Text

Description automatically generated

**Faculty of Science, Engineering and Computing**

**School of Computing and Information Systems**

**MSc in Data Science**

**Module: CI7340 Applied Data Programming**

**Individual report**

**Submitted by: Granthik Anjara**

**KU ID: K2169338**

**Date: 9th January 2023**

**Topic: Application of Machine Learning algorithms for Covid-19 predictions**

**Supervisor: David F. Wertheim**

Warranty Statement

This is a student project. Therefore, neither the student nor Kingston University makes any warranty, express or implied, as to the accuracy of the data or conclusion of the work performed in the project and will not be held responsible for any consequences arising out of any inaccuracies or omissions therein.

DECLARATION

I hereby declare that this submission is my own work towards the degree in Master of science in Software Engineering at Kingston University, London. To the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for the award of any degree of the university elsewhere, except where due acknowledgement has been made in the text.

ACKNOWLEDGEMENTS

First, I want to say thank you to my supervisor Professor David Wertheim for introducing me to this topic and for his guidance and support throughout this process. I'm so grateful for his guidance and his suggestions from the beginning until the end. I want to say thank you to Kingston University, especially the Faculty of Science, Engineering, and Computing for the services and support they have provided. I want to say thank you to the Module Leader Dr. Nada Philip for guiding me in this entire module of the project dissertation. Lastly, I want to say thank you to my family and friends for their encouragement and support.

LIST OF FIGURES

[Fig. 3.1 Machine Learning 25](https://d.docs.live.net/eb6ffc91981582a1/Desktop/Dissertation%20K2169338.docx#_Toc125414842)

[Fig. 3.2 Polynomial Regression (Selvi T., 2021) 26](https://d.docs.live.net/eb6ffc91981582a1/Desktop/Dissertation%20K2169338.docx#_Toc125414843)

[Fig. 3.3 Support Vector Machines (Alam, B., 2022) 26](https://d.docs.live.net/eb6ffc91981582a1/Desktop/Dissertation%20K2169338.docx#_Toc125414844)

[Fig. 5.1 Ascending graph of cases for all the countries 33](https://d.docs.live.net/eb6ffc91981582a1/Desktop/Dissertation%20K2169338.docx#_Toc125414845)

[Fig. 5.2 Colour gradient for top 20 country 34](#_Toc125414846)

[Fig. 5.3 Worldwide cases 35](#_Toc125414847)

[Fig. 5.4 Worldwide deaths 35](#_Toc125414848)

[Fig. 5.5 Everyday increase with moving average 36](#_Toc125414849)

[Fig. 5.6 Everyday deaths 36](#_Toc125414850)

[Fig. 5.7 India's cases 37](#_Toc125414851)

[Fig. 5.8 India's deaths 37](#_Toc125414852)

[Fig. 5.9 UK's cases 38](#_Toc125414853)

[Fig. 5.10 UK's Deaths 38](#_Toc125414854)

[Fig. 5.11 US's Cases 39](#_Toc125414855)

[Fig. 5.12 US's Deaths 39](#_Toc125414856)

[Fig. 5.13 Top 10 countries 40](#_Toc125414857)

[Fig. 5.14 India vs US vs others 40](#_Toc125414858)

[Fig. 5.15 Pie chart for top countries 41](#_Toc125414859)

[Fig. 5.16 Country's positive cases comparison 42](#_Toc125414860)

[Fig. 5.17 Country's deaths comparison 42](#_Toc125414861)

[Fig. 5.18 Mortality Rate 43](#_Toc125414862)

[Fig. 6.1 Polynomial Regression Predictions 45](#_Toc125414863)

[Fig. 6.2 Bayesian Ridge Predictions 47](#_Toc125414864)

[Fig. 6.3 SVM Predictions 49](#_Toc125414865)

[Fig. 6.4 Polynomial Regression Future Predictions 51](#_Toc125414866)

[Fig. 6.5 Bayesian Ridge Future Predictions 52](#_Toc125414867)

[Fig. 6.6 SVM Future Predictions 53](#_Toc125414868)

[Fig. 6.7 Predictions Comparison 54](#_Toc125414869)

[Fig. 7.1 Future Cases Prediction Comparison 57](#_Toc125414870)

LIST OF TABLES

[Table 4.1 Description of dataset 29](#_Toc125414878)

