ABSTRACT

Novel Corona Virus (COVID-19 or 2019-nCoV) pandemic has neither clinically proven vaccine nor drugs; however, its patients are recovering with the aid of antibiotic medications, anti-viral drugs, and chloroquine as well as vitamin C supplementation. It is now evident that the world needs a speedy and quicker solution to contain and tackle the further spread of COVID-19 across the world with the aid of non-clinical approaches such as data mining approaches, augmented intelligence and other artificial intelligence techniques so as to mitigate the huge burden on the healthcare system while providing the best possible means for patients' diagnosis and prognosis of the 2019-nCoV pandemic effectively.

In this project, data mining models were developed for the prediction of COVID-19 infected patients' recovery and prevalence using epidemiological dataset of COVID-19. Machine Learning algorithms such as Linear Regression, Support Vector Machine Regression, Polynomial Regression, ARIMA and FB Prophet were applied directly on the dataset using python programming language to develop the models.

The models forecast the number of Confirmed, Recovered and Death cases for the next week. Predicting the prevalence and incidence of this disease throughout the world is crucial to helping health professionals make key decisions about the disease.

Keywords **–** COVID-19, Linear Regression, Support Vector Regression, Polynomial regression, ARIMA, Prophet, RMSE, MAE, MAPE.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1. Introduction:

Severe Acute Respiratory Syndrome Coronavirus Two (SARS-CoV-2), the causative agent of novel coronavirus (COVID-19 or 2019-nCoV), has emerged in late 2019 which is believed to be originated from Hubei Province, China called Wuhan. 2019-nCoV or COVID-19 is rapidly spreading in humans. The major symptoms of SARS-CoV-2 include fever, cough, and shortness of breath which in many instances appeared to be similar to that flu. COVID-19 had since reached a decisive point and pandemic potential which claimed the lives of many people across the world and human-to-human transmission of COVID-19 from infected individuals with mild symptoms have been reported. However, there is no drug or vaccine clinically proven to treat COVID-19 pandemic, therefore other non-clinical or non-medical therapeutic techniques are urgently needed to contain and prevent further outbreak of COVID-19 pandemic such as data mining techniques, machine learning and expert system among other artificial intelligence techniques.

Data mining (DM) is an advanced artificial intelligence (AI) technique that is used for discovering novel, useful, and valid hidden patterns or knowledge from dataset. The technique reveals relationships and knowledge or patterns among the dataset in several or single datasets. It has also widely used for the prognosis and diagnosis of many diseases including Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) that were so far discovered in 2003 and 2012, respectively. As huge dataset generated around the world related to 2019-nCoV pandemic every day is a treasured resource to be mined and analysed for useful, valid, and novel knowledge or patterns extraction for better decision-making to contain the outbreak of COVID-19 pandemic. In the healthcare sector, data mining has been widely applied in many different applications such as predicting patient outcomes, modeling health outcomes, hospital ranking, and evaluation of treatment effectiveness and infection control, stability, and recovery.

In this study, we develop several data mining models for the prediction of 2019-nCoV-infected patients' recovery. The models predict when COVID-19 infected patients would be recovered and released from isolation centers as well as patients that may likely not be recovered and lost their lives to COVID-19 pandemic. The models help the health workers to determine the recovery and stability of the newly infected persons with pandemic COVID-19. Data mining algorithm which includes linear regression, support vector machine regression, etc. were applied directly on the dataset using python programming language to develop the models.

1.2. Data Mining:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. The knowledge or extracted information can be used to predict results in near future based on discovered patterns. Data mining is the analysis step of the "knowledge discovery in databases" process or KDD. Aside from the raw analysis

step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

1.2.1. Types of Data that can be Mined?

Data mining can be performed on the following types of data:

- **Relational Database:** A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.
- **Data warehouses:** A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.
- **Data Repositories:** The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.
- **Object-Relational Database:** A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc. One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.
- **Transactional Database:** A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

