Visvesvaraya Technological University

BELAGAVI, KARNATAKA - 590 014.

PROJECT REPORT
ON

"PREDICTIVE ANALYSIS AND FORECASTING OF COVID-19 OUTBREAK"

Submitted By

AKANKSHA J	**[4PM17IS005]**
AKSHATHA S M	**[4PM17IS006]**
RAJESHWARI M HEBBALLI	**[4PM17IS034]**

Submitted in partial fulfillment of the requirement for the final year Internship of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

Under the Guidance of

Mr. Vinay S K
Asst .Professor, Dept. of IS&E.
PESITM, Shivamogga

PES Institute of Technology and Management

Department of Information Science & Engineering

July-2021

PES Institute of Technology & Management

NH-206, Sagar Road, Shivamogga-577 204

(Affiliated to Visvesvaraya Technological University, Belagavi)

Department of Information Science and Engineering

PESITM
education for the real world

CERTIFICATE

Certified that the project work entitled "*Predictive Analysis and Forecasting of COVID-19 Outbreak*" carried out jointly by

AKANKSHA J	[4PM17IS005]
AKSHATHA S M	[4PM17IS006]
RAJESHWARI M HEBBALLI	[4PM17IS034]

bonafide students of *PES INSTITUTE OF TECHNOLOGY & MANAGEMENT* in partial fulfillment for the award of Bachelor of Engineering in *INFORMATION SCIENCE & ENGINEERING* of the Visvesvaraya Technological University, Belagavi during the year *2020-21*. It is certified that all corrections/suggestions indicated forInternal Assessment have been incorporated in the report deposited in the department library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Project Guide

11/8/2021

Mr. Vinay S K
Asst .Professor, Dept. of IS&E.
PESITM, Shivamogga.

Project Coordinator

Dr. Pramod
Assoc .Professor, Dept. of IS&E.
PESITM, Shivamogga.

Head of the Department

Dr. Prasanna Kumar H.R.
Professor & Head, Dept. of IS&E
PESITM, Shivamogga.

Principal

24/8/21

Dr. Chaitanya Kumar M V
Principal,
PESITM, Shivamogga

Name of Examiner

1.....

2.....

Signature with date

.....

.....

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude to **Dr. Chaitanya Kumar M V,** Principal, PESITM, Shivamogga, for his kind support, guidance and encouragement throughout the course of this dissertation work.

I am highly grateful to **Dr. Prasanna Kumar H R,** Head, Department of IS&E, PESITM, Shivamogga for his kind support and encouragement throughout the course of this work.

I am highly grateful to **Mr. Vinay S K,** Asst.Professor, Department of IS&E, PESITM, Shivamogga for his careful and precious guidance which were extremely valuable for my study both theoretically and practically.

I would like to thank all the teaching and non-teaching staff of Department of IS&E for their kind Co-operation during the course of the work. The support provided by the College and Departmental library is gratefully acknowledged.

I am grateful to my parents and friends, who helped me in one way or the other throughout my work.

Akanksha J
Akshatha S M
Rajeshwari M Hebballi

ABSTRACT

Outbreak of Novel Corona Virus has created a disastrous situation in many countries around the world and caused millions of deaths. Prediction of the future trend of the disease in different countries can be useful for managing the outbreak. In this project, Data Analysis on World and India Dataset has been performed. Different visualization techniques are applied to the data set to visualize the trend of the confirmed covid-19 cases. This data uses features of past data for the future prediction. In our project the machine learning (ML)-guided linear regression model and multivariate regression model has been used to predict the future cases. The linear regression model and multivariate regression model has been fitted into the dataset to deal with the total number of confirmed, recovered, and death cases. The experimental results of corona virus cases prediction are obtained and the accuracy of the model is calculated using cross validation function in the prophet linear model.

TABLE OF CONTENTS

LIST OF FIGUERS

Chapter 1

INTRODUCTION

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. The outbreak in early December 2019 in the Hubei province of the People's Republic of China has spread worldwide. COVID-19 is now the name of the biggest panic. Nowadays on TV, newspaper, online, social media and every people think of this virus in their mouths and this virus makes people panic, why not, all the countries of the world become helpless. The whole world is stunned by this epidemic virus. The big countries are struggling today because of the coronavirus. The biggest force in the coronavirus is the human transmission. The most common symptoms of coronavirus are fever, tiredness, dry cough.

Machine learning (ML) proved to be an important study field, addressing many incredibly complicated and elaborate problems in the real world. Nearly all real-world areas, including healthcare, autonomous vehicles (AV) Enterprises software, NLP, smart robotics, sports, climatic simulation, voicing and image processing were included in the application fields. Important ML areas are expected. Many standard ML algorithms were used to direct potential outcomes in many application fields, including temperature, disease forecasting, stock market prediction, and disease prediction. Different models of regression and neural networks are broadly applicable to forecasts of future disease conditions in patients. Diverse experiments were undertaken to predict various ML disorders, such as cardiovascular disease, and coronary heart disease. And this project focuses in particular on COVID-19 analysis and the study focuses even on the prediction and early response of COVID-19 outbreaks to monitor the current situation by decision making through the prediction process in order to direct earlier interventions in the efficient management of the disease. This project intends to apply the machine learning models simultaneously with the forecast of expected reachability of the COVID-19 over the nations by using the real-time data.

Data Science and Machine Learning make use of different methods, processes, algorithms and system to extract knowledge or insights with its goal to discover hidden patterns from the raw data. It is about finding and exploring the data in the real world and then use it to solve some real world problems. With the help of datasets, Prediction and Forecasting techniques will be deployed to study the probable number of cases in the near future.

1.1 Machine Learning

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

1.2 Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and data. Data science is a “concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyze actual phenomena" with data.

It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

1.3 Dataset

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database table, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. Data sets can also consist of a collection of documents or files. Dataset may be in comma separated values (csv) or in tables. While handling the data, the data set can be a bunch of tables, schema and other objects. The data are essentially organized to a certain model that helps to process the needed information. The set of data is any permanently saved collection of information that usually contains either case-level, gathered data, or statistical guidance level data.

1.4 Data Analytics

Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze various behavioral patterns. The techniques and the tools used vary according to the organization or individual. Data Analytics has a key role in improving your business as it is used to gather hidden insights, generate reports, perform market analysis, and improve business requirements. Data is basically categorized as Structured and Unstructured, where structured data is highly organized and usually in the form of tables. Structured data statistics are for data visualization to find solution for business problems. Examples of structured data includes names, dates, and addresses, credit card numbers and many more. On the other had Unstructured data is most often categorized as qualitative data, and it cannot be processed and analyzed using conventional data tools and methods. Examples of unstructured data include text, video files, audio files, mobile activity, social med and many more.

1.5 Data Visualization

Data visualization is an interdisciplinary field that deals with the graphic representation of data. It is a particularly efficient way of communicating when the data is numerous as for example a Time Series. From an academic point of view, this representation can be considered as a mapping between the original data (usually numerical) and graphic elements (for example, lines or points in a chart). The mapping determines how the attributes of these elements vary according to the data. In this light, a bar chart is a mapping of the length of a bar to a magnitude of a variable. Since the graphic design of the mapping can adversely affect the readability of a chart mapping is a core competency of Data visualization.

1.6 Problem Statement

The Corona virus outbreak has already taken grip over people's life. The question that we are trying to answer is *"*will it be able to tackle this COVID-19 pandemic*"*. So, a step should be taken in helping people to understand, how the virus could spread across different countries and region.

1.7 Objectives

- To Analyze and Visualize COVID-19 Data.
- To Forecast the COVID-19 cases using Time Series Analysis.
- To build a model that will predict the next n day's Corona virus cases.

Chapter 2

LITERATURE SURVEY

Literature survey review refers to the content getting from the books which is related to the topic. It should be referred from the some research paper which is related to the topic which is given to the student. Any material which is related to the paper from internet and which is valuable for student and that literature review helped the student to enhance the report status and calculation, analysis and tabulation also strong which majorly reflects in the report.

2.1 Papers Referred

1. *"Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee"* A H M Jakaria , Md Mosharaf Hossain , Mohammad Ashiqur Rahman , November 2018, Research Gate.

Weather predictions are performed with the help of large complex models of physics, which utilize different atmospheric conditions over a long period of time. These conditions are often unstable because of perturbations of the weather system, causing the models to provide inaccurate forecasts. The models are generally run on hundreds of nodes in a large High Performance Computing (HPC) environment which consumes a large amount of energy. Here weather prediction technique that utilizes historical data from multiple weather stations to train simple machine learning models, which can provide usable forecasts about certain weather conditions for the near future within a very short period of time. The models can be run on much less resource intensive environments. A technology is presented to utilize machine learning techniques to provide weather forecasts. Machine learning technology can provide intelligent models, which are much simpler than traditional physical models. They are less resource-hungry and can easily be run on almost any computer including mobile devices. Evaluation results show that these machine learning models can predict weather features accurately enough to compete with traditional models and also can utilize the historical data from surrounding areas to predict weather of a particular area. Further with the utilization of low-cost Internet of Things (IoT) devices such as temperature and humidity sensors, the use of different sensors could increase the number of local features in the training dataset. This data, along with the weather station data, will further improve the performance of prediction model.

2. *"The Research of Regression Model in Machine Learning Field"* Shen Rong, Zhang Bao-wen, 2018, IFID.

The paper herein will analyze the sale of iced products affected by variation of temperature. Firstly, we will collect the data of the forecast temperature last year and the sale of iced products and then conduct data compilation and cleansing. Finally, we will set up the mathematical regression analysis model based on the cleansed data by means of data mining theory. Regression analysis refers to the method of studying the relationship between independent variable and dependent variable. Linear regression model that corresponds to the practical situation is proposed in the paper, which is to set up simple linear regression model based on practical problem and then to implement the following with the help of the latest and most popular Python3.6. Python3.6 boasts the features of pure object-oriented, platform independence and concise and elegant language. So we will call the corresponding library function to predict the sale of iced products according to the variation of temperature, which will provide the foundation for the company to adjust its production each month, or even each week and each day. As a result, the situation of overproduction can be avoided. Moreover, the other situation as the profit will be affected by the lack of production since the rise of temperature will also be avoided. So the regression model also has reference value for the other fields of marketing.

3. *"Crime Prediction and Analysis Using Machine Learning"* Alkesh Bharati1, Dr Sarvanaguru RA.K, Sep 2018, IRJET.

Crime is one of the biggest and dominating problem in our society and its prevention is an important task. Daily there are huge numbers of crimes committed frequently. This require keeping track of all the crimes and maintaining a database for same which may be used for future reference. The current problem faced are maintaining of proper dataset of crime and analyzing this data to help in predicting and solving crimes in future. The objective of this project is to analyze dataset which consist of numerous crimes and predicting the type of crime which may happen in future depending upon various conditions. In this project, we will be using the technique of machine learning and data science for crime prediction of Chicago crime data set. The crime data is extracted from the official portal of Chicago police. It consists of crime information like location description, type of crime, date, time, latitude, longitude.

Before training of the model data preprocessing will be done following this feature selection and scaling will be done so that accuracy obtain will be high. The K-Nearest Neighbor (KNN) classification and various other algorithms will be tested for crime prediction and one with better accuracy will be used for training. Visualization of dataset will be done in terms of graphical representation of many cases for example at which time the criminal rates are high or at which month the criminal activities are high. The soul purpose of this project is to give a jest idea of how machine learning can be used by the law enforcement agencies to detect, predict and solve crimes at a much faster rate and thus reduces the crime rate. It not restricted to Chicago, this can be used in other states or countries depending upon the availability of the dataset.

4. *"Global Warming Prediction in India Using Machine Learning",* D. Deva Hema, Anirban Pal, Vineet Loyer, Rajeev Gaurav 2019, IJEAT.

Long term global warming prediction can be of major importance in various sectors like climate related studies, agricultural, energy, medical and many more. This paper evaluates the performance of several Machine Learning algorithm (Linear Regression, Multi-Regression tree, Support Vector Regression (SVR), lasso) in problem of annual global warming prediction, from previous measured values over India. The first challenge dwells on creating a reliable, efficient statistical reliable data model on large data set and accurately capture relationship between average annual temperature and potential factors such as concentration of carbon dioxide, methane, nitrous oxide. The data is predicted and forecasted by linear regression because it is obtaining the highest accuracy for greenhouse gases and temperature among all the technologies which can be used. It was also found that CO2 is the plays the role of major contributor temperature change, followed by CH4, then by N20. After seeing the analysed and predicted data of the greenhouse gases and temperature, the global warming can be reduced comparatively within few years. The reduction of global temperature can help the whole world because not only human but also different animals are suffering from the global temperature. In this paper, the data (temperature and greenhouse gases) of 100-150 years is analyzed. Linear Regression and Linear model are used to predict and forecast the temperature and greenhouse gases for the next 10 years in average. The matplot library is used to plot the predicted and the forecasted data. So, at last the following conclusion can be drawn. A model for forecasting data for next 10 years is trained and tested with different input variables like temperature, carbon di-oxide, methane, nitrous oxide by linear regression. Some graphs are plotted as a graphical interface for the predicted and forecasted data for all the inputs with the help of matplot library.

5. *"Predicting Heart Diseases in Logistic Regression of Machine Learning Algorithms By Python Jupyterlab"* A. S. Thanuja Nishadi, Aug 2019, IJARP.

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper authors have proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM). Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible.

6. *"Prediction of Rainfall Using Machine Learning Techniques",* Moulana Mohammed, Roshitha Kolapalli, Niharika Golla, Siva Sai Maturi, 01 January 2020, International Journal of Scientific & Technology Research.

Rainfall prediction is important as heavy rainfall can lead to many disasters. The prediction helps people to take preventive measures and moreover the prediction should be accurate. There are two types of prediction short term rainfall prediction and long-term rainfall. Prediction mostly short-term prediction can give us the accurate result. The main challenge is to build a model for long term rainfall prediction. Heavy precipitation prediction could be a major drawback for earth science department because it is closely associated with the economy and lifetime of human. It's a cause for natural disasters like flood and drought that square measure encountered by individuals across the world each year.

Accuracy of rainfall statement has nice importance for countries like India whose economy is basically dependent on agriculture. The dynamic nature of atmosphere, applied mathematics techniques fail to provide sensible accuracy for precipitation statement. The prediction of precipitation using machine learning techniques may use regression. Intention of this project is to offer non experts' easy access to the techniques; approaches utilized in the sector of precipitation prediction and provide a comparative study among the various machine learning techniques. This project concentrated on estimation of rainfall and it is estimated that SVR is a valuable and adaptable strategy, helping the client to manage the impediments relating to distributional properties of fundamental factors, geometry of the information and the normal issue of model over fitting. The decision of bit capacity is basic for SVR displaying. We prescribe tenderfoots to utilize straight and RBF piece for direct and non-straight relationship individually. We see that SVR is better than MLR as an expectation strategy. MLR can't catch the non-linearity in a data set and SVR winds up helpful in such circumstances. We additionally process Mean Absolute Error (MAE) for both MLR and SVR models to assess execution of the models. At last, we look at the presentation of SLR, SVR and tuned SVR model. True to form, the tuned SVR model gives the best expectation.