[Table 4.2 Description of Latest dataset 30](#_Toc125414879)

[Table 5.1 Country’s Confirmed cases 32](#_Toc125414880)

[Table 7.1 MAE and MSE Comparison 58](#_Toc125414881)

LIST OF ABBREVATIONS

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Abbreviations** | **Full Forms** |
| 1 | Covid-19 | Coronavirus Disease 2019 |
| 2 | SEIR | Susceptible Exposed Infectious Removed |
| 3 | SVM | Support Vector Machine |
| 4 | RF | Random Forest |
| 5 | SARS | Severe Acute Respiratory Syndrome |
| 6 | WHO | World Health Organization |
| 7 | AI | Artificial Intelligence |
| 8 | LSTM | Long Short-Term Memory |
| 9 | SIR | Susceptible Infected Recovered |
| 10 | JHU | John Hopkin’s University |
| 11 | DNN | Deep Neural Network |
| 12 | RNN | Recurrent Neural Network |
| 13 | GAN | Generative adversarial networks |
| 14 | CNN | Convolutional Neural Network |
| 15 | ML | Machine Learning |
| 16 | DL | Deep Learning |
| 17 | ARIMA | Autoregressive Integrated Moving Average. |
| 18 | VGG16 | Visual Geometry Group |
| 19 | MAE | Mean Absolute Error |
| 20 | MSE | Mean Squared Error |

ABSTRACT

A novel virus causes the inflammatory disease coronavirus disease (COVID-19).   
The current COVID-19 disease pandemic's causative agent, "SARS-CoV-2," was initially discovered in Wuhan, China, on December 31, 2019. The virus is believed to have originated in bats and has since spread to humans. The illness results in a respiratory condition (like influenza) with symptoms including a cold, a cough, and a fever, and in increasingly extreme situations, a breathing issue. Since then, a number of quick and severe countermeasures have been implemented in an effort to stop its global spread. It is impossible to test for Coronavirus due to time and cost constraints as the number of cases is increasing rapidly. Officials from all across the world are using a number of COVID-19 epidemic prediction models to make well-informed choices and enact necessary controls. The media and authorities are more drawn to simple epidemiological and statistical models when it comes to predicting the global spread of COVID-19.

Machine learning has become increasingly relied upon in the medical industry in recent decade. We present a machine learning approach for predicting the onset of Covid-19. Machine learning can help to speed up the turnaround time for test results and help medical personnel to treat patients who may have COVID-19. The basic generalisation and robustness abilities of current models need to be enhanced. The literature has numerous attempts to solve this issue.   
This study compares machine learning and soft computing methods to forecast the COVID-19 outbreak. SIR and SEIR models are two methods that are often used to forecast outbreaks, but they have limitations. Machine learning and soft computing methods are newer methods that may be able to overcome these limitations. The paper discusses how machine learning can be used to improve predictions for future pandemics. It explains that by combining machine learning with SEIR models, it is possible to achieve a more accurate prediction of when and where a pandemic will occur. Our model is based on a deep learning architecture that predicts the probability of infection based on data from the John Hopkins University’s website and other public sources. We evaluate our model on a held-out dataset. This thesis' major objective is to create a machine learning model to accurately forecast the spread and severity of COVID-19 cases in a given population. Identify high-risk groups and areas that are more susceptible to COVID-19 infection. Inform public health strategies and policies that can help mitigate the impact of the pandemic. Assist with resource allocation and planning, such as determining the need for hospital beds, medical supplies, and other resources. Improve the understanding of the transmission dynamics of COVID-19 and inform efforts to contain its spread. Inform research efforts on the development of vaccines, treatments, and other interventions to combat COVID-19. A literature review and experiment are planned to find a viable algorithm for such a model to evaluate the characteristics that affect the prediction model. The best algorithms for the prediction model are determined by conducting a comprehensive literature review. An experimental model is conducted for prediction of coronavirus and to identify the features that effect the model. From the literature review, a group of algorithms that are good for prediction were found. These include Polynomial Regression, Bayesian Ridge and SVM. To determine how important a feature is to the forecast, significance values are generated. The use of machine learning to predict COVID-19 may speed up the diagnosis of diseases and lower death rates.

Keywords: Machine Learning, Classification Methods, COVID-19, Prediction, Supervised Learning.