1.3. Machine Learning:

Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given. Machine Learning is a field which is raised out of Artificial Intelligence (AI). Applying AI, we wanted to build better and intelligent machines. But except for few mere tasks such as finding the shortest path between point A and B, we were unable to program more complex and constantly evolving challenges. There was a realisation that the only way to be able to achieve this task was to let machine learn from itself. This sounds similar to a child learning from its self. So, machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of

technology, that did not even realise it while using it. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to carry out tasks without explicit instructions, such as by using pattern recognition and inference. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical 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 email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. 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. Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. For the early tasks that humans wanted computers to accomplish, it was possible to create algorithms that enabled the machine to execute all the steps needed to solve the problem in hand. So on the computer's part, no learning was needed. For certain advanced tasks, facial recognition for example, it is not easy to create the needed algorithms, partly as it's not easy for humans to precisely define how we recognise faces. Abundant face related data exists however. So far, compared to the difficulty in directly creating the required algorithms, it's turned out in practice to be easier to assist computers to learn themselves how to recognise faces from available data. The discipline of machine learning develops various approaches for computers to learn to accomplish tasks for which no algorithm exists.

1.3.1. Regression In ML:

Regression is a statistical technique that helps in qualifying the relationship between the interrelated economic variables. The first step involves estimating the coefficient of the independent variable and then measuring the reliability of the estimated coefficient. This requires formulating a hypothesis, and based on the hypothesis, we can create a function. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. To exemplify, given data about the size of houses on the real estate market, try to predict their price. In the same way regression algorithms were used to predict the COVID-19 trends for next coming week.

1.3.2. Time Series Forecasting:

Making predictions about the future is called extrapolation in the classical statistical handling of time series data. More modern fields focus on the topic and refer to it as time series forecasting. Forecasting involves taking models fit on historical data and using them to predict future observations. Descriptive models can borrow for the future (i.e., to smooth or remove noise), they only seek to best describe the data. An important distinction in forecasting is that the future is completely unavailable and must only be

estimated from what has already happened. The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanisms that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series The skill of a time series forecasting model is determined by its performance at predicting the future. This is often at the expense of being able to explain why a specific prediction was made, confidence intervals and even better understanding the underlying causes behind the problem.

1.4. Problem Statement:

In this project, predictive analysis on Covid-19 datasets of confirmed, recovered, and death cases individually will be performed using data mining and ML algorithms. These datasets consist the numerical data regarding the Covid-19 disease out-spread. There will be construction of different statistical representations for the data given in the data set. Prevalence of the Covid-19 will be analyzed and the respective cases will be forecasted for a week using time series forecasting and regression models.

1.5. Motivation of the Work:

The high prevalence Covid-19 has made it a new pandemic. Covid-19 requires a global and integrated response of all national medical and healthcare systems. Covid-19 exposed the need for timely response and data sharing on fast spreading global pandemics. It is a new virus with its own characteristics. The Covid-19 virus is unique among human corona viruses which has capacity of high transmissibility, substantial fatal deaths in some high-risk groups, and ability to cause huge societal and economic disruption in the world.

The pandemic has already taken grip over peoples' life. Since the start of the pandemic, some countries are facing problem of ever-increasing cases. Through the data analysis of cases one can analyse how countries all over the world are doing in terms of controlling the pandemic. Analysing 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 individual/country/organisations, 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.

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multidimensional database, Data visualization, Soft- computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Machine Learning provides methods, techniques, and tools that can help solving diagnostic and prognostic problems in a variety of medical domains. ML is being used for the analysis of the importance of clinical parameters and of their combinations for prognosis, e.g., prediction of disease progression, for the extraction of medical knowledge for outcomes research, for therapy planning and support, and for overall patient management. ML is also being used for data analysis, such as

and Bayesian imputation to study the cases for COVID-19 in India. They have also studied the impact of social distancing and lockdown in India. They suggest sustained lockdown with periodic relaxation. Botha and Dednam (2020) developed a simple 3-dimensional iterative map model to forecast the Coronavirus disease. From their perspective lockdown measures are effective in postponing the large peak. However, if the relaxation is given there are high chances that the virus would exponentially grow. Peng, Yang, Zhang, Zhuge, and Hong (2020) examined the public data of the National Health Commission of China from January 20 to February 9,, 2020 to make predictions in five different regions.