7. ***"Forecasting COVID-19 cases in India Using Machine Learning Models",*** **Dr. Vakula Rani J, Aishwarya Jakka, July 2021, International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE).**

COVID-19 pandemic has affected the economy and changed the human way of life, disrupting everyone's mental, physical, and financial well-being. Many of the fastest growing economies are strained owing to the severity and communicability of the epidemic. Because of the increasing diversity of cases and the resulting burden on healthcare practitioners and the government, therefore, predicting the number of infected COVID-19 cases which could be useful in planning the required hospital resources in the future. In this paper, we focussed on information-led methods of estimating the numbers of COVID-19 confirmed cases in the country and their implications in the future, using different learning models such as Sigmoid modelling, ARIMA, SEIR model and LSTM, for protective measures, such as social isolation or the lockout of COVID-19. . Use of raw data by separating an event from the previous event in order to set the time series. The computation of number of positive incidents, number of re referred incidents are reliable within a limited range. A datadriven forecasting method has been used to approximate the total confirmed cases in coming months.

These LSTM model gave very promising results than other models. Hence, this work would help the decision makers to understand the upcoming of the pandemic trajectory in the country and take necessary actions for the effect of interventions.

8 . "COVID-19 Future Forecasting Using Supervised Machine Learning Model", **Furqan Rustam,Aijaz Ahmad Reshi, Arif Mehmood,Saleeem Ullah ,Byung-Won On, Waqar Aslam,and Gyu Sang Choi, May 2020, National Research of Korea (NRF) grant funded by Korea government (MSIT).**

Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes to improve the decision making on the future course of actions. The ML models have long been used in many application domains which needed the identifcation and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.

9. "A Comparative Approach To Predict Corona Virus Using Machine Learning", **Rohini, Naveena,Jothipriya.G,Kameshwaran,Jagadeeswari M A,2021, International Conference on Artificial Intelligence and Smart Systems (ICAIS).**

Coronavirus disease (COVID-19), is one of the most infectious diseases which reshaped our everyday lives globally in the 21st century. Technology progressions have a rapid effect on every field of life, be it the medical domain or any other. More than 250 countries have been affected by COVID in no matter of time.

The Indian government is making the necessary steps to control the spread of virus in the society. People all over the world are vulnerable to its consequences in the future. In a pandemic like this, people often worry whether they show a symptom of COVID-19 or not. Various AI methods have been applied successfully in epidemic studies. This paper presents the prediction and analysis of COVID-19 using various machine learning algorithms. In the present study, ML-based enhanced model is implemented to predict the possible threat of COVID-19 all over the world and the algorithms used in these models classifies the COVID patients based on several subsets of features and predicts their likeliness to get affected to this disease. This model uses 20 metrics including the patient's geographical location, travel history, health record statistics, etc., to predict the severity of the case and the feasible outcome. This research finds the patients exposed to Covid-19 and could be used as a reference, by the patients before consulting further with the doctor. The model developed using K-Nearest Neighbors (KNN) is effective with a prediction accuracy of 98.34%, Recall of 97% and an F1- Score of 0.97. Overall, this paper proposes a simple and practicable method to quickly identify and predict the high risk patients and provide priority to them for treatment so that the fatality rate can be decreased.

10. "Forecasting&Severity Analysis of COVID-19 Using Machine Learning Approach with AdvancedDataVisualization", **Ovi Sarkar, Md Faysal Ahamed, Pallab Chowdhury,Dec 2020, International Conference on Computer and Information Technology (ICCIT).**

SARS-CoV-2 (n-coronavirus) is a global pandemic that causes the deaths of millions of people worldwide. It can cause Pneumonia and severe acute respiratory syndrome (SARS) and lead to death in severe cases. It is an asymptomatic disease that hardens our life and work conditions. As there is no effective treatment available, many scientists and researchers are trying their best to fight the pandemic. This paper focused on the coronavirus pandemic situation in the global and Bangladesh region and its related effects and future status. We have utilized different information representation and machine learning calculations to recreate the affirmed, recuperated, and passing cases. We believe the research will help scientists, researchers, and ordinary people predict and analyze this pandemic's impact. Finally, the comparison and analysis of different models and algorithms successfully showed our visualization and prediction success.

11. "*Machine Learning Approaches for COVID-19 Forecasting*" , Othman Istaiteh , Tala Owais , Nailah Al-Madi Saleh Abu-Soud, 2020, International Conference on Intelligent Data Science Technologies and Applications (IDSTA)

COVID-19 (Coronavirus) pandemic tends to be one of the most global serious issues in the last century. Furthermore, the world did not face any similar experience regarding the spread of the virus and its economic and political impacts. Forecasting the number of COVID-19 cases in advance could help the decision-makers to take proactive measures and plans. This paper aims to provide a global forecasting tool that predicts the COVID-19 confirmed cases for the next seven days in all over the world. This paper applies four different machine learning algorithms; the autoregressive integrated moving average (ARIMA), artificial neural network (ANN), long-short term memory (LSTM), and convolutional neural network (CNN) to predict the COVID-19 cases in each country for the next seven days. The fine-tuning process of each model is described in this paper and numerical comparisons between the four models are concluded using different evaluation measures; mean absolute error (MAPE), root mean squared logarithmic error (RMSLE) and mean squared logarithmic error (MSLE).

***12. "Applying Machine Learning to Software Fault Prediction",* Bartłomiej Wojcicki , Robert Dabrowski.**

Software engineering continuously suffers from inadequate software testing. The automated prediction of possibly faulty fragments of source code allows developers to focus development efforts on fault-prone fragments first. Fault prediction has been a topic of many studies concentrating on C/C++ and Java programs, with little focus on such programming languages as Python. In this study the authors verify whether the type of approach used in former fault prediction studies can be applied to Python. More precisely, the primary objective is conducting preliminary research using simple methods that would support the expectation that predicting faults in Python programs is also feasible. The research demonstrates experimental evidence that fault-prediction methods similar to those developed for C/C++ and Java programs can be successfully applied to Python programs. The tool resulting from this research is a fault prediction tool for Python projects its performance was experimentally assessed on five open-source projects. On selected projects the tool achieved recall up to 0.64 with false positive rate 0.2.

13. "Flood Prediction Using Machine Learning Models", **Amir Mosavi, PinarOzturk and Kwok-wing Chau.**

Floods are among the most destructive natural disasters, which are highly complex to model. The research on the advancement of flood prediction models contributed to risk reduction, minimization of the loss of human life, and reduction of the property damage associated with floods. To mimic the complex mathematical expressions of physical processes of floods, during the past two decades, machine learning methods contributed highly in the advancement of prediction systems providing better performance and cost-effective solutions. Due to the vast benefits and potential of ML, its popularity dramatically increased among hydrologists. Researchers through introducing novel ML methods and hybridizing of the existing ones aim at discovering more accurate and efficient prediction models. The main contribution of this paper is to demonstrate the state of the art of ML models in flood prediction and to give insight into the most suitable models. The performance comparison of ML models presents an in-depth understanding of the different techniques within the framework of a comprehensive evaluation and discussion. As a result, this paper introduces the most promising prediction methods for both long-term and short-term floods, the major trends in improving the quality of the flood prediction models are investigated. This survey can be used as a guideline for hydrologists as well as climate scientists in choosing the proper ML method according to the prediction task.

14. "Temperature Forecast in Buildings Using Machine Learning Techniques", **Fernando Mateo , Juan J. Carrasco1 , Monica Millan-Giraldo, AbderrahimSellami, Pablo Escandell-Montero, Jose M. Martınez and Emilio Soria-Olivas.**

Energy efficiency in buildings requires having good prediction of the variables that define the power consumption in the building. Temperature is the most relevant of these variables because it affects the operation of the cooling systems in summer and the heating systems in winter, while being also the main variable that defines comfort. In the recent years, the interest in energy efficiency in large buildings has grown considerably. This interest is due to several reasons: the price growth of energy sources such as oil, the increasing need to preserve resources by shifting to renewable energy sources, the economic reasons like the increase in overall machine performance and industry productivity and the appeals and mandates imposed by governments to protect ecological systems and to ensure environmental sustainability.

One of the most energy-intensive elements is the HVAC (Heating, Ventilation and Air Conditioning) system. HVAC control systems should be able to predict the temperature response to changes in the input data (meteorological and environmental conditions, changing seasons, holiday periods, etc.). Temperature prediction is crucial for the management of energy efficiency in large buildings. This prediction can be made using different linear and non-linear techniques, several of which have been compared and tested using a simulated one-year temperature record for a particular building. An exhaustive analysis of the data has been made as a first step before building these models. The results show that linear regression methods produce similar errors, with a slight advantage for robust models.

15. "Applications of Machine Learning in Cancer Prediction and Prognosis", **Joseph A. Cruz, David S. Wishart.**

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer. Diagnosis and detection. More recently machine learning has been applied to cancer prognosis and prediction. This latter approach is particularly interesting as it is part of a growing trend towards personalized, predictive medicine. In assembling this review we conducted a broad survey of the different types of machine learning methods being used, the types of data being integrated and the performance of these methods in cancer prediction and prognosis.a strong bias towards applications in prostate and breast cancer, and a heavy reliance on "older" technologies such artificial neural networks (ANNs) instead of more recently developed or more easily interpretable machine learning methods. At a more fundamental level, it is also evident that machine learning is also helping to improve our basic understanding of cancer development and progression. In this review we have attempted toexplain,compare and assess the performance of different machine learning that are being applied to cancer prediction and prognosis. Specifically we identified a number of trends with respect to the types of machine learning methods being used,the types of training data being integrated, the kinds of end point predictions being made, the types of cancers being studied and the overall performance of these methods in predicting cancer susceptibility or outcomes.

It is also clear that machine learning methods generally improve the performance or predictive accuracy of most prognoses, especially when compared to conventional statistical or expert-based systems. we believe that if the quality of studies continues to improve, it is likely that the use of machine learning classifier will become much more common place in many clinical and hospital settings.

16. "Machine Learning Methods for Earthquake Prediction", **Alyona Galkina& Natalia Grafeeva.**

Earthquakes are one of the most dangerous natural disasters, primarily due to the fact that they often occur without an explicit warning, leaving no time to react. This fact makes the problem of earthquake prediction extremely important for the safety of humankind. Despite the continuing interest in this topic from the scientific community, there is no consensus as to whether it is possible to find the solution with sufficient accuracy. However, successful application of machine learning techniques to different fields of research indicates that it would be possible to use them to make more accurate short term forecasts. This paper reviews recent publications where application of various machine learning based approaches to earthquake prediction was studied. The aim is to systematize the methods used and analyze the main trends in making predictions. We believe that this research will be useful and encouraging for both earthquake scientists and beginner researchers in this field. In this research, the main approaches in application of machine learning methods to a problem of earthquake prediction are observed. The main open-source earthquake catalogs and databases are described. The definition of main metrics used for performance evaluation is given. A detailed review of published works is presented, which highlights the way of development of scientific methods in this area of research. Finally, during the discussion of the results achieved, further directions of research in the field of earthquake prediction are proposed. These are: Creating a "benchmark" earthquake dataset, which can be used to assess the quality of various predictor systems. The dataset includes frequently observed seismic zones and seismically active areas of East Asia and Europe, such as Central Japan and Sicily Island. The performance of previously proposed methods can also be evaluated using the «benchmark» dataset. Focusing on the most complex and important task of predicting earthquakes of high and extreme magnitudes (equal to or greater than 5.5). Making attempts to solve the problem of earthquake prediction in its original form, as determined by earthquake scientists; namely, the simultaneous specification of time, place and magnitude of seismic events with a certain probability.

***17. "Machine Learning Model for Stock Market Prediction",* Osman Hegazy , Omar S. Soliman and Mustafa Abdul Salam.**

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on a financial exchange. The successful prediction of a stock's future price will maximize investor's gains. This paper proposes a machine learning model to predict stock market price. The proposed algorithm integrates Particle swarm optimization (PSO) and least square support vector machine (LS-SVM). The PSO algorithm is employed to optimize LS-SVM to predict the daily stock prices. Proposed model is based on the study of stocks historical data and technical indicators. PSO algorithm selects best free parameters combination for LS-SVM to avoid over-fitting and local minima problems and improve prediction accuracy. The proposed model was applied and evaluated using thirteen benchmark financials datasets and compared with artificial neural network with Levenberg-Marquardt (LM) algorithm. The obtained results showed that the proposed model has better prediction accuracy and the potential of PSO algorithm in optimizing LS-SVM. This paper, proposed a machine learning model that integrates particle swarm optimization (PSO) algorithm and LS-SVM for stock price prediction using financial technical indicators.

These indicators include relative strength index, money flow index, exponential moving average, stochastic oscillator and moving average convergence/divergence The PSO is employed iteratively as global optimization algorithm to optimize LS-SVM for stock price prediction. Also, PSO algorithm used in selection of LS-SVM free parameters C (cost penalty), ϵ (insensitive-loss function) and γ (kernel parameter). The proposed LS-SVM-PSO model convergence International Journal of Computer **S**cience and Telecommunications [Volume 4, Issue 12, December 2013] 23 to the global minimum. Also, it is capable to overcome the over-fitting problem which found in ANN, especially in case of fluctuations in stock sector. -LS-SVM algorithm parameters can be tuned. The performance of the proposed model is better than LS-SVM and compared algorithms. LS-SVM-PSO achieves the lowest error value followed by single LS-SVM, while ANN-BP algorithm is the worst one.

18. "Forecasting of Photovoltaic power generation using Machine Learning", **Di Su, Efstratios Batzelis, Bikash Pal.**

Due to the intrinsic intermittency and stochastic nature of solar power, accurate forecasting of the photovoltaic (PV) generation is crucial for the operation and planning of PV intensive power systems. Several PV forecasting methods based on machine learning algorithms have recently emerged, but a complete assessment of their performance on a common framework is still missing from the literature. In this paper, a comprehensive comparative analysis is performed, evaluating ten recent neural networks and intelligent algorithms of the literature in short-term PV forecasting. All methods are properly fine-tuned and assessed on a one-year dataset of a 406 MWp PV plant in the UK. Furthermore, a new hybrid prediction strategy is proposed and evaluated, derived as an aggregation of the most well-performing forecasting models. Simulation results in MATLAB show that the season of the year affects the accuracy of all methods, the proposed hybrid one performing most favorably overall. In this paper, a comprehensive performance assessment among some of the most popular PV power forecasting methods is performed on a common dataset. NARXNN is found to be superior over other neural networks due to its dynamic feedback mechanism. RF performs the best among the intelligence algorithms, since it is a combination of uncorrelated decision trees that exhibits bad data tolerance. There is a seasonal effect on the forecasting problem; summer and autumn are easier to forecast than spring and winter. The training process of a neural network exhibits great randomness, while intelligent algorithms are generally more robust. The proposed Hybrid method performs most favourably among all methods, correcting erroneous fluctuations and negative forecasting. In fact, a major conclusion from this investigation is that simple combination of several good models can generate a more reliable prediction than any single method on its own. This may be found useful especially when there is no complete data for model training.

19." Machine-Learning Models for Sales Time Series Forecasting", **Bohdan M. Pavlyshenko**

In this paper, we study the usage of machine-learning models for sales predictive analytics. The main goal of this paper is to consider main approaches and case studies of using machine learning for sales forecasting. The effect of machine-learning generalization has been considered. This effect can be used to make sales predictions when there is a small amount of historical data for specific sales time series in the case when a new product or store is launched.

A stacking approach for building regression ensemble of single models has been studied. The results show that using stacking techniques, we can improve the performance of predictive models for sales time series forecasting. In our case study, we considered different machine-learning approaches for time series forecasting. Sales prediction is rather a regression problem than a time series problem. The use of regression approaches for sales forecasting can often give us better results compared to time series methods. One of the main assumptions of regression methods is that the patterns in the historical data will be repeated in future. The accuracy on the validation set is an important indicator for choosing an optimal number of iterations of machine-learning algorithms. The effect of machine-learning generalization consists in the fact of capturing the patterns in the whole set of data. This effect can be used to make sales prediction when there is a small number of historical data for specific sales time series in the case when a new product or store is launched. In stacking approach, the results of multiple model predictions on the validation set are treated as input regressors for the next level models.