Table of Contents

[1 Chapter 1: INTRODUCTION AND BACKGROUND 11](#_Toc125414219)

[1.1 Background 11](#_Toc125414220)

[1.1.1 Background Research 11](#_Toc125414221)

[1.1.2 Parallel work 12](#_Toc125414222)

[1.1.3 Marketplace trends 13](#_Toc125414223)

[1.2 Motivation 13](#_Toc125414224)

[1.2.1 General relevance 14](#_Toc125414225)

[1.3 Aims and Objectives 14](#_Toc125414226)

[1.3.1 Aim 14](#_Toc125414227)

[1.3.2 Objectives 14](#_Toc125414228)

[1.4 Ethical considerations 15](#_Toc125414229)

[2 CHAPTER 2: LITERATURE REVIEW 16](#_Toc125414230)

[2.1 INTRODUCTION 16](#_Toc125414231)

[2.2 RELATED WORK: 17](#_Toc125414232)

[3 Chapter 3: Contributions 23](#_Toc125414233)

[3.1 Suitability and Justification of Artifact 23](#_Toc125414234)

[3.2 MODEL COMPLEXITY 24](#_Toc125414235)

[3.3 Proposed Method 25](#_Toc125414236)

[3.3.1 Overview of Machine Learning Algorithms 25](#_Toc125414237)

[3.3.2 Polynomial regression 25](#_Toc125414238)

[3.3.3 SVM 26](#_Toc125414239)

[3.3.4 Bayesian Ridge 27](#_Toc125414240)

[3.4 Outline for the research paper 28](#_Toc125414241)

[4 Chapter 4: Dataset 29](#_Toc125414242)

[4.1 Data collection 29](#_Toc125414243)

[5 Chapter 5: Exploratory Data Analysis 31](#_Toc125414244)

[5.1 Software environment 31](#_Toc125414245)

[5.2 MAE and MSE 31](#_Toc125414246)

[5.3 Data Exploring 32](#_Toc125414247)

[5.3.1 Mortality Rate 43](#_Toc125414248)

[6 Chapter 6: Implementation and evaluation 44](#_Toc125414249)

[6.1 Train-test split 44](#_Toc125414250)

[6.2 Implementation and evaluation of polynomial regression 44](#_Toc125414251)

[6.3 Implementation and evaluation of Bayesian Ridge Regression 46](#_Toc125414252)

[6.4 Implementation and evaluation of SVM 48](#_Toc125414253)

[6.5 Future predictions 50](#_Toc125414254)

[6.5.1 Predictions from polynomial regression model 51](#_Toc125414255)

[6.5.2 Predictions from Bayesian Ridge model 52](#_Toc125414256)

[6.5.3 Predictions from SVM model 53](#_Toc125414257)

[6.6 Comparing different models 54](#_Toc125414258)

[6.7 Discussion 55](#_Toc125414259)

[7 Chapter 7: Conclusion and Future Work 56](#_Toc125414260)

[7.1 Conclusion 56](#_Toc125414261)

[7.2 Limitations 58](#_Toc125414262)

[7.3 Future Work 58](#_Toc125414263)

[8 Chapter 8: REFERENCES 60](#_Toc125414264)

# Chapter 1: INTRODUCTION AND BACKGROUND

*This chapter discusses the research that has been conducted in preparation for this study. It mentions the cutting-edge theoretical ideas that will be used and any pertinent prior work that has been done in parallel. The market trends that drove this initiative are clearly visible. This establishes the backdrop of the issue that has to be resolved and the driving force behind the project, demonstrating its overall relevance and significance. The aims and objectives also cover the anticipated benefits and contribution. This chapter's final section offers discussion of the project's ethical considerations.*

## Background

### Background Research

A wide family of viruses known as corona viruses has been linked to a variety of illnesses, from the common cold to more serious conditions like the Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The corona viruses known as MERS-CoV and SARS-CoV are the carriers of these two illnesses. MERS was initially detected in Saudi Arabia in 2012, whereas SARS was originally detected in China in 2002. The SARS-COV-2 virus, which causes corona virus, was most recently discovered in Wuhan, China. Given how swiftly COVID-19 spread over the world after emerging in the start of 2020, it has had a significant influence on life as we know it. Sadly, it has resulted in the loss of many lives while upending economies and testing healthcare systems globally.

