Eng. Empower. Biomed. Technol. Better Futur. IBIOMED 2016*, no. October, pp. 5–6, 2017.

[38] E. H. A. Rady, H. Fawzy, and A. M. A. Fattah, "Time series forecasting using tree based methods," *J. Stat. Appl. Probab.*, vol. 10, no. 1, pp. 229–244, 2021.

[39] A. K. Sahai, N. Rath, V. Sood, and M. P. Singh, "ARIMA modelling & forecasting of COVID-19 in top five affected countries," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 1419–1427, 2020.

[40] F. M. Khan and R. Gupta, "ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India," *J. Saf. Sci. Resil.*, vol. 1, no. 1, pp. 12–18, Sep. 2020.

[41] H. R. Wang, C. Wang, X. Lin, and J. Kang, "An improved ARIMA model for precipitation simulations," *Nonlinear Process. Geophys.*, vol. 21, no. 6, pp. 1159–1168, 2014.

[42] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches," *Chaos, Solitons and Fractals*, vol. 138, Sep. 2020.

[43] P. Bühlmann, "Bagging, Boosting and Ensemble Learning," *Handb. Comput. Stat.*, no. 1, pp. 1–38, 2012, [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-21551-3_33.

[44] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results Phys.*, vol. 27, p. 104462, 2021.

[45] D. B. Vukovic, K. Romanyuk, S. Ivashchenko, and E. M. Grigorieva, "Are CDS spreads predictable during the Covid-19 pandemic? Forecasting based on

References

SVM, GMDH, LSTM and Markov switching autoregression," *Expert Syst. Appl.*, vol. 194, no. January, p. 116553, 2022.

[46] N. Mehdiyev, D. Enke, P. Fettke, and P. Loos, "Evaluating Forecasting Methods by Considering Different Accuracy Measures," *Procedia Comput. Sci.*, vol. 95, no. December, pp. 264–271, 2016.

[47] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, no. 1, 2018.

[48] H. Liu, A. Gegov, and M. Cocea, *Rule Based Systems for Big Data*, vol. 13. 2016.

[49] B. Qin, Y. Xia, S. Prabhakar, and Y. Tu, "A rule-based classification algorithm for uncertain data," *Proc. - Int. Conf. Data Eng.*, no. March, pp. 1633–1640, 2009.

[50] H. L. Viktor and E. Paquet, "Visualization Techniques for Data Mining," *Encycl. Data Warehous. Min.*, no. January, 2011.

[51] W. Zhou, *GIS for Earth Sciences*, 2nd ed. Elsevier Inc., 2021.

[52] D. Forrest, *Thematic Maps in Geography*, Second Edi., vol. 23. Elsevier, 2015.

[53] Kraemer M U G *et al.*, "The effect of human mobility and control measures on the COVID 19 epidemic in China," *Science*, vol. 4218, no. February 2019. pp. 1–13, 2020.

الخلاصة

تعد دراسة نمذجة انتشار الوباء والتنبوء بالاصابات المستقبلية من المواضيع المهمة خاصة فى حقل انظمة الرعاية الصحية حيث جذبت إنتباه واهتمام العديد من الباحثين خلال العقود القليلة الماضية للخوض فى هذا المجال ، خاصه وباء العصر كوفيد-19 .

يتكون البحث المقترح من ثلاث نماذج رئيسية : نموذج إكتشاف المعرفة، ونموذج التنبؤ، ونموذج التمثيل الصورى أو المرئى تم إستخدام نموذج إكتشاف المعرفة للعثور على العلاقة بين التنقل البشري وانتشاركوفيد -19 خلال فترات زمنية مختلفة. وتشير نتائج الدراسة لوجود علاقة واضحة وقوية بينهما . بالاضافة لتفسير منطقى لانماط تلك العلاقة بناءا على قواعد الاستدلال و استعراض بعض العوامل المؤثرة على النتائج خلال تلك الفترات الزمنية.

تم استخدام نموذج الغابات العشوائيه للتنبؤ بالسلوك العشوائى و الغير منتظم للإصابات اليومية ب كوفيد - 19 بالاضافة الى التنبؤ بتنقلات الإنسان بهيئاتها المختلفة (القيادة والمشى)، وتم تطبيق نظام التنبؤ على مجموعتى بيانات وهى مجموعة بيانات منظمة الصحة العالمية للاصابات اليومية بفايروس كورونا، ومجموعة بيانات شركة ابل للتنقل البشرى لسبع دول تضمنتها الدراسة وهى (الإمارات العربية المتحدة ، الصين ، جنوب إفريقيا ، المملكة المتحدة ، ألمانيا ، الدنمارك والبرازيل)، ومن اجل تقييم نموذج التنبؤ تم استخدام متوسط الخطأ المطلق وخطأ متوسط الجذر. علاوة على ذلك تمت مقارنة نموذج التنبؤ المقترح مع تقنيات شائعة اخرى وهى الإنحدار الخطى وشبكة الذاكرة طويلة المدى ونموذج الانحدار الذاتى المتكامل .

أظهرت النتائج أن أداء التقنية المقترحة يتفوق على أداء التقنيات الاخرى للاصابات اليومية بفايروس كورونا حيث وصلت افضل قيم لMAE و RMSEللنموذج المقترح

(0.004 و 0.002 للحالات الجديده الموزونة لافريقيا الجنوبية ، و (0.009 و 0.006) للتنقل البشري (القيادة) للامارات العربية المتحدة ، و (0.011 و 0.007) للتنقل البشري (المشي) للامارات العربية المتحدة.

من خلال نموذج التمثيل المرئى ، تم إنشاء الخريطة التصحيحية التى تصور كيفية اختلاف متغير عبر منطقه جغرافية أو إظهار مستوى التباين داخل منطقه ما بناءا على المواقع الجغرافية لدول الدراسة. تقدم الدراسه عرضا مفيداً من خلال خريطه GIS

حيث تم إنشاء نوعين من الخرائط التصحيحية: خريطة توضح انتشار المرض و تنقل الإنسان وخريطة توضح تنبؤات الاصابات الجديدة بالوباء بالاضافة الى تنبؤات حركة الانسان . تلعب هذه الخطوة أيضا دورا مهما في عملية التقييم من خلال مقارنة التنبؤات بالقيم الفعلية ، بالإضافة إلى تقديم رؤية أخرى لأنماط العلاقة بين انتشار الوباء وحركة الانسان .

جمهورية العراق
وزارة التعليم العالي و البحث العلمي
جامعة بابل/ كلية تكنولوجيا المعلومات

التنقل وإنتشار جائحة فيروس كورونا : تنبؤ، إكتشاف أنماط العلاقة، وتمثيل مرئى بإستخدام تقنيات تعلم الآلة

رسالة

الى مجلس كلية تكنولوجيا المعلومات / جامعة بابل وهي جزء من متطلبات نيل شهادة الماجستير في تكنولوجيا المعلومات / برمجيات

من قبل

ضحى حسين محمد جواد

بإشراف

أ.د. ايمان صالح صكبان

1443هـ **2022 م**