***20." Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process"*, E. Chandra Blessie, R. Rekha**.

Extending credits to corporates and individuals for the smooth functioning of growing economies like India is inevitable. As increasing number of customers apply for loans in the banks and non-banking financial companies (NBFC), it is really challenging for banks and NBFCs with limited capital to device a standard resolution and safe procedure to lend money to its borrowers for their financial needs. In addition, in recent times NBFC inventories have suffered a significant downfall in terms of the stock price. It has contributed to a contagion that has also spread to other financial stocks, adversely affecting the benchmark in recent times. In this paper, an attempt is made to condense the risk involved in selecting the suitable person who could repay the loan on time thereby keeping the bank's non-performing assets (NPA) on the hold. This is achieved by feeding the past records of the customer who acquired loans from the bank into a trained machine learning model which could yield an accurate result. The prime focus of the paper is to determine whether or not it will be safe to allocate the loan to a particular person. This paper has the following sections (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.

By properly analysing positive qualities and constraints, it can be concluded with confidence that the Naïve Bayes model is extremely efficient and gives a better result when compared to other models. It works correctly and fulfils all requirements of bankers and can be connected to many other systems. There were multiple malfunctions in the computers, content errors and fixing of weight in computerized prediction systems. In the near term, the banking software could be more reliable, accurate, and dynamic in nature and can be fit in with an automated processing unit. The old data sets are initially fed into the system for training purpose and then the new data sets. Machine learning helps to understand the factors which affect the specific outcomes most. Other models like neutral network and discriminate analysis can be used individually or combined for enhancing reliability and accuracy prediction.

***21. "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs",* Jie Xu, Member, IEEE, Kyeong Ho Moon, Student Member, IEEE, and Mihaela van der Schaar, Fellow, IEEE.**

Accurately predicting students' future performance based on their ongoing academic records is crucial for effectively carrying out necessary pedagogical interventions to ensure students' on-time and satisfactory graduation. Although there is a rich literature on predicting student performance when solving problems or studying for courses using data-driven approaches, predicting student performance in completing degrees (e.g. college programs) is much less studied and faces new challenges: (1) Students differ tremendously in terms of backgrounds and selected courses; (2) Courses are not equally informative for making accurate predictions; (3) Students' evolving progress needs to be incorporated into the prediction. In this paper, we develop a novel machine learning method for predicting student performance in degree programs that is able to address these key challenges. The proposed method has two major features. First, a bilayer structure comprising of multiple base predictors and a cascade of ensemble predictors is developed for making predictions based on students' evolving performance states. Second, a data-driven approach based on latent factor models and probabilistic matrix factorization is proposed to discover course relevance, which is important for constructing efficient base predictors. Through extensive simulations on an undergraduate student dataset collected over three years at UCLA, we show that the proposed method achieves superior performance to benchmark approaches. In this paper, we proposed a novel method for predicting students' future performance in degree programs given their current and past performance. A latent factor model-based course clustering method was developed to discover relevant courses for constructing base predictors.

An ensemble-based progressive prediction architecture was developed to incorporate students' evolving performance into the prediction. These data-driven methods can be used in conjunction with other pedagogical methods for evaluating students' performance and provide valuable information for academic advisors to recommend subsequent courses to students and carry out pedagogical intervention measures if necessary. Additionally, this work will also impact curriculum design in degree programs and education policy design in general. Future work includes extending the performance prediction to elective courses and using the prediction results to recommend courses to students.

***22. "A Time-Series Water Level Forecasting Model Based on Imputation and Variable Selection Method",* Jun-He Yang, Ching-Hsue Cheng, and Chia-Pan Chan.**

Reservoirs are important for households and impact the national economy. This paper proposed a time-series forecasting model based on estimating a missing value followed by variable selection to forecast the reservoir's water level. This study collected data from the Taiwan Shimen Reservoir as well as daily atmospheric data from 2008 to 2015. The two datasets are concatenated into an integrated dataset based on ordering of the data as a research dataset. The proposed time-series forecasting model summarily has three foci. First, this study uses five imputation methods to directly delete the missing value. Second, we identified the key variable via factor analysis and then deleted the unimportant variables sequentially via the variable selection method. Finally, the proposed model uses a Random Forest to build the forecasting model of the reservoir's water level. This was done to compare with the listing method under the forecasting error. These experimental results indicate that the Random Forest forecasting model when applied to variable selection with full variables has better forecasting performance than the listing model. In addition, this experiment shows that the proposed variable selection can help determine five forecast methods used here to improve the forecasting capability. This study proposed a time-series forecasting model for water level forecasting in Taiwan's Shimen Reservoir. The experiments showed that the mean of nearby points' imputation method has the best performance. These experimental results indicate that the Random Forest forecasting model when applied to variable selection with full variables has better forecasting performance than the listing model. The key variables are Reservoir IN, Temperature, Reservoir OUT, Pressure, Rainfall, and Relative Humidity. The proposed time-series forecasting model with/without variable selection has better forecasting performance than the listing models using the five evaluation indices. This shows that the proposed time-series forecasting model is feasible for forecasting water levels in Shimen Reservoir.

Chapter 3

SYSTEM ANALYSIS

Systems analysis is "the process of studying a procedure or business in order to identify its goals and purposes and create systems and procedures that will achieve them in an efficient way". Another view sees system analysis as a problem-solving technique that breaks down a system into its component pieces for the purpose of the studying how well those component parts work and interact to accomplish their purpose. The field of system analysis relates closely to requirements analysis or to operations research. It is also "an explicit formal inquiry carried out to help a decision maker identify a better course of action and make a better decision than they might otherwise have made."

Analysis is defined as "the procedure by which we break down an intellectual or substantial whole into parts," while synthesis means "the procedure by which we combine separate elements or components in order to form a coherent whole". System analysis is used in every field where something is developed. Analysis can also be a series of components that perform organic functions together, such as system engineering. System engineering is an interdisciplinary field of engineering that focuses on how complex engineering projects should be designed and managed.

3.1 Existing System

Several outbreak prediction models for COVID-19 are being used by officials around the world to make informed decisions and enforce relevant control measures [1]. Among the standard models for COVID-19 global pandemic prediction, simple epidemiological and statistical models have received more attention by authorities, and these models are popular in the media. Due to a high level of uncertainty and lack of essential data, standard models have shown low accuracy for long-term prediction. Although the literature includes several attempts to address this issue, the essential generalization and robustness abilities of existing models need to be improved [2].

3.2 Proposed System

The COVID-19 pandemic has already taken grip over people's life. Since the start of the pandemic, whole world is facing problem of ever-increasing cases. Through the data analysis of cases one can analyze how countries all over the world are doing in terms of controlling the pandemic.

Analyzing data leads to adapt the prevention model of the countries that are doing great in terms of lowering the graph. Predictions are made with the dataset available to the country/organizations, thus helping them to decide how far they are able to control the pandemic or up to how much extent they should guide preventive measures. Through this project, a step towards helping people to understand the spread and predict the cases all over the World and in India will be done.

3.3 Methodology

The collected COVID-19 dataset is preprocessed using various techniques like data cleaning, data reduction and data transformation. The preprocessed data is analyzed and visualized and later fed into Linear Regression Model. The designed model predicts the probable number of cases in the upcoming days.

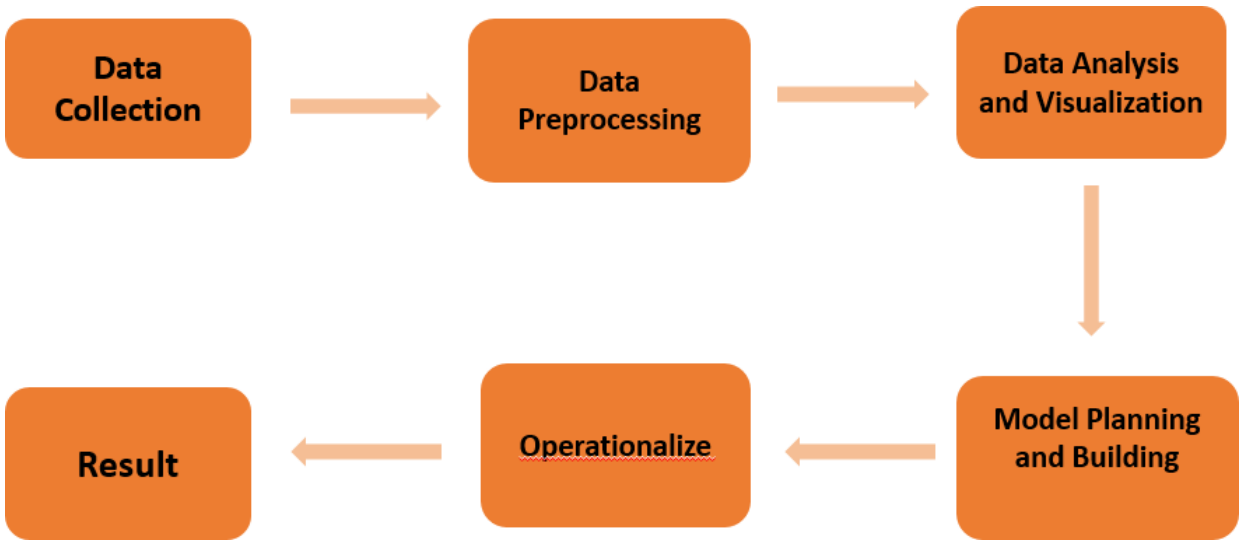


Fig 3.1 Methodology

Chapter 4

SYSTEM REQUIREMENTS AND SPECIFICATIONS

A System Requirement Specification (SRS) is a comprehensive description of the intended purpose and environment for software under development. The SRS fully describes what the software will do and how it will be expected to perform. An SRS minimizes the time and effort required by developers to achieve desired goals and also minimizes the development cost. A good SRS defines how an application will interact with the system hardware, other programs and human users in a wide variety of real-world situations.

4.1 Software Requirements

- Operating system
  - Windows 7 or above
- Web Application
  - Google Collab
- Web Browser
  - Google Chrome, Mozilla Firefox

4.2 Hardware Requirements

- Processor
  - Core i3 or above
- RAM
  - 4GB or above
- Disk Space
  - 3GB or above

Chapter 5

SYSTEM DESIGN

System Design is a broad term for describing methodologies for developing high quality information system which combines Information Technology, people and Data to support business requirement. System design is the phase that bridges the gap between problem domain and the existing system in a manageable way. This phase focuses on the solution domain, it is the phase where the System Requirement Specification (SRS) document is converted into a format that can be implemented and decides how the system will operate. In this phase, the complex activity of system development is divided into several smaller sub-activities, which coordinate with each other to achieve the main objective of system development.

5.1 System Architecture

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. The architecture of a system describes its major components, their relationships (structures), and how they interact with each other. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system.

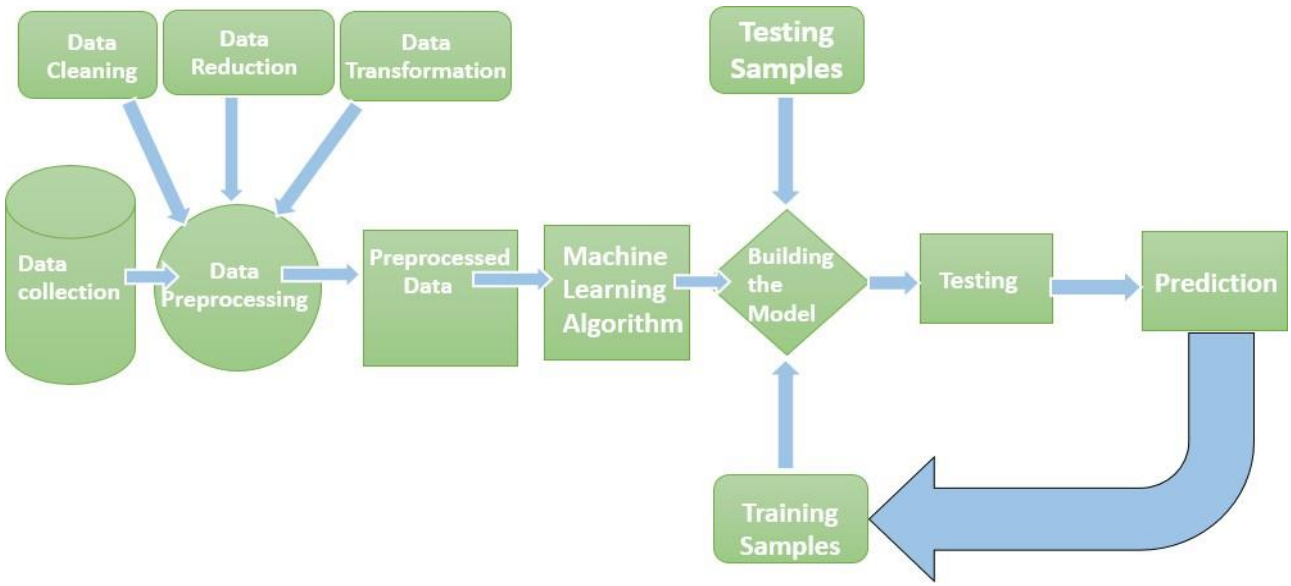


Fig 5.1 System Architecture

5.2 Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential orderi.e. the order in which these interactions take place. The sequence diagram represents the flow of messages in the system and is also termed as an event diagram. It helps in envisioning several dynamic scenarios. It portrays the communication between any two lifelines as a time-ordered sequence of events, such that these lifelines took part at the run time. In UML, the lifeline is represented by a vertical bar, whereas the message flow is represented by a vertical dotted line that extends across the bottom of the page. It incorporates the iterations as well as branching. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function.

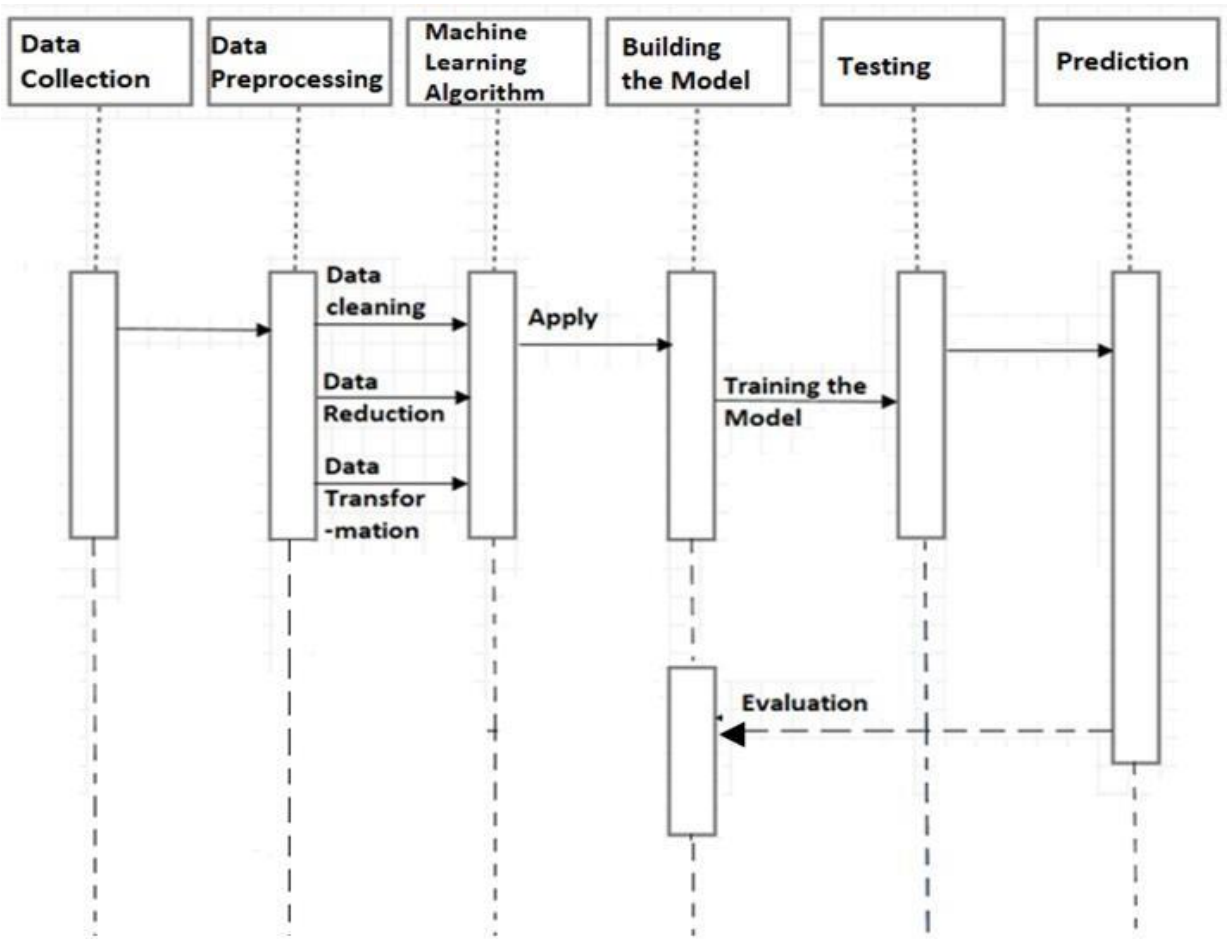


Fig 5.2 Sequence diagram

5.3 Data Flow Diagram

Data flow diagrams also known as DFD, are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchy of detailed diagrams.