The coronavirus (COVID-19) pandemic has had an unprecedented impact on the world and its inhabitants. One of the most pressing questions is how weather might affect the spread of this virus. COVID-19 outbreaks have been reported in many countries around the world, with different regions experiencing different weather conditions. There is no definitive answer yet, but research suggests that certain aspects of weather can play a role in determining how quickly or slowly COVID-19 spreads. Temperature, humidity, air pressure, wind speed and direction are all thought to have an impact. Temperature appears to be one of the most significant factors when considering how climate affects coronavirus cases. Weather conditions have a strong impact on the spread of coronavirus disease 2019 (COVID-19), which is an airborne virus. The hotter and drier the weather, the more easily the virus can spread. Studies have shown that viruses tend to survive better at lower temperatures than higher ones; therefore, it stands to reason that colder climates may lead to more widespread transmission rates due to people spending more time indoors where they are more likely to come into contact with each other and share germs. In the summertime, when it is hot and dry, the virus can spread more easily because people are more likely to congregate in close quarters, such as in airplanes, trains, and buses. In addition, heat and dehydration can weaken the immune system, making people more susceptible to the virus. In the wintertime, when it is colder and wetter, the virus can spread less easily because people are less likely to congregate in close quarters and are more likely to wear masks and gloves. In addition, the cold weather can help to kill the virus. This would explain why countries like Russia and Canada have seen higher levels of infection compared with warmer regions such as India or Brazil. Humidity also plays a part in influencing coronavirus cases since it affects how moisture droplets linger in the air after someone coughs or sneezes which could potentially transmit the virus from person-to-person if not properly contained by face masks etcetera. Higher relative humidity tends to keep these droplets suspended for longer periods while low relative humidity ensures they evaporate faster so it’s important for governments around the world to consider both extremes when deciding on lockdown policies etcetera depending on their local climate conditions. Air pressure can also influence viral transmission rates since high pressures tend reduce airborne particles whereas low pressures increase them thus making them easier for people within close proximity inhale and become infected by any lingering virus particles present in their environment. Therefore, areas experiencing extreme changes in barometric force should take extra precautions against potential outbreaks even if temperatures remain relatively mild year round as fluctuations between high/low pressures could still cause dangerous spikes in infections over time without proper containment measures being taken beforehand.

Wind speed & direction are two additional elements worth examining since strong winds blowing from one location towards another can carry infectious particles which then spread further away from their original source increasing risk exposure among larger populations who otherwise wouldn’t been exposed directly through personal contact alone. This means places prone frequent gusty winds must look into ways reducing outdoor activities until conditions improve again lest they want to see rapid increases new transmissions occurring during those times despite taking precautionary steps prevent them elsewhere throughout day/night cycles respectively. While we don't know exactly what effect weather will have on coronavirus cases going forward, there is enough evidence to suggest that environmental factors like temperature, humidity, and air pressure can all contribute to the spread of the disease. This can be either positive or negative depending on the individual circumstances. Thus authorities need pay close attention data available order ensure maximum safety public health whilst minimizing risks associated prolonged lockdowns harsh restrictions imposed upon citizens almost every corner globe right now .

Furthermore, high temperatures have also been found effective against certain viruses such as SARS which could suggest similar benefits against COVID-19 although this needs further research before any conclusions can be drawn definitively about its effects on this particular strain of coronavirus infections. Overall, while evidence suggests that extreme hot or cold climates do not necessarily affect COVID-19 itself directly making it critical for us all take into account these factors when implementing strategies aimed at curbing its spread going forward.

### Parallel work

There is a lot of ongoing work in the field of predicting the spread and impact of corona virus. Researchers and organizations all around the globe are using various approaches, such as modeling the transmission dynamics of the virus and analyzing data on cases, hospitalizations, and deaths, to predict the course of the pandemic and inform public health policies. The study found that the vaccination was not only for COVID-19, but also for other diseases such as Japanese encephalitis, malaria, etc. The study also found that a majority of the people who are vaccinated for COVID-19 are those who are working in the health sector. A lot of researchers have come forward discussing the Machine learning methods for checking and 19analysing what does local public thing about COVID-19 vaccines in social media platforms. There are also few articles that medico-political discussion about COVID-19. One example of a research project in this area is the COVID-19 Forecast Hub, which is a collaboration between the IHME (Institute for Health Metrics and Evaluation) at (UoW)University of Washington and the Microsoft Research-Inria Joint Centre. The Forecast Hub provides real-time forecasts for corona virus, hospitalizations and deaths at global and national levels, using a range of modeling approaches. Other organizations, such as CDC (Centres for Disease Control and Prevention) and WHO, also produce forecasts and projections of the impact of the pandemic. These forecasts are based on a variety of factors, including the sum of reported cases and deaths, effectiveness due to public health measures (such as mask-wearing and social distancing), and the availability of vaccines and other interventions. Overall, the goal of this work is to better understand the spread of the virus and its impact on populations, in order to inform decision-making and mitigate the impact of the pandemic.