The authors have also made an inference that Beijing and Shanghai will soon stay safe and Wuhan will turn more severe till April.

Bayes and Valdivieso (2020) constructs a predictive Bayesian nonlinear model for the number of COVID-19 deaths in Peru. Liu et al. (2020) presented a timely and novel methodology that combines disease estimates from mechanistic models with digital traces with machine learning mythologies to forecast COVID-19 activity in the Chinese province. The model was able to give forecast 2 days ahead of the current time. Shim, Tariq, Choi, Lee, and Chowell (2020) studied the growth rate of the outbreak of COVID −19 in South Korea by identifying major clusters. They studied the transmission rate and estimated the reproduction number at 1.5 on average. Their estimates supported the implementation of social distancing measures in Korea. Perc, Gorišek Miksić, Slavinec, and Stožer (2020) proposed forecasts obtained with a simple iteration method that needs only the daily values of confirmed cases as input. The results show that the daily growth rates should be kept at least below 5% plateaus are to be seen anytime soon. Perone (2020) in his article forecasted the epidemic trend throughout April using Autoregressive integrated moving average (ARIMA). This model helps in understanding basic trends by suggesting the hypothetic epidemic's inflection point and final size.

“Fatemeh Ahouz, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran & Amin Golabpour proposed a machine learning model to forecast the number of COVID-19 patients’ across the world”. “Data Modelling & Analysing Coronavirus (COVID19) Spread using Data Science & Data Analytics in Python Code” By Jatin Chaudhary utilizes the SIR model to predict the trends of COVID-19 across India.

Based on the aforesaid literature, it is observed that there are several methods to forecast COVID −19. However, this paper focuses on a simple machine learning model that forecasts the number of positive cases along with also focuses on discussing the impact it shows on the lockdown. The project proposes to comparing the COVID-19 case with other recent outbreaks and then develop a forecasting model that can help predict the number of positive cases more effectively. There are several forecasting models built using effective statistical methods. This project focuses on using machine learning methods that can handle the nonlinearity better and produce more effective results. Also, the project suggests some key guidelines appropriate to tackle the current crisis. The analysis will be helpful to policymakers and health authorities to allocate resources rightly in the next few days/weeks.

3. SYSTEM ARCHITECTURE AND WORKING

In this section, we explain our project which we are working on in detail. The study aims to analyze COVID-19 data and future forecasting the trends of COVID-19 (Confirmed cases, Death cases, and Recovered cases). The dataset used in this study was taken from the repository of GitHub provided by the "Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)". This repository contains the three datasets for confirmed cases, death cases, and the recovered cases respectively. The data are updated every day for all daily cases. The data from these datasets were merged to obtain the parameterized datasets of the world from January 22, 2020, till March 11, 2021. These were used for the data analysis and visualization module. To get better decisions from this dataset, irrelevant columns were removed. Data cleaning was performed by removing rows with zeros from the dataset. The dataset used in the prediction module is considered for 345 days from January 22, 2020, to December 31, 2021. The architecture of this research work is shown in Figure 3.1. The whole process goes according to the following steps:

1. Dataset preparation and pre-processing.
2. Perform Exploratory Data Analysis (EDA).
3. Build and Train a Model.
4. Evaluate the Model.
5. Compare different models and arrive at the best model with a good performance.
6. Perform predictions using the accurate Model.