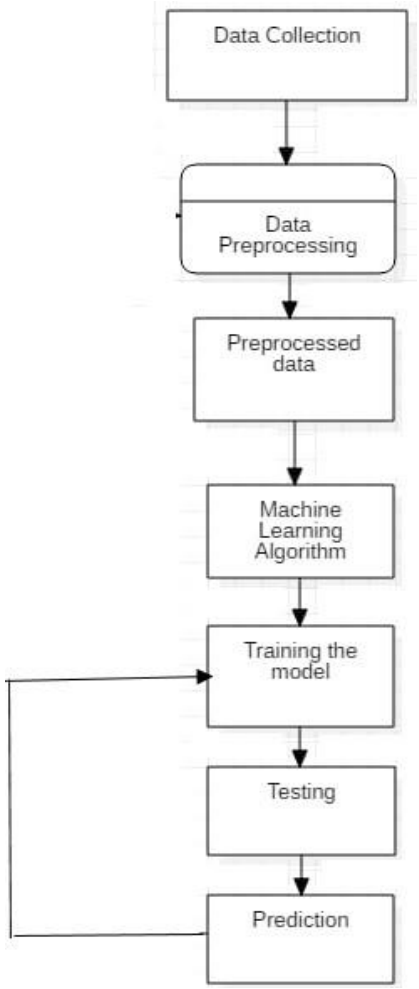


Fig 5.3 Data Flow diagram

Chapter 6

IMPLEMENTATION

The implementation phase involves putting the project plan, idea, model, design, specification, algorithms into action. As shown in the project design, our project consists of several steps. They are as follows:

1. Collection of Dataset
2. Data Preprocessing
3. Applying preprocessed data to the Algorithm
4. Building the model
5. Prediction

Collecting data is the first step in data processing. Data is pulled from available sources. Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as "pre-processing". Here raw data is cleaned up and organized. During preparation, raw data is checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data to procure best results.

6.1 Analysis of Worldwide and India's Dataset

- Covid-19 Worldwide dataset is collected for analysis.
- As mentioned before, Pandas, matplotlib, Seaborn, Numpy, Plotly libraries are imported.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import plotly.express as px
```

- Importing the dataset

```
path = 'https://raw.githubusercontent.com/datasets/covid-19/main/data/countries-aggregated.csv'
df = pd.read_csv(path)
df.head( )
```

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- Checking the information of the data.

```
df.info ()
```

- Transferring data into the required format.

```
df = pd.read_csv(path,parse_dates=['Date'])
df.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 96693 entries, 0 to 96692
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       96693 non-null  datetime64[ns]
 1   Country    96693 non-null  object
 2   Confirmed  96693 non-null  int64
 3   Recovered  96693 non-null  int64
 4   Deaths     96693 non-null  int64
dtypes: datetime64[ns](1), int64(3), object(1)
memory usage: 3.7+ MB
```

- Calculating the Active cases.

```
active = df['Confirmed'] - df['Recovered'] - df['Deaths']
df['Active'] = active
df.head()
```

	Date	Country	Confirmed	Recovered	Deaths	Active
96688	2021-06-01	Zimbabwe	38998	36624	1599	775
96689	2021-06-02	Zimbabwe	39031	36661	1599	771
96690	2021-06-03	Zimbabwe	39092	36680	1604	808
96691	2021-06-04	Zimbabwe	39144	36690	1605	849
96692	2021-06-05	Zimbabwe	39168	36698	1605	865

- Displaying the latest date data

```
top = df[df['Date']== df['Date'].max()]
top.head()
```

	Date	Country	Confirmed	Recovered	Deaths	Active
500	2021-06-05	Afghanistan	77963	58265	3104	16594
1001	2021-06-05	Albania	132374	129627	2451	296
1502	2021-06-05	Algeria	130681	90995	3510	36176
2003	2021-06-05	Andorra	13758	13527	127	104
2504	2021-06-05	Angola	35594	28866	794	5934

- Removing the duplicate values

```
world = top.groupby('Country')['Active','Confirmed','Recovered','Deaths'].sum()
world.head()
```

Country	Active	Confirmed	Recovered	Deaths
Afghanistan	16594	77963	58265	3104
Albania	296	132374	129627	2451
Algeria	36176	130681	90995	3510
Andorra	104	13758	13527	127
Angola	5934	35594	28866	794

- Resetting the Index

```
world = top.groupby('Country')['Active', 'Confirmed', 'Recovered', 'Deaths']
aths'].sum().reset_index()
world.head()
```

	Country	Active	Confirmed	Recovered	Deaths
0	Afghanistan	16594	77963	58265	3104
1	Albania	296	132374	129627	2451
2	Algeria	36176	130681	90995	3510
3	Andorra	104	13758	13527	127
4	Angola	5934	35594	28866	794

- Plotting Active, Confirmed, Recovered and Death cases on the world map

```
figure = px.choropleth(world,locations='Country',locationmode='country names'
,color='Active',hover_name='Country',range_color=[1,2000],color_continuous_
scale='purp',title='Countries With Active Cases')
 figure.show()
```

- plotting worldwide active cases over date using point plot

```
plt.figure(figsize=(5,5))
plt.xticks(rotation=90,fontsize=10)
```

```
sns.pointplot(total_active_cases['Date'].dt.date.tail(5),total_active_cases['Active'])
plt.show()
```

- Top 20 Countries based on the total active cases.

```
top_active = top.groupby('Country')['Active'].sum().sort_values
(ascending=False).reset_index()
top_20_a = top_active.head(20)
top_20_a
```

- Plotting top 20 countries based on the total active cases on the bar plot

```
plt.figure(figsize=(12,10))
sns.barplot(top_20_a['Active'],top_20_a['Country'])
plt.show()
```

- plotting worldwide active cases over date using point plot

```
plt.figure(figsize=(5,5))
plt.xticks(rotation=90,fontsize=10)
sns.pointplot(total_active_cases['Date'].dt.date.tail(5),total_active_cases['Active'])
plt.show()
```

- Top 20 Countries based on the total active cases.

```
top_active = top.groupby('Country')['Active'].sum().sort_values
(ascending=False).reset_index()
 top_20_a = top_active.head(20)
 top_20_a
```

- Plotting top 20 countries based on the total active cases on the bar

```
plotplt.figure(figsize=(12,10))
sns.barplot(top_20_a['Active'],top_20_a['Country'])
plt.show()
```

- Covid-19 India's dataset is collected for analysis.
- Importing libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import plotly.express as px
```

- Importing the dataset

```
from google.colab import files
uploaded = files.upload()
import io
df = pd.read_csv(io.BytesIO(uploaded['C19India.csv']))
df.head()
```

- Checking the format of data and transforming them

```
df.info()
df = pd.read_csv('C19India.csv',parse_dates=['Date'])
df.head()
```

- Removing the unwanted columns

```
df.drop (['ConfirmedIndianNational','Sno','ConfirmedForeignNational',
' Time'],axis = 1, inplace= True)
df.head()
```

- Renaming the column

```
df.rename(columns ={"state/unionteritory":"state","cured":"recovered"},inplace=True)
df.head()
```

- Calculating the active cases

```
active = df['Confirmed'] - df['Recovered'] - df['Deaths']
df['Active'] = active
df.head()
```

- Sorting the values

```
data = df[['State','Confirmed','Active','Deaths','Recovered','Date']]
data.tail( )
 data.sort_values('Active',ascending=False,inplace=True)
 data.head( )
 data.sort_values('Confirmed',ascending=False,inplace=True)
 data.head()
 data.sort_values('Recovered',ascending=False,inplace=True)
 data.head()
 data.sort_values('Deaths',ascending=False,inplace=True)
 data.head()
 df.tail( )
```

6.2 Forecasting of Confirmed Cases

- Importing libraries

```
Import pandas as pd
Import matplotlib.pyplot as plt
Import seaborn as sns
Import numpy as np
Import plotly.express as px
```

- Path of dataset

```
Path=‘http”//raw.githubusercontent.com/datasets/covid19/main/datacountrieaggrega
ted.csv’
df= pd.read_csv(path)
df.head()
```

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- checking the information of the data in the datset

```
df.info()
df = pd.read_csv(path,parse_dates=['Date'])
df.head()
```

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- adding new column Active by calculating active

```
active = df['Confirmed'] - df['Recovered'] - df['Deaths']
df['Active'] = active
df.tail()
```

	Date	Country	Confirmed	Recovered	Deaths	Active
101785	2021-06-22	Zimbabwe	42714	37288	1691	3735
101786	2021-06-23	Zimbabwe	43480	37477	1692	4311
101787	2021-06-24	Zimbabwe	44306	37524	1709	5073
101788	2021-06-25	Zimbabwe	45217	37604	1721	5892
101789	2021-06-26	Zimbabwe	46018	37761	1725	6532

- active = df.groupby('Date')['Active'].sum().reset_index()
  active.tail()

	Date	Active
517	2021-06-22	58168650
518	2021-06-23	58196639
519	2021-06-24	58302153
520	2021-06-25	58385964
521	2021-06-26	58455776

- confirmed = df.groupby('Date')['Confirmed'].sum().reset_index()
  confirmed.tail()

	Date	Confirmed
517	2021-06-22	179158296
518	2021-06-23	179595551
519	2021-06-24	179999388
520	2021-06-25	180421128
521	2021-06-26	180733756

- from fbprophet import Prophet,library in python created by facebook for the time series analysis
- confirmed.rename(columns={'Date':"ds","Confirmed":"y"},inplace=True)
  confirmed

	ds	y
0	2020-01-22	557
1	2020-01-23	655
2	2020-01-24	941
3	2020-01-25	1433
4	2020-01-26	2118
...	...	...
517	2021-06-22	179158296
518	2021-06-23	179595551
519	2021-06-24	179999388
520	2021-06-25	180421128
521	2021-06-26	180733756

522 rows × 2 columns

- building the model

model = Prophet(interval_width=0.95)

- apply the model (train the model)

model.fit(confirmed)

- future = model.make_future_dataframe (periods=7)

future.tail (7)

- forecast=model.predict(future)
- forecast[['ds','yhat','yhat_lower','yhat_upper']].tail(7)

	ds	yhat	yhat_lower	yhat_upper
522	2021-06-27	1.863922e+08	1.845765e+08	1.882844e+08
523	2021-06-28	1.869639e+08	1.850823e+08	1.887361e+08
524	2021-06-29	1.875874e+08	1.856718e+08	1.894774e+08
525	2021-06-30	1.882436e+08	1.863920e+08	1.900772e+08
526	2021-07-01	1.889109e+08	1.870480e+08	1.907515e+08
527	2021-07-02	1.895649e+08	1.876426e+08	1.914339e+08
528	2021-07-03	1.901855e+08	1.883117e+08	1.922256e+08

- pd.concat([confirmed.set_index('ds')['y'],forecast.set_index('ds')['yhat']],axis=1).plot()<matplotlib.axes._subplots.AxesSubplot at 0x7fa47b7cd550>
- confirmed_plot = model.plot(forecast)
  confirmed_plot_weekly=model.plot_components(forecast)
- Forecasting confirmed
  India
- df_india = df[df['Country']=='India']
  df_india.tail()

	Date	Country	Confirmed	Recovered	Deaths	Active
41755	2021-06-22	India	30028709	28994855	390660	643194
41756	2021-06-23	India	30082778	29063740	391981	627057
41757	2021-06-24	India	30134445	29128267	393310	612868
41758	2021-06-25	India	30183143	29193085	394493	595565
41759	2021-06-26	India	30183143	29193085	394493	595565

- active = df_india.groupby('Date')['Active'].sum().reset_index()
  confirmed=df_india.groupby('Date'['Confirmed'].sum().reset_index()
- confirmed.rename(columns={'Date':"ds","Confirmed":"y"},inplace=True)
  confirmed.tail()
- Building the Model
  m = Prophet(interval_width=0.95)
  m.fit(confirmed)
- future = m.make_future_dataframe(periods = 7)
  future.head()
- future.tail()
- forecast = m.predict(future)
- forecast[['ds','yhat','yhat_lower','yhat_upper']].tail(7)

	ds	yhat	yhat_lower	yhat_upper
522	2021-06-27	3.232743e+07	3.101770e+07	3.373650e+07
523	2021-06-28	3.252870e+07	3.122119e+07	3.390117e+07
524	2021-06-29	3.273539e+07	3.149386e+07	3.409114e+07
525	2021-06-30	3.294769e+07	3.157827e+07	3.428564e+07
526	2021-07-01	3.315794e+07	3.188063e+07	3.450559e+07
527	2021-07-02	3.336824e+07	3.199458e+07	3.476245e+07
528	2021-07-03	3.357855e+07	3.222411e+07	3.497786e+07

- india_plot = m.plot(forecast)
- confirmed_plot_weekly=m.plot_components(forecast)
- from fbprophet.diagnostics import cross_validation

df_cv=cross_validation(model=m,initial='180 days', horizon='90 days')

- df_cv.head()

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2020-08-16	2.452140e+06	2.368877e+06	2.528655e+06	2647663	2020-08-15
1	2020-08-17	2.495629e+06	2.413550e+06	2.582295e+06	2702681	2020-08-15
2	2020-08-18	2.539781e+06	2.458320e+06	2.625281e+06	2767253	2020-08-15
3	2020-08-19	2.584870e+06	2.509085e+06	2.664091e+06	2836925	2020-08-15
4	2020-08-20	2.630345e+06	2.554656e+06	2.712398e+06	2905825	2020-08-15