### Marketplace trends

There are several marketplace trends in the field of COVID-19 prediction:

* Increased demand for data analytics and AI-powered solutions: As the COVID-19 pandemic continues to evolve, there is a growing demand for data analytics and artificial intelligence (AI) solutions that can help predict the spread of the virus and its impact on various sectors.
* Emergence of new startups and companies: The COVID-19 pandemic has led to the emergence of many new startups and companies that are focused on developing innovative solutions for predicting the spread of the virus. These companies are using various data sources and technologies, such as machine learning and predictive analytics, to develop accurate and reliable models.
* Collaboration between academia and industry: There has been a significant increase in the collaboration between academia and industry in the field of COVID-19 prediction. Researchers and scientists are working with companies and organizations to develop and deploy advanced prediction models that can help inform decision-making and strategy.
* Use of big data and real-time data sources: To accurately predict the spread and impact of COVID-19, organizations are increasingly relying on big data and real-time data sources, such as social media, search data, and IoT data, to inform their models and forecasts.
* Government and policy implications: As the COVID-19 pandemic continues to evolve, governments and policy makers are using prediction models and forecasts to inform their decisions and strategies. This has led to an increased focus on the reliability and accuracy of these models and the importance of transparent data and methodology.

## Motivation

* There are several motivations for predicting COVID-19 cases and trends:
* Public health and safety: Accurate prediction of COVID-19 cases helps governments and health organizations prepare for potential outbreaks and implement necessary measures, such as lockdowns or vaccine distribution, to prevent the spread of the virus.
* Economic impacts: COVID-19 has had a significant impact on businesses and industries, and predicting trends can help decision-makers plan for future economic outcomes.
* Medical resource allocation: Predicting COVID-19 cases and trends can help healthcare systems allocate resources, such as hospital beds and personal protective equipment, in an efficient and effective manner.
* Personal preparedness: Accurate predictions can help individuals make informed decisions about their own safety, such as whether to travel or attend large events.
* Scientific curiosity: There is a strong desire among scientists and researchers to better understand the spread and impact of COVID-19, and predicting trends is a key part of this pursuit.

### General relevance

The field of COVID-19 prediction is of immense importance in the current global pandemic. Accurate predictions of the spread and impact of the virus are crucial for informing decision-making at all levels, from individual behavior changes to global policy responses. Predictive models help policymakers and public health officials to anticipate the impact of various interventions and to allocate resources appropriately. They also allow for the identification of high-risk populations and the targeting of prevention and mitigation efforts. In addition, accurate COVID-19 predictions can help individuals and communities to plan and prepare for potential outbreaks, and to make informed decisions about their own risk and behaviors. Overall, the ability to accurately predict the course of the pandemic is essential for effective and targeted response efforts, and for minimizing the negative impacts on public health, the economy, and society as a whole.

## Aims and Objectives

### Aim

The aim of COVID-19 prediction is to forecast the spread and impact of the virus in order to inform public health policy and decision-making. This includes predicting the number of cases and deaths in a given region. By accurately predicting the trajectory of the virus, authorities can allocate resources and implement strategies to mitigate the impact of the pandemic on communities and individuals.

### Objectives

* To accurately forecast the spread and severity of COVID-19 cases in a given population.
* To identify high-risk groups and areas that are more susceptible to COVID-19 infection.
* To inform public health strategies and policies that can help mitigate the impact of the pandemic.
* To assist with resource allocation and planning, such as determining the need for hospital beds, medical supplies, and other resources.
* To improve the understanding of the transmission dynamics of COVID-19 and inform efforts to contain its spread.
* To inform research efforts on the development of vaccines, treatments, and other interventions to combat COVID-19.