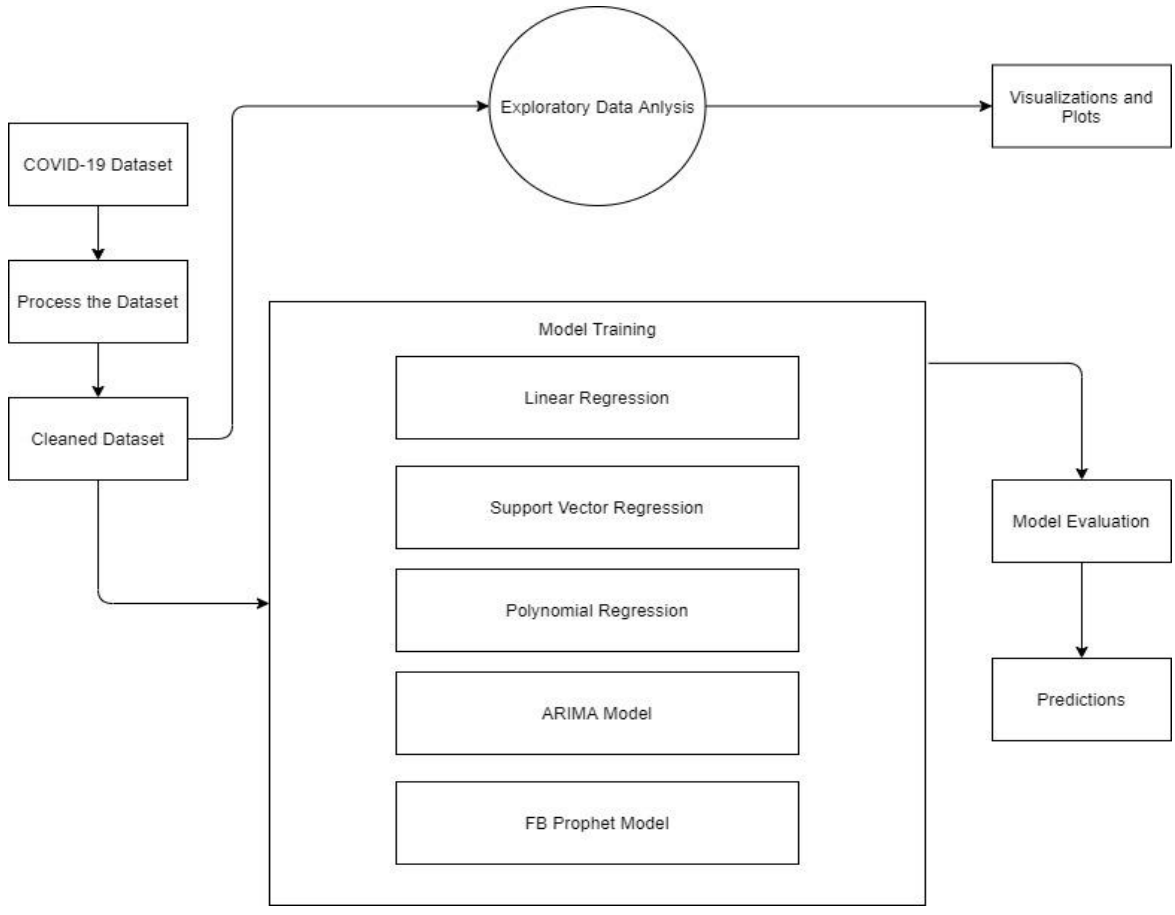

Figure 3.1. System Architecture.

3.1. Dataset preparation and data pre-processing

The data set used in this project was the real-time data from Johns Hopkins Centre for Systems Science and Engineering. It is the real time data of the patients. The data consists of the regions of the state, the country, the latitude, the symptoms, the longitude, and the number of cases for each day starting from 22/01/2020. We have considered the updated data set for the project until 11/03/2021 for EDA module and until 31/12/2020 for the Prediction module. The steps followed for pre-processing of the data is as follows:

- The data was cleansed and the null values were replaced by averaging the column. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data.
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!) or perform other exploratory analysis.
- The data was transformed using the Standard Scaler object in Python to achieve a Gaussian distribution for predicting the spread of the pandemic.
- The data were normalized using the logarithmic scale for removing the outliers.
- Split into training and evaluation sets.

3.2. Data Analysis and Data Mining:

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

Data mining is a particular data analysis technique that focuses on statistical modeling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information. In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses. Predictive analytics focuses on the application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis.

3.3. Building and Training a Model:

1. **Choose a Model:**

   Different algorithms are for different tasks. We will be choosing some from them.