- from fbprophet.diagnostics import performance_metrics

df_p= performance_metrics(df_cv)

df_p

	horizon	mse	rmse	mae	mape	mdape	coverage
0	9 days	1.530935e+11	3.912716e+05	3.335030e+05	0.043386	0.036743	0.074074
1	10 days	1.806192e+11	4.249931e+05	3.573683e+05	0.045804	0.038667	0.092593
2	11 days	2.133742e+11	4.619245e+05	3.830941e+05	0.048411	0.042652	0.111111
3	12 days	2.512065e+11	5.012050e+05	4.092493e+05	0.050990	0.045136	0.129630
4	13 days	2.952874e+11	5.434035e+05	4.364265e+05	0.053576	0.047687	0.148148
...	...	...	...	...	...	...	...
77	86 days	6.293511e+13	7.933165e+06	5.911787e+06	0.334079	0.306240	0.000000
78	87 days	6.427769e+13	8.017337e+06	5.997329e+06	0.337028	0.307549	0.000000
79	88 days	6.563858e+13	8.101764e+06	6.081327e+06	0.339858	0.310838	0.000000
80	89 days	6.702389e+13	8.186812e+06	6.164268e+06	0.342581	0.315802	0.000000
81	90 days	6.841037e+13	8.271056e+06	6.245524e+06	0.345191	0.320822	0.018519

82 rows × 7 columns

6.3 Multivariate Forecasting

- Importing the libraries
  import pandas as pd
  import matplotlib.pyplot as plt
  import seaborn as sns
  import numpy as np
  import plotly.express as px
- Path of the raw dataset

path = 'https://raw.githubusercontent.com/datasets/covid- 19/main/data/countries aggregated.csv'

- reading the dataset using pandas and displaying top 5 dataset

```
df = pd.read_csv(path)
df.head()
```

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- displaying bottom 5 dataset

```
df.tail()
```

- Checking the information of the data in the datset

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102375 entries, 0 to 10237
Data columns (total 5 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   Date        102375 non-null  objec
 1   Country     102375 non-null  objec
 2   Confirmed   102375 non-null  int64
 3   Recovered   102375 non-null  int64
 4   Deaths      102375 non-null  int64
dtypes: int64(3), object(2)
memory usage: 3.9+ MB
```

- converting data into date-time format

```
df = pd.read_csv(path,parse_dates=['Date'])
df.head()
```

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- df.info()
- df.head()

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- adding new column Active by calculating active

```
active = df['Confirmed'] - df['Recovered'] - df['Deaths']
df['Active'] = active
df.tail()
```

	Date	Country	Confirmed	Recovered	Deaths	Active
102370	2021-06-25	Zimbabwe	45217	37604	1721	5892
102371	2021-06-26	Zimbabwe	46018	37761	1725	6532
102372	2021-06-27	Zimbabwe	46442	37817	1736	6889
102373	2021-06-28	Zimbabwe	47284	37949	1749	7586
102374	2021-06-29	Zimbabwe	48533	38323	1761	8449

- India