## Ethical considerations

Lack of COVID-19 data can significantly affect the accuracy and reliability of predictions related to the spread and impact of the virus. Without sufficient data, it is difficult to accurately model the transmission and spread of the virus, which can lead to inaccurate predictions about the number of cases and deaths that may occur in a given area or over a specific time period. Additionally, a lack of data can make it difficult to identify trends and patterns in the spread of the virus, which is essential for developing effective strategies to contain and mitigate its impact. Furthermore, a lack of data can also make it challenging to accurately assess the effectiveness of various prevention and treatment measures, as well as to identify the most vulnerable populations and areas that may be at higher risk for severe illness or death. A lack of COVID-19 data can significantly hinder our ability to effectively predict and respond to the ongoing pandemic.

* **Privacy**: Predictive models rely on data and gathering and analyzing this data can potentially invade individuals' privacy. Therefore, it is important to ensure that personal data is collected, stored, and used in a transparent and ethical manner.
* **Bias**: Predictive models can perpetuate or even amplify existing biases and inequalities if they are not designed and tested with diverse data sets. It is essential to ensure that models are not biased against certain groups of people, such as those from marginalized communities or with pre-existing health conditions.
* **Responsibility**: Predictive models can influence decisions made by governments and other organizations, so it is important to ensure that they are accurate and reliable. If a model is found to be faulty or misleading, those who developed and used it should be held accountable.
* **Transparency**: Predictive models should be transparent, with their algorithms and assumptions clearly explained to allow for independent evaluation and scrutiny.
* **Inclusivity**: Predictive models should be designed and tested with input from a diverse group of stakeholders, including those who may be impacted by the predictions. This helps to ensure that the models are fair and relevant to all populations.

# CHAPTER 2: LITERATURE REVIEW

*This chapter will review similar work on the topic coronavirus prediction using machine learning algorithms and different methods to rectify that problem and similar work done by other scholars.*

There are few steps that we performed in this paper to achieve a systematic literature review they are as follows:

**Keywords**: Machine learning algorithms, COVID-19, prediction, classification.

**Looking for literatures**: A search was performed on platforms like Google Scholar using the previously identified keywords and search strings to look for literatures.

**Evaluation**: After looking up for literatures further evaluation and selection is done that would be included in our literature review.

**Summarising**: Summarising the selected literature and including them in our literature review part.

## INTRODUCTION

A literature review is a written summary of research that has been published on a particular topic. It is an important part of the research process, as it helps to identify what has already been studied and what gaps or questions remain. A literature review helps to place the research being conducted within the broader context of what is already known about the topic. By reviewing the existing research, a literature review can identify areas where there is a lack of understanding or where further research is needed. A literature review allows researchers to evaluate the quality and relevance of previous research on a topic, helping them to determine the validity and reliability of their own findings. A literature review can help researchers to develop hypotheses and research questions for future studies, and to identify appropriate methods and designs for their research. Overall, a literature review is an important tool for synthesizing and organizing existing knowledge on a topic, and for informing and guiding future research.

There has been a vast amount of research published on Covid-19 in the scientific literature, covering a wide range of topics including the epidemiology, transmission, clinical features, diagnosis, treatment, and prevention of the disease. One key area of research has been the investigation of the viral and host factors that contribute to the severity of Covid-19. Studies have shown that older age, male sex, underlying health conditions such as hypertension, diabetes, and obesity, and certain genetic factors can increase the risk of severe illness and death from Covid-19. Other research has focused on the development and effectiveness of vaccines and treatments for the disease. There are now several vaccines that have been approved for use in different countries, and clinical trials are ongoing for various treatments, including antiviral drugs, monoclonal antibodies, and other therapeutic approaches. In addition, there has been a significant amount of research on the social and economic impacts of the pandemic, including the effects on mental health, the economic consequences of lockdowns and other public health measures, and the disparities in the burden of disease among different population groups. Overall, the research on Covid-19 has contributed significantly to our understanding of the disease and has informed the development of strategies to control the pandemic and protect public health. A literature review on Covid-19 is important because it allows for a comprehensive understanding of the current state of research on the virus. It can help to provide a broad overview of the current research on the virus, which can help to inform future research and policy decisions. Additionally, a literature review on Covid-19 can help to identify gaps in the current research, which can help to guide future research efforts. Finally, a literature review on Covid-19 can help to inform the general public about the latest research on the virus. By synthesizing the findings from a variety of sources, a literature review can provide a more complete picture of the virus and its potential implications. Additionally, a literature review can help to identify gaps in the research that may need to be addressed in order to better understand and combat Covid-19.