```
df_india = df[df['Country']=='India']
df_india.tail()
```

	Date	Country	Confirmed	Recovered	Deaths	Active
41995	2021-06-25	India	30183143	29193085	394493	595565
41996	2021-06-26	India	30233183	29251029	395751	586403
41997	2021-06-27	India	30279331	29309607	396730	572994
41998	2021-06-28	India	30316897	29366601	397637	552659
41999	2021-06-29	India	30316897	29366601	397637	552659

- df_india.groupby('Date')['Confirmed'].sum().reset_index()

	Date	Confirmed
0	2020-01-22	0
1	2020-01-23	0
2	2020-01-24	0
3	2020-01-25	0
4	2020-01-26	0
...	...	...
520	2021-06-25	30183143
521	2021-06-26	30233183
522	2021-06-27	30279331
523	2021-06-28	30316897
524	2021-06-29	30316897

525 rows × 2 columns

- confirmed=df_india.reset_index()[['Date','Confirmed','Recovered','Deaths']].rename({'Date': 'ds','Confirmed':'y','Recovered':'rec','Deaths':'d})
- Confirmed

	ds	y	rec	deaths
0	2020-01-22	0	0	0
1	2020-01-23	0	0	0
2	2020-01-24	0	0	0
3	2020-01-25	0	0	0
4	2020-01-26	0	0	0
...	...	...	...	...
520	2021-06-25	30183143	29193085	394493
521	2021-06-26	30233183	29251029	395751
522	2021-06-27	30279331	29309607	396730
523	2021-06-28	30316897	29366601	397637
524	2021-06-29	30316897	29366601	397637

525 rows × 4 columns

- train=confirmed[(confirmed['ds']>='2020-01-22') & (confirmed['ds']<='2021-03-15')]
  test=confirmed[(confirmed['ds']>'2021-03-15')& (confirmed['ds']<='2021-06-26')]
- train.shape
  (419, 4)
- test.shape
  (103, 4)
- test.head()

	ds	y	rec	deaths
419	2021-03-16	11438734	11045284	159044
420	2021-03-17	11474605	11063025	159216
421	2021-03-18	11514331	11083679	159370
422	2021-03-19	11555284	11107332	159558
423	2021-03-20	11599130	11130288	159755

- from fbprophet import Prophet
  a library in python created by facebook for the time series analysis (forecasting)

- Building the Model
  m = Prophet(interval_width=0.95)
- m.add_regressor('rec',standardize=False)
  m.add_regressor('deaths',standardize=False
- m.fit(train)
- m.params
- future=m.make_future_dataframe(periods=103)
  future.tail()
- future['rec']=confirmed ['rec']
- future['deaths']=confirmed ['deaths']
- future

- forecast = m.predict(future)
  forecast[['ds','yhat','yhat_lower','yhat_upper']].tail()

	ds	yhat	yhat_lower	yhat_upper
517	2021-06-22	2.100418e+07	1.950874e+07	2.264488e+07
518	2021-06-23	2.105136e+07	1.953040e+07	2.273277e+07
519	2021-06-24	2.109636e+07	1.955807e+07	2.279520e+07
520	2021-06-25	2.114036e+07	1.955610e+07	2.285477e+07
521	2021-06-26	2.118339e+07	1.957690e+07	2.290631e+07

- fig1=m.plot(forecast)

- fig2=m.plot_components(forecast)

- from fbprophet.diagnostics import cross_validation, performance_metrics
  df_cv=cross_validation(model=m,initial='180 days', horizon='90 days')
  df_p = performance_metrics(df_cv)
  df_p

	horizon	mse	rmse	mae	mape	mdape	coverage
0	9 days	8.179244e+09	9.043917e+04	7.072691e+04	0.016646	0.008169	0.25
1	10 days	1.064061e+10	1.031533e+05	8.165934e+04	0.019118	0.008623	0.25
2	11 days	1.327492e+10	1.152168e+05	9.207960e+04	0.021345	0.010882	0.25
3	12 days	1.639482e+10	1.280423e+05	1.025836e+05	0.023550	0.013524	0.25
4	13 days	1.989785e+10	1.410597e+05	1.133672e+05	0.025736	0.017884	0.25
...	...	...	...	...	...	...	...
77	86 days	1.520913e+12	1.233253e+06	9.996644e+05	0.113900	0.081899	0.50
78	87 days	1.561636e+12	1.249654e+06	1.010505e+06	0.114691	0.082066	0.50
79	88 days	1.602062e+12	1.265726e+06	1.020902e+06	0.115438	0.082140	0.50
80	89 days	1.640890e+12	1.280972e+06	1.030596e+06	0.116105	0.082177	0.50
81	90 days	1.678312e+12	1.295497e+06	1.039673e+06	0.116702	0.082325	0.50

82 rows × 7 columns

- from fbprophet.plot import plot_cross_validation_metric
  fig3=plot_cross_validation_metric(df_cv,metric='mape')

Chapter 7

RESULTS AND DISCUSSION

In the analysis of worldwide dataset, various countries active cases are visualized on the world map using plotly library. Purple color has been divided into 2000 samples, where dark shade represent the more active number of cases and light shade represents less number of active cases.

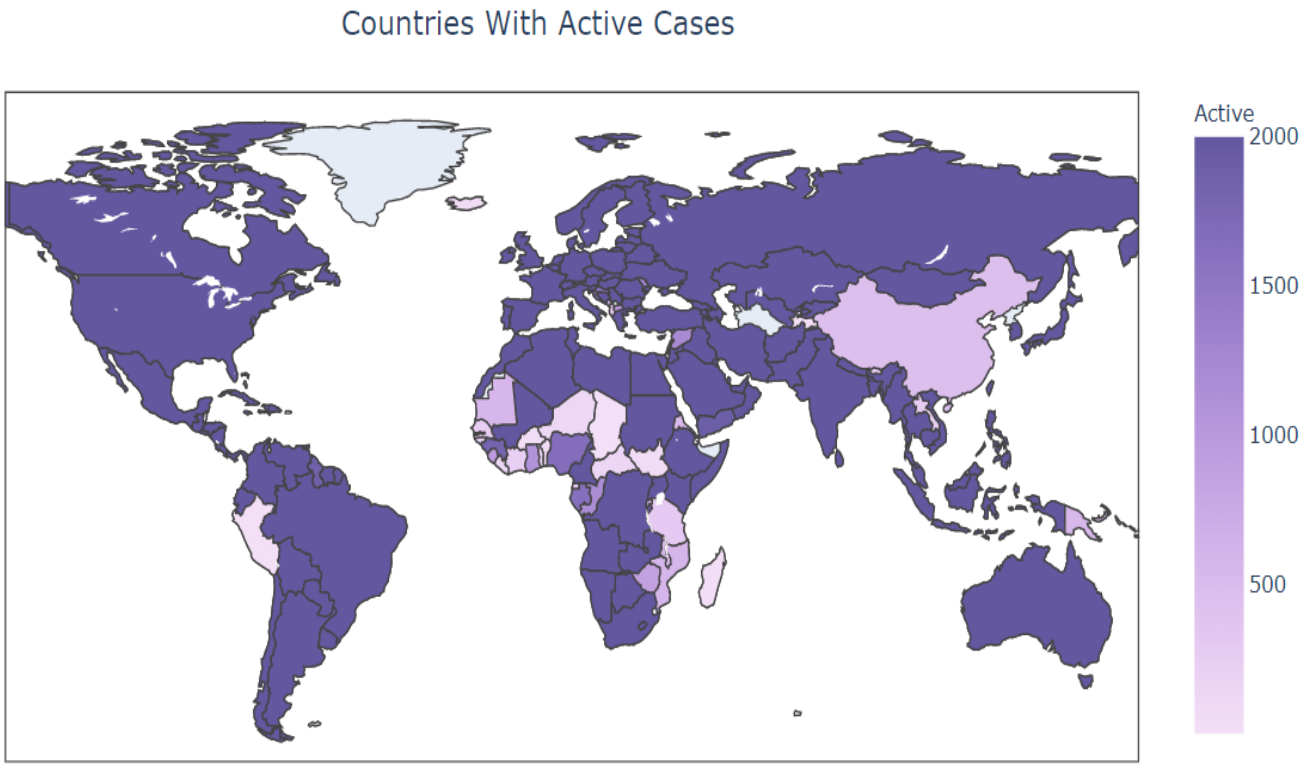


Fig 7.1 Total Active Cases plotted on World map

Analysis of the worldwide dataset the countries confirmed cases are visualized on the world map using plotly library. Red color has been divided into 2000 samples, where dark shade represent the more confirmed number of cases and light shade represents less number of confirmed cases.

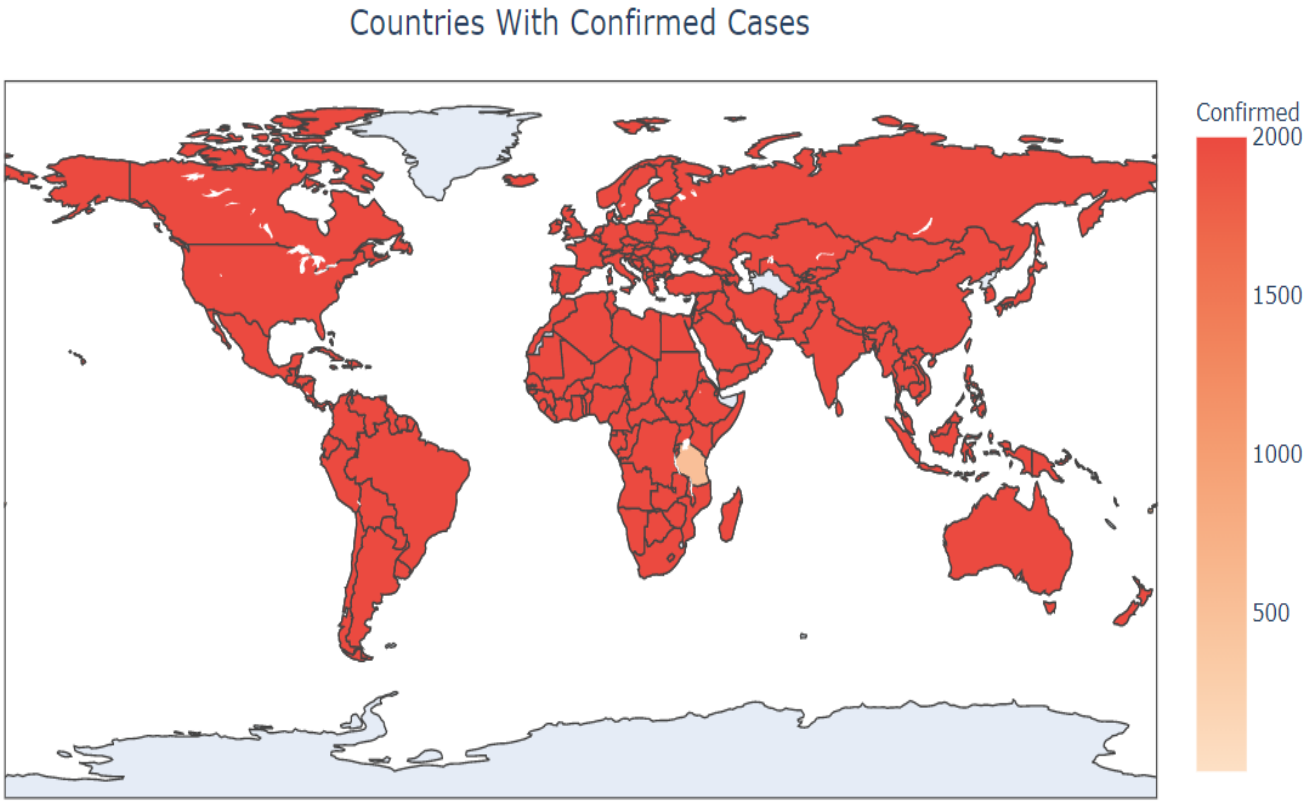


Fig 7.2 Total Confirmed Cases plotted on World map

Analysis of the worldwide dataset the countries confirmed cases are plotted on the world map using plotly library. Green color has been divided into 2000 samples, where dark shade represent the more recovered number of cases and light shade represents less number of recovered cases.

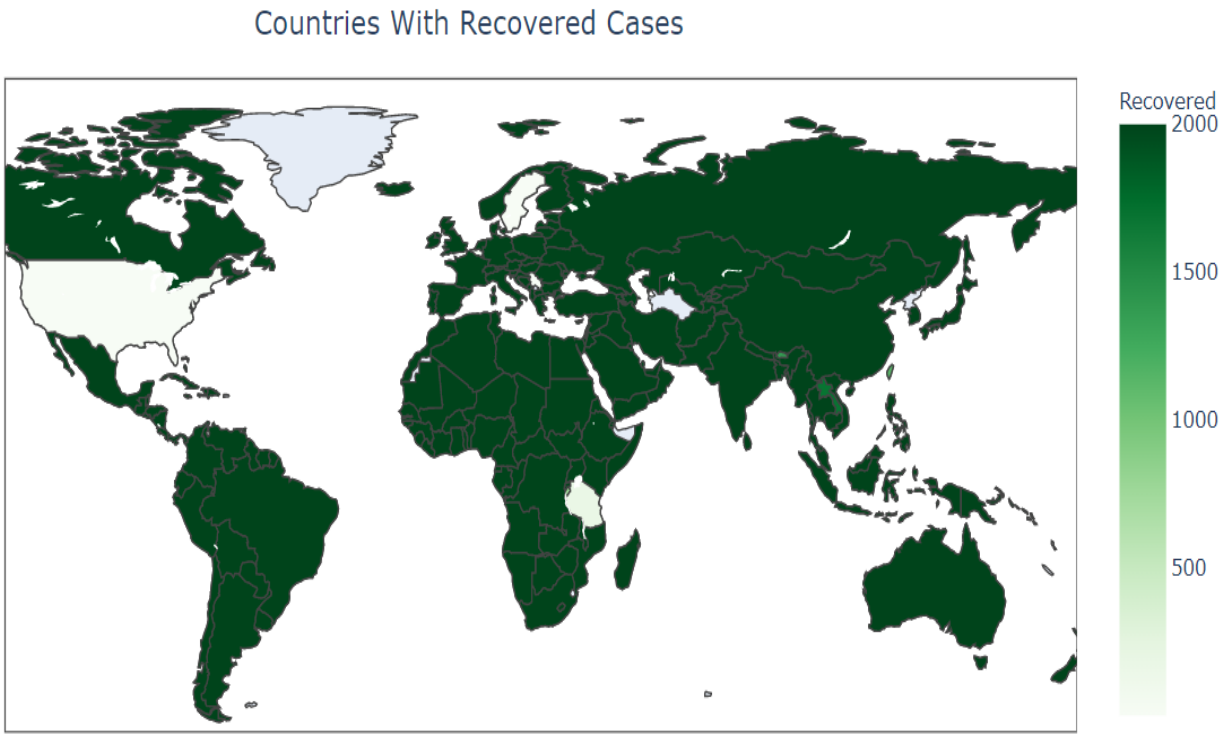


Fig 7.3 Total Recovered Cases plotted on World map

Analysis of worldwide dataset where the countries death cases are visualized on the world map using plotly library. Viridis color scale has been divided into 2000 samples, where dark shade represent the more death number of cases and light shade represents less number of death cases.

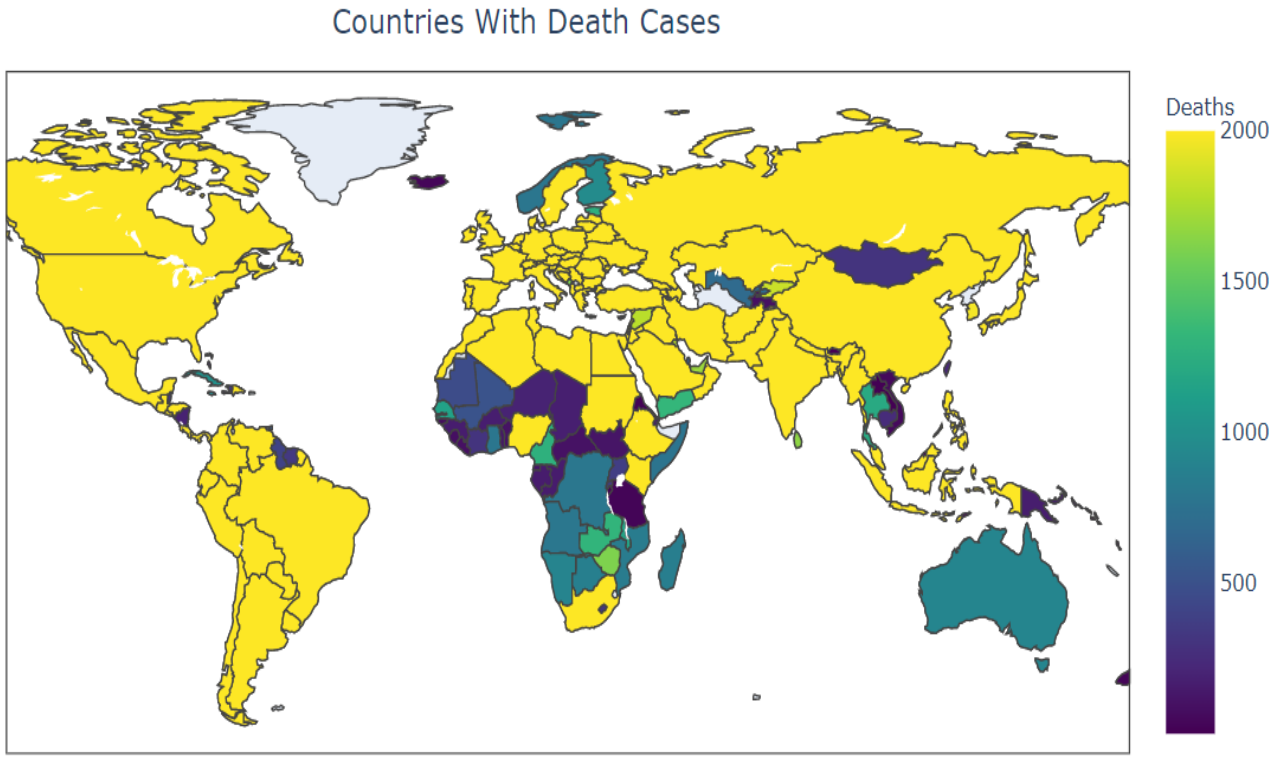
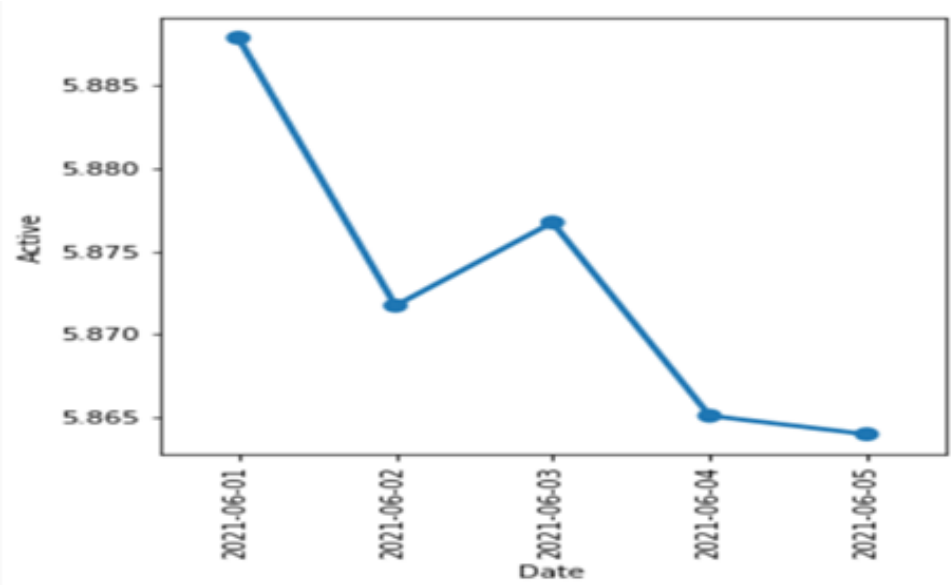


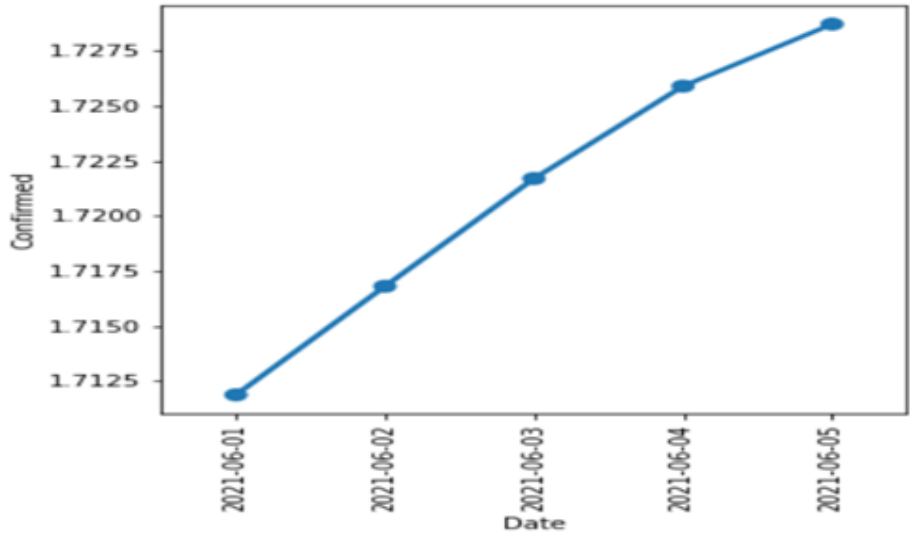
Fig 7.4 Total Death Cases plotted on World map

Plotting worldwide Active cases verses Date using point plot visual element.



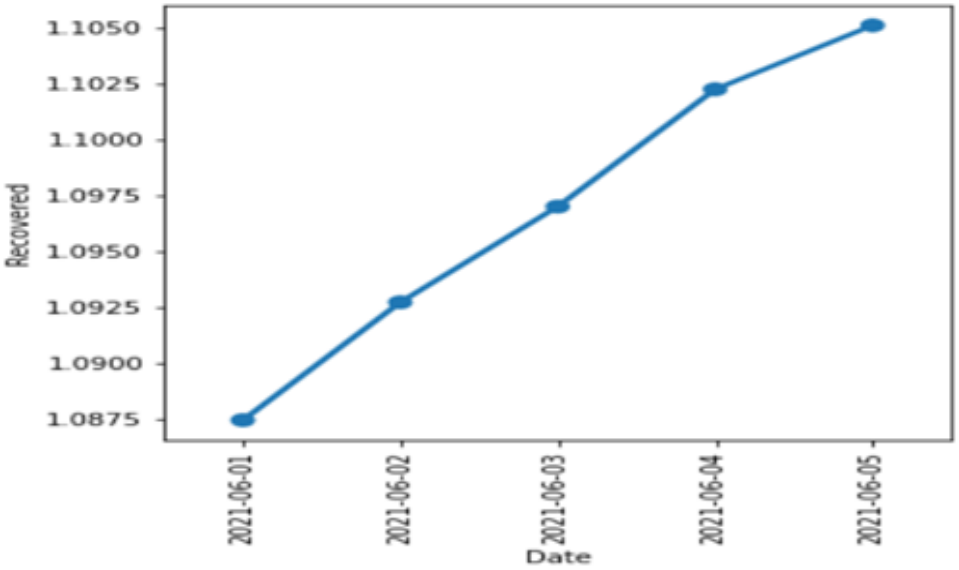
7.5 Date vs Total Active cases

Plotting worldwide Confirmed cases verses Date using point plot visual element.



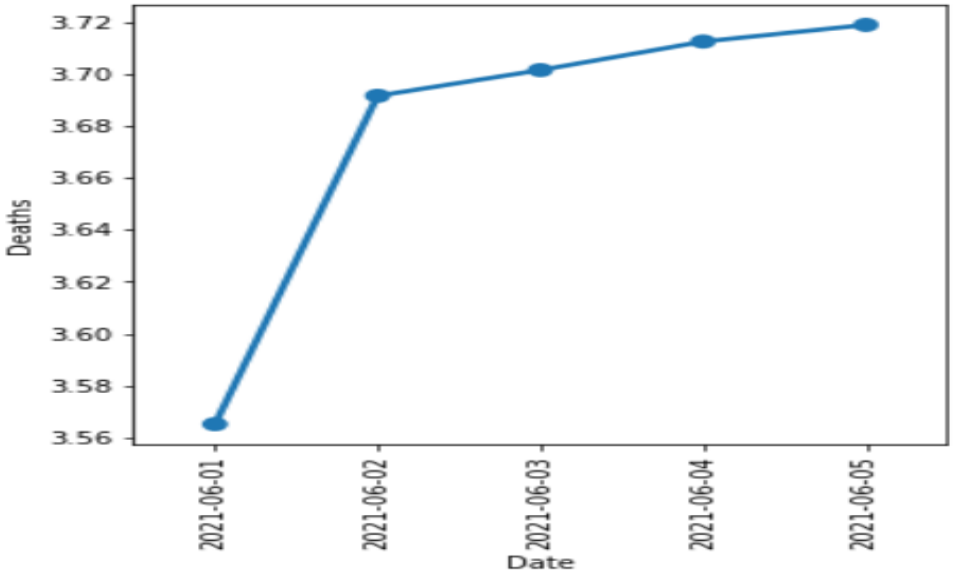
7.6 Date vs Total Confirmed cases

Plotting worldwide Recovered cases verses Date using point plot visual element.



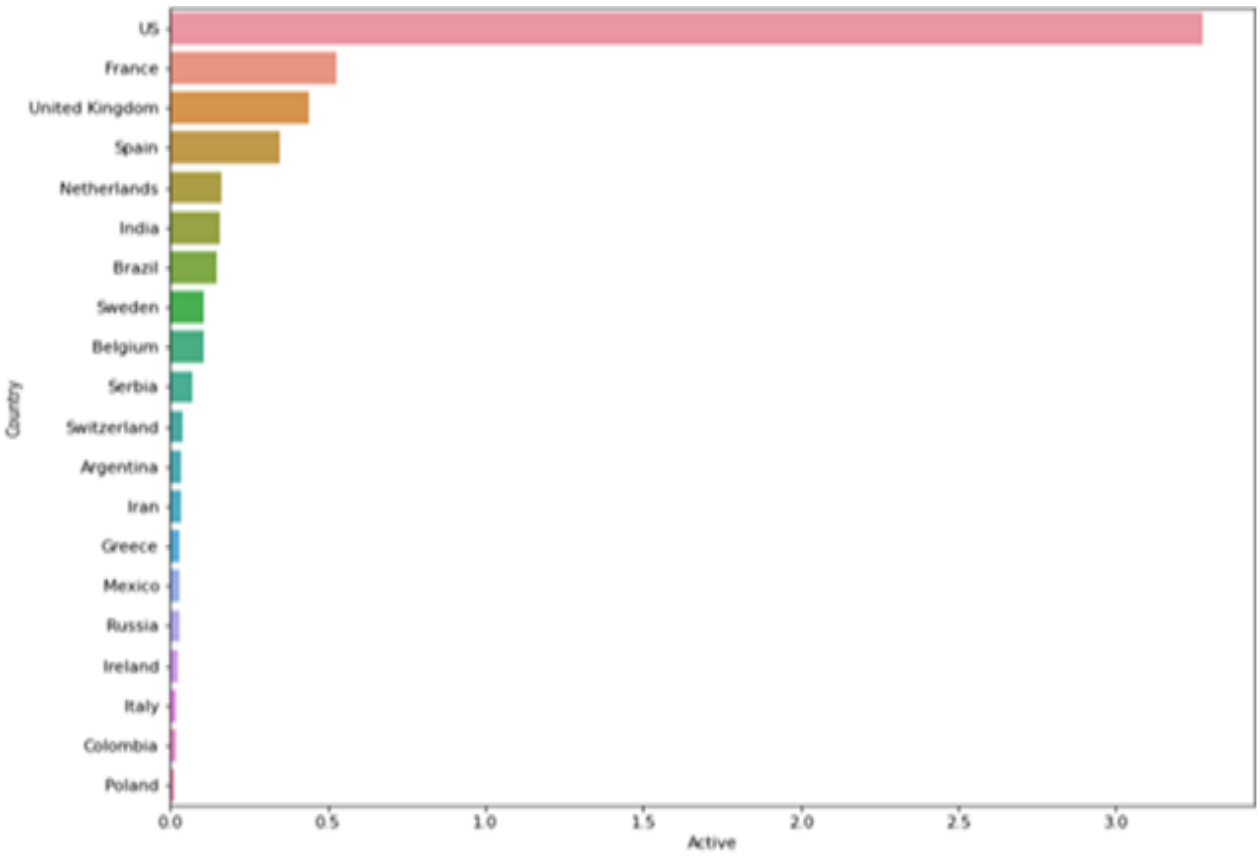
7.7 Date vs Total Recovered cases

Plotting worldwide death cases verses date on the visual element using point plot.



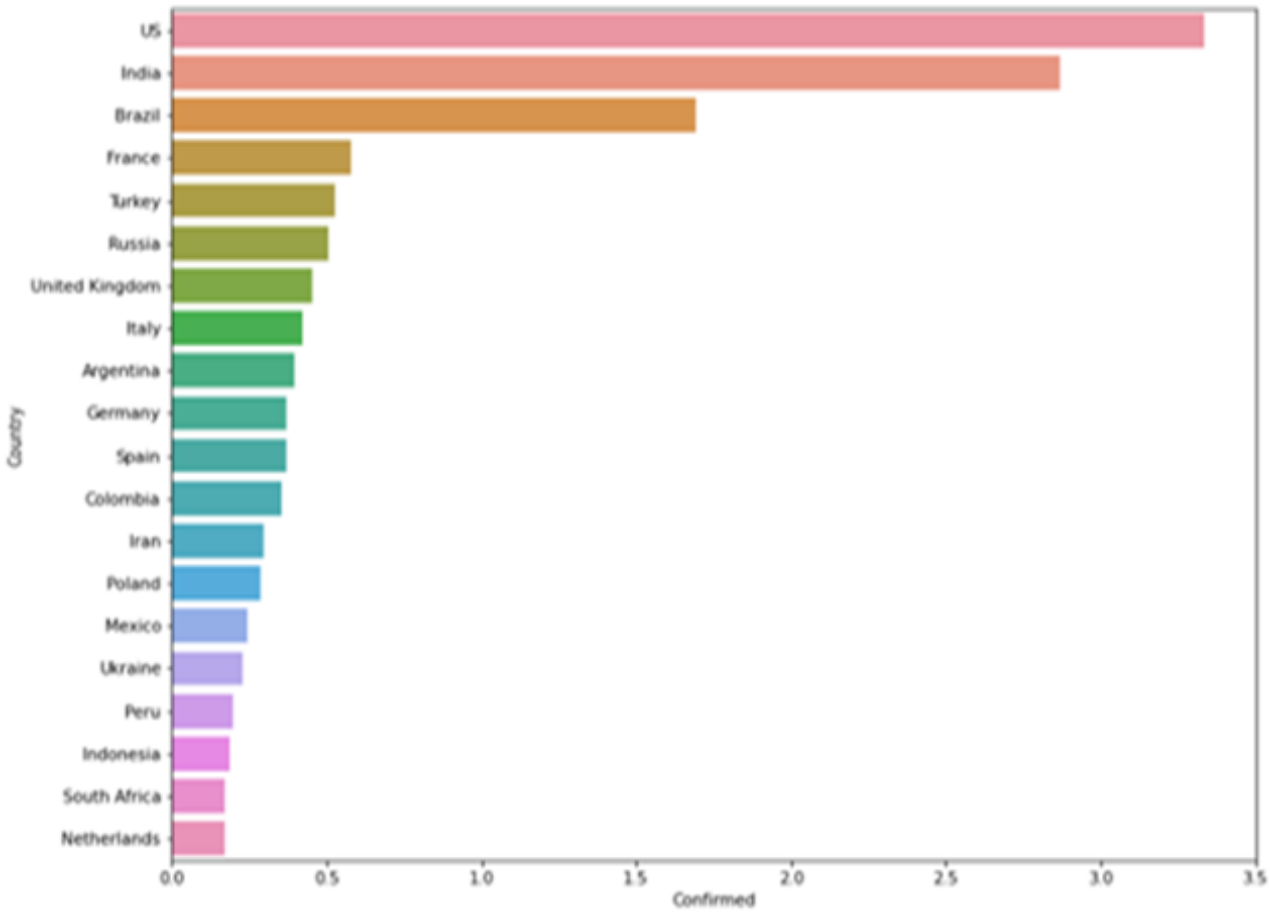
7. 8 Date vs Total Death cases

Plotting top 20 countries with maximum number of active cases on bar plot using seaborn library.



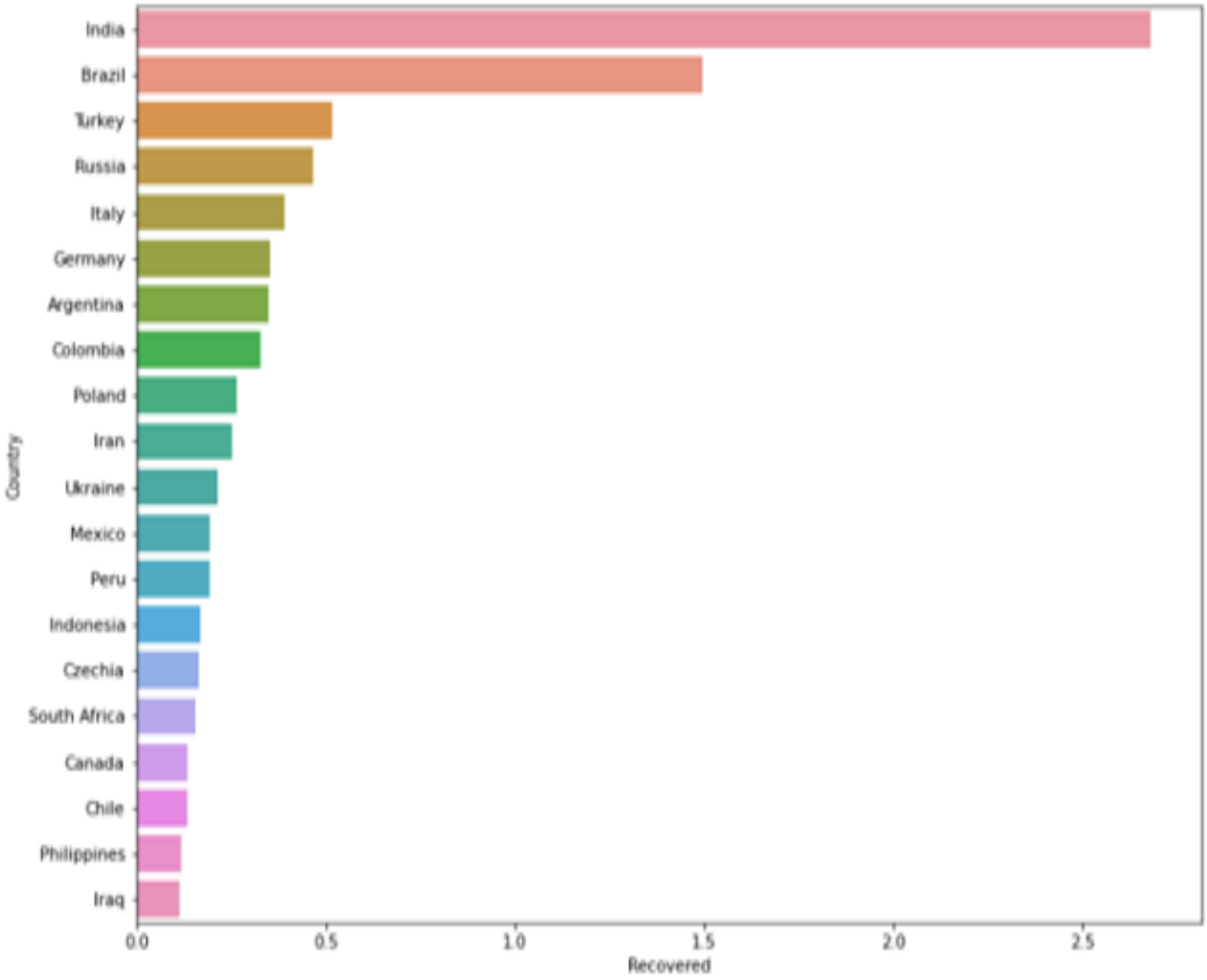
7.9 Active cases vs Top 20 countries

Plotting top 20 countries with maximum number of confirmed cases on bar plot using seaborn library.



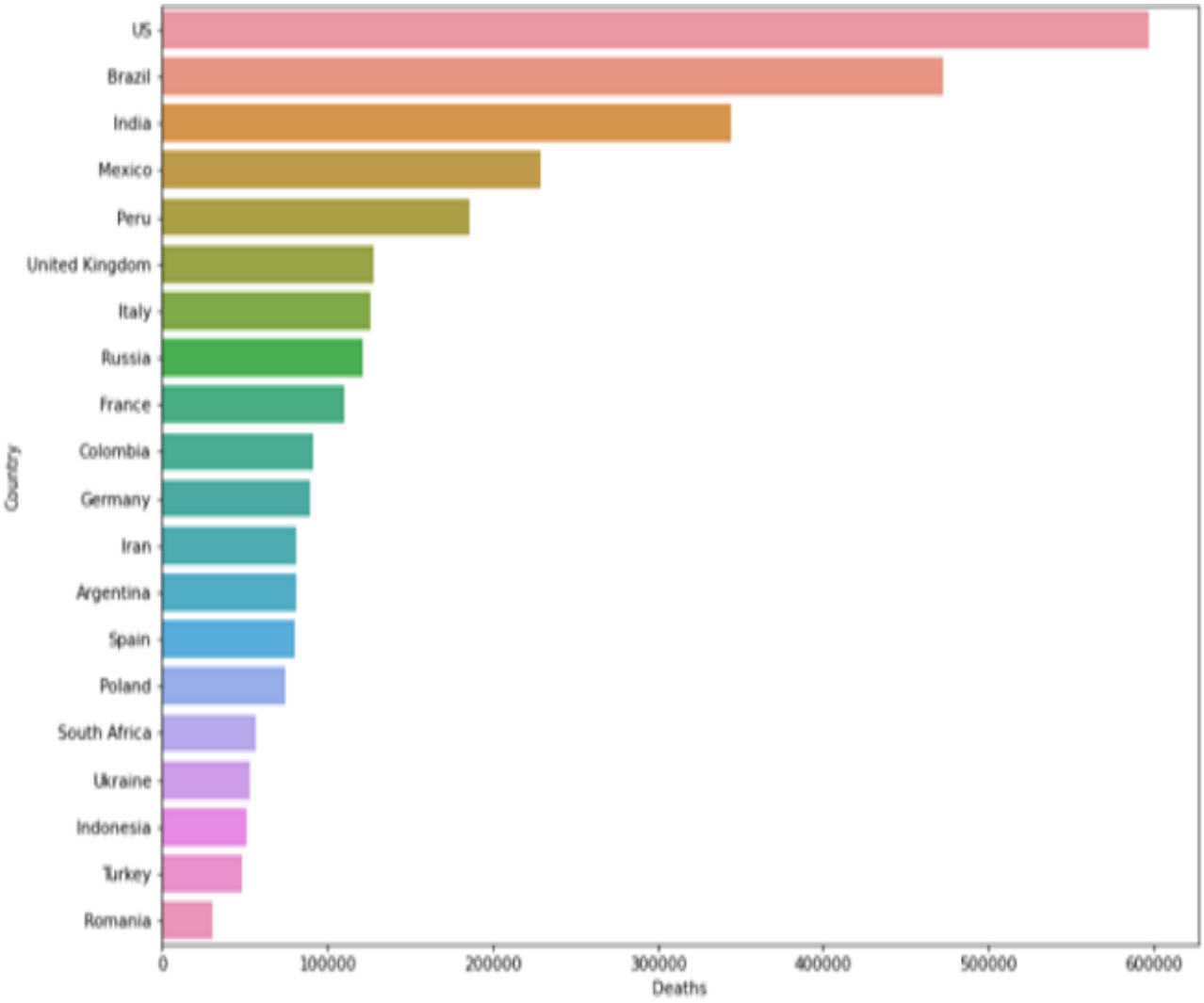
7.10 Confirmed cases vs Top 20 countries

Plotting top 20 countries with maximum number of recovered cases on bar plot using seaborn library.



7.11 Recovered cases vs Top 20 countries

Plotting top 20 countries with maximum number of death cases on bar plot using seaborn library.



7.12 Death cases vs Top 20 countries

Graphical representation of the predicted result of the COVID19 confirmed cases by Linear Model. Here y refers to the confirmed cases and ds represents the date. Confirmed cases over Date from January 2020 to July 2021.

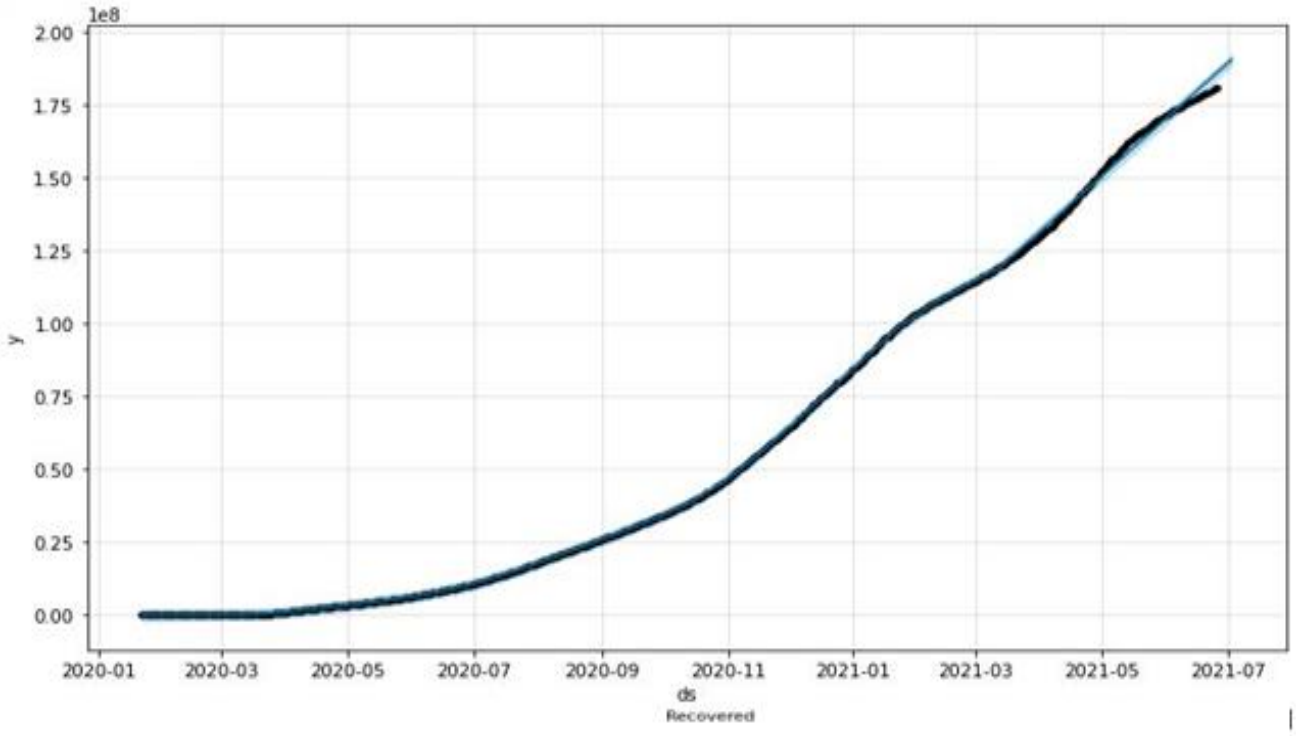


Fig 7.13 Date vs Confirmed cases

Different Trend components of corona virus cases over the date have been represented graphically. Linear trend is observed over the date. On considering weekly trend Sunday and Friday has more confirmed cases.

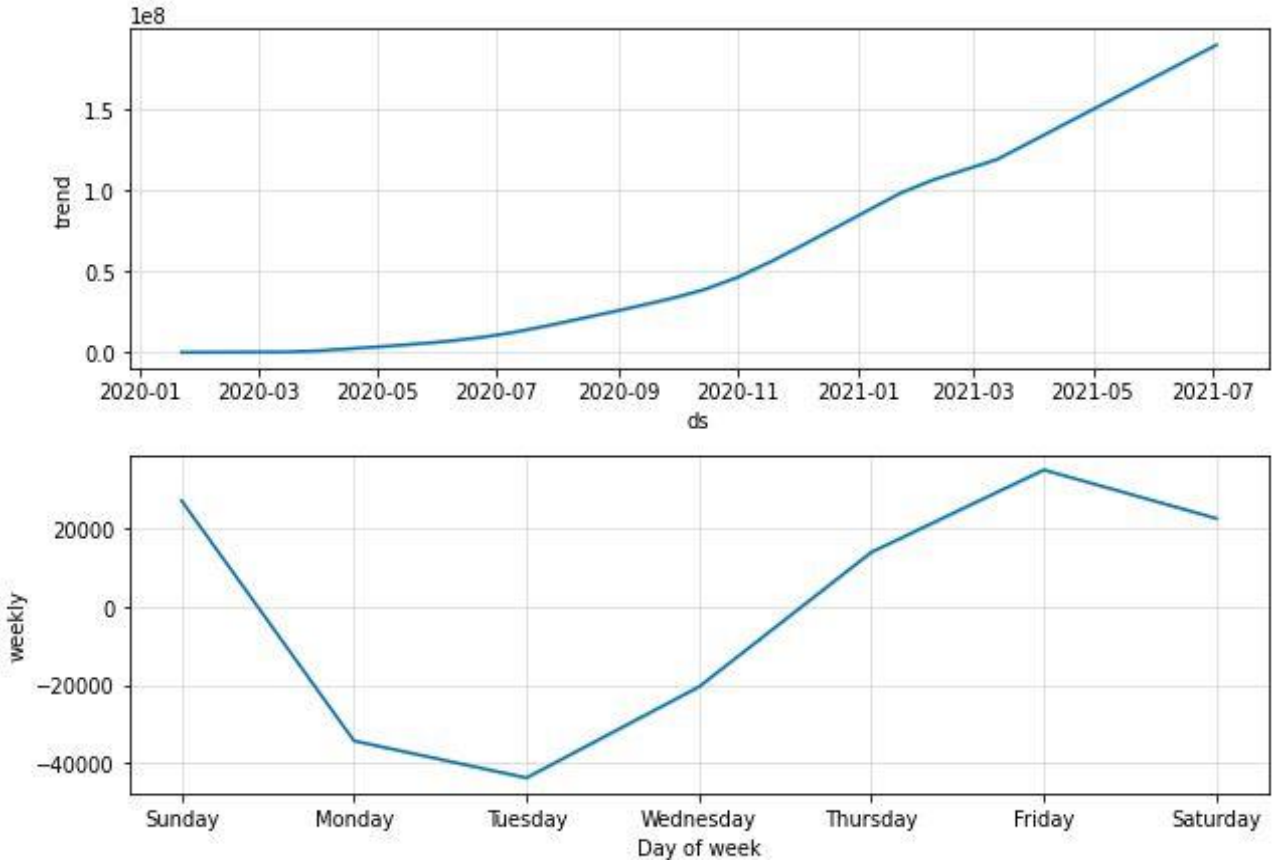


Fig 7.14 Components

Graphical representation of the predicted result of the COVID19 confirmed cases by Multivariate Model. Here y refers to the confirmed cases and ds represents the date. Confirmed cases over Date from February 2020 to June 2021.

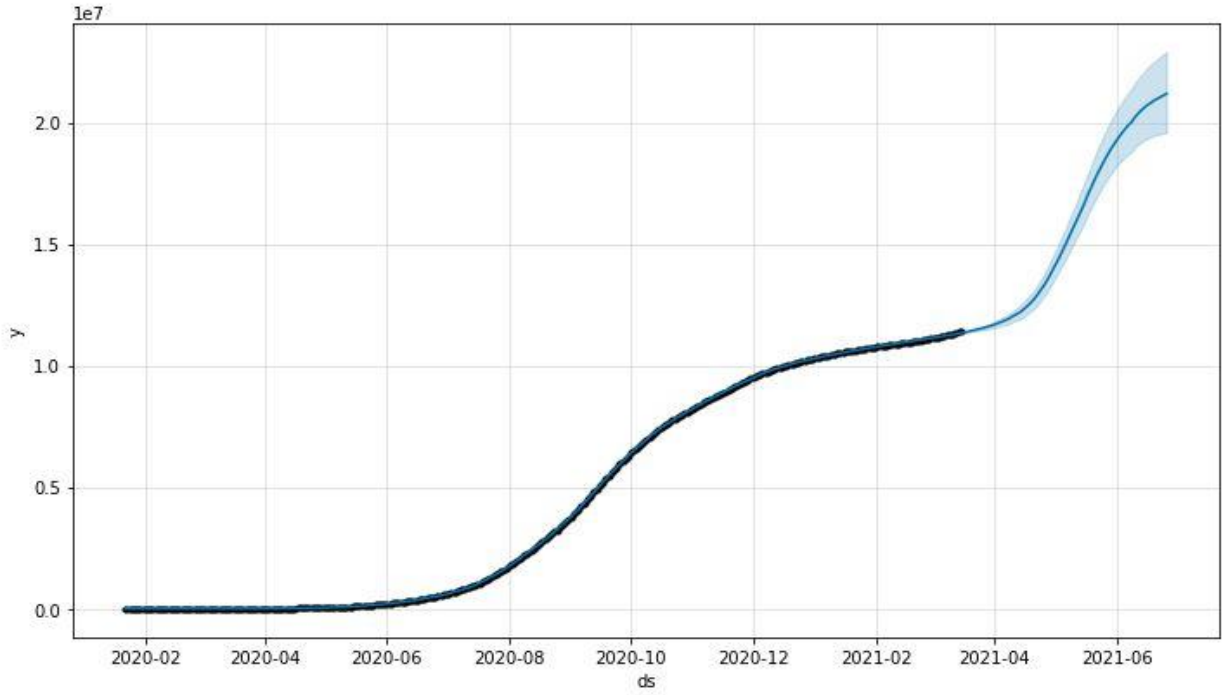


Fig 7.15 Multivariate Plot

Different Trend components of corona virus cases over the date have been represented graphically. Considering weekly trend Sunday and Monday has more confirmed cases.

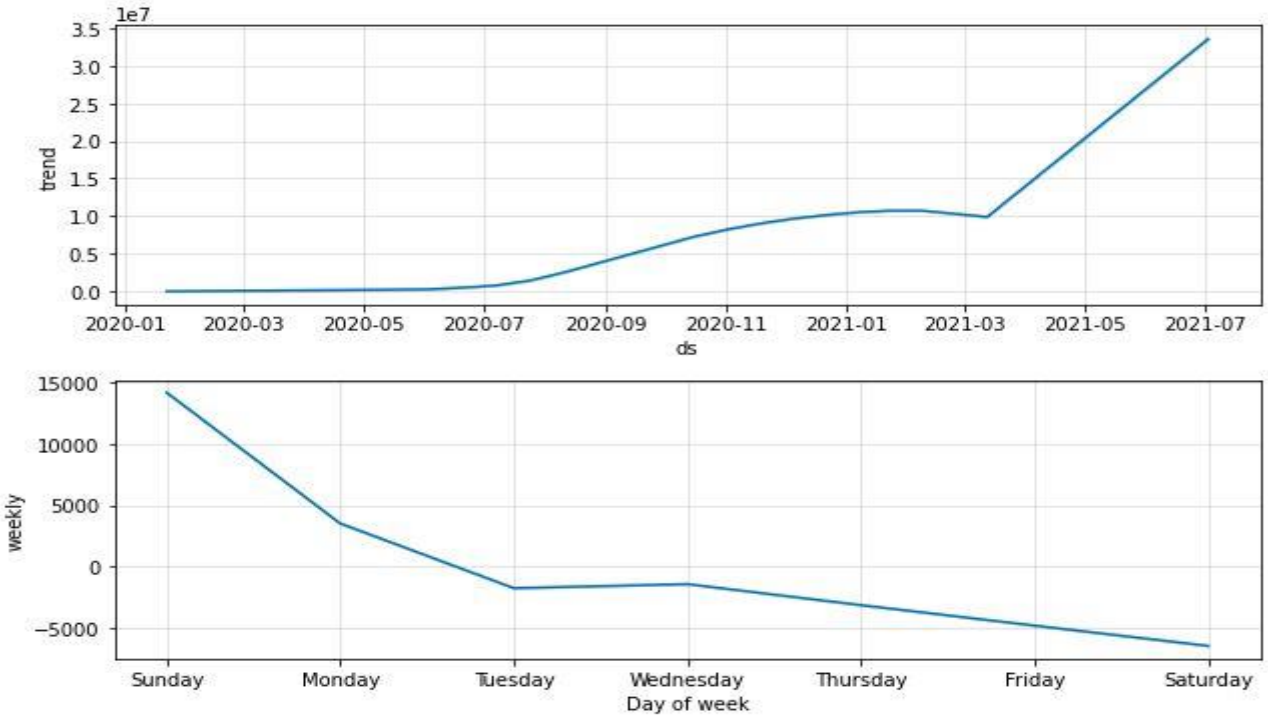


Fig 7.16 Multivariate Component

CONCLUSION AND FUTURE WORK

In this project, the analysis of COVID-19 data is performed and the pandemic spread is compared between different countries. Data Visualization techniques are applied to give a clear look on the trend of the data, that how the virus is spreading, which countries are getting affected mostly and how different countries are recovering. On applying Linear Regression Model and Multivariate Regression Model the possible number of COVID-19 cases are obtained. 37120600 and 34143420 are the upper and lower bound of the linear model. Through this analysis, it has shown that the number of cases are going to decrease in the coming days. Accuracy of the model has been calculated for both the models using cross validation function in the prophet model. Among Linear and Multivariate Regression Models, linear model has performed well and has shown the better accuracy than multivariate model. By considering different or more features multivariate model can be improvised.

REFERENCES

1. Ashutosh kumar, *"COVID-19:Analysis, Prediction, Plotting"*, June 2020,ResearchGate.
2. Sina F. Ardabili, Amir Mosavi, PedramGhamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy,Uwe Reuter, Timon Rabczuk, Peter M. Atkinson , "*COVID-19 Outbreak Prediction with Machine Learning",* March 2020,ResearchGate.
3. Mohammadreza Nemati, Jamal Ansary, Nazafarin Nemati, *"Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data"*,4 July 2020, Science Direct.
4. R. Sujath, Jyotir Moy Chatterjee & Aboul Ella Hassanien, "*A machine learning forecasting model for COVID-19 pandemic in India*", July 2020 , Springer link.
5. Rajan Gupta, Gaurav PandeyPoonam Chaudhary, Saibal K. Pal, "*Machine Learning Models for Government to Predict COVID-19 Outbreak* ", Aug 2020, ACM digital library.
6. Othman Istaiteh, Tala Owais, Nailah Al-Madi Saleh Abu-Soud, *"Machine Learning Approaches for COVID-19 Forecasting",*2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA).
7. Dr. Vakula Rani J, Aishwarya Jakka, *"Forecasting COVID-19 cases in India Using Machine Learning Models"* , July 2021, International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE 2020).
8. Saud Sheikh, Jaini Gala, Aishita Jain, Sunny Advani, Sagar Jaidhara, *"Analysis and Prediction of Covid-19 using Regression models and Time series Forecasting",* 2021, 11th International Conference on Cloud computing, Data Science and Engineering(Confluence 2021).
9. Ovi Sarkar, Md Faysal Ahamed, Pallab Chowdhury, *"Forecasting & Severity Analysis of COVID-19 Using Machine Learning Approach with Advanced Data Visualization"* , Dec 2020, 23rd International Conference on Computer and Information Technology (ICCIT).

10. Furqan Rustam, Aijaz Ahmad Reshi 2, Arif Mehmood,Saleem Ullah ,Byung-Won On, Waqar Aslam,And Gyu Sang Choi, *"COVID-19 Future Forecasting Using Supervised Machine Learning Models"* ,June 2020, National Research of Korea (NRF) grant funded by Korea government (MSIT).
11. Zhihao Yang,Kang'an Chen, *"Machine Learning Methods on COVID-19 Situation Prediction"* ,2020 ,International Conference on Artificial Intelligence and Computer Engineering (ICAICE).
12. Vartika Bhadana, Anand singh jalal, Pooja pathak, *"A comparitive study of Machine Learning methods for Covid-19 prediction in India"*,2020.
13. Deepak Painuli, Divya Mishra, Suyash Bhardwaj,Mayank Aggarwal, *"Forecast and prediction of COVID-19 using machine learning"* ,2021, Elseveir Public health Emergency Collection.
14. Md. Shahriare Satu, Koushik Chandra Howlader, Mufti Mahmud, M. Shamim Kaiser ,Sheikh Mohammad Shariful Islam , Julian M. W. Quinn , Salem A. Alyami and Mohammad Ali Moni , *" Short-Term Prediction of COVID-19 Cases Using MachineLearning Models"* , 2021, MDPI
15. Sina F. Ardabili,Amir Mosavi,Pedram Ghamisi,Filip Ferdinand,Annamaria R.Varkonyi-Koczy, Uwe Reuter,Timon Rabczuk,Peter M. Atkinson, *" COVID-19 Outbreak Prediction with Machine Learning"*, 2020.
16. Senthilkumar Mohan, John A, Ahed Abugabah, Adimoolam M, Shubham Kumar Singh, Ali kashif Bashir, Louis Sanzogni, *"An approach to forecast impact of Covid-19 using supervised machine learning model"* ,2021.
17. Sina F. Ardabili,Amir Mosavi,Pedram Ghamisi 5,Filip Ferdinand ,Annamaria R. Varkonyi-Kocz,Uwe Reuter ,Timon Rabczuk and Peter M. Atkinson , *"COVID-19 Outbreak Prediction with Machine Learning"* ,Septemper 2020,MDPI
18. Iman Rahimi, Fang Chen & Amir H. Gandomi, *" A review on COVID-19 forecasting models"* ,2021,Neural Computing and Applications.

19. Abdel kader,Dairia Fouzi ,Harroub Abdelhafid ,Zeroualcd Mohamad Mazen Hittaweb,Ying Sun, *"Comparative study of machine learning methods for COVID-19 transmission forecasting"* ,June 2021, Elsevier.
20. Rohini, Naveena, Jothipriya.G3, Kameshwaran, Jagadeeswari.M, *"A Comparative Approach To Predict Corona Virus Using Machine Learning"* ,2021,International Conference on Artificial Intelligence and Smart Systems (ICAIS).
21. G.Monika, Dr. M.Bharathi Devi, *"Using Machine Learning Approach to Predict Covid-19 Progress"*, 2020, International Journal for Modern Trends in Science and Technology.
22. Ekta Gambhir,Ritika Jain, Alankrit Gupta, Uma Tomer, *" Regression Analysis of COVID-19 using Machine Learning Algorithms"* , 2020, Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020).