## RELATED WORK:

A deep discussing on the COVID\_19 prediction literature followed in scientific literature, and the opinions on the existing work done. It discusses various methods used by the fellow other researchers in the past to analyse the public option on it.

(Pal et al., 2020) For the scientific community, A lot many changes have been created due to the ongoing roll out of coronavirus(COVID-19) disease. Characteristics, likelihood and out comes of such kind of epidemic can be forecasted using AI techniques. These kind of forecast can help in monitoring and preventing the spread of this novel disease. Some major barriers to implementing artificial intelligence are its unpredictable nature and the small amount of data. In this study, they have introduced a trivial LSTM # based neural network for the prediction of a county’s number of cases. In this paper, they’ve developed a Bayesian optimization model to optimise as well as automatically generate networks customised for various countries. The findings demonstrated that the suggested pipeline performed better than state of the art method and can count as an asset for these kind of problems.

(Vega et al., 2022) The article demonstrates how challenging it is to use the SIMLR model to accurately forecast how many people would become infected during an outbreak. In the SIMLR model, machine learning improves the epidemiological SIR model. The ability of the SIR model is to track changes in governmental policy to predict the frequency of new infections which can be used to estimate the time-varying parameters of the model. The possibility that a change in policy will occur on each of these upcoming events can also be predicted by the model. Using data from Canada and the US, it was shown that the mean average percentage error of the model is comparable to that of state-of-the-art forecasting models. The model is believed to be useful in forecasting the progression of other infectious diseases in addition to projecting COVID-19 infections.

(Liu et al., 2022) The goal of the was to find out what outcomes did coronavirus had over stock market in different countries that are affected from this novel virus. The analysis revealed that the investing market in this case the stock market in all the participating countries had a considerable negative impact on returns. Asian countries responded to the outbreak quicker than other countries, but their financial markets did not recover as swiftly. The investigation also found that the confirmed coronacirus instances had a sizeable detrimental effect over major market indices. Investor dread mood was found to be the primary conciliator and imparting pathway for the adverse effect of this disease on the investment financial market. Findings of the study have important innuendo for suck policies. It is important to consider how official statements may affect people emotionally and psychologically. One of the study's many flaws is that it only examined COVID-19's short- and immediate-term effects on the country’s stock market. Health professionals must consider how their statements will make people feel and think. One of the drawbacks of our research it only took under consideration the current and short term impact of coronavirus over the stock market on bif countries because of the small wave of impact and the dynamic nature of this viral propagation. They were unable to look at demographic parameters due to the lack of data. This is another another restriction. The analysis offers a rudimentary understanding of the pandemic issue, but there is still plenty of need for more digging into investors confidence within and among international markets. Lastly, investigations may leverage the study depending on the mood of the investor and unreliability.

(Dong el at., 2020) In response to the present communal health calamities of COVID-19, JHU (Johns Hopkins University) has created online dashboard to visualise and check the reported cases worldwide. The dashboard is designed to help keep the public informed about the latest developments related to the outbreak. The place and the total sum of the positive cases, fatalities, recoveries for every impacted country are shown on their website from 22nd January, 2020. In China, cases are reported on the dashboard for province level, while countries like Canada, the US and Australia cases for each city are reported and for rest of the countries, it’s on country level. Johns Hopkins University introduced a partially automated living data stream strategy on February 1 to replace the manual reporting process. DXY is an online website created by the Chinese medical team gathers the data from the local media and information from the government. It refreshes and updates every 15 minutes to show the total number of cases for all the affected Chinese provinces as well as other countries and regions. To keep track of any new incidents, Johns Hopkins University monitored a variety of news, messages received through their website and through twitter.

(Zaid el at., 2018) The COVID-19 epidemic identification and diagnosis utilising ML and DL algorithms are discussed in this paper. The bulk of studies utilising coronavirus for ML come from countries like china, but majority of papers using COVID-19 for deep learning come from India. Generally, the coronavirus overview findings are evaluated using approaches such as ML, DL and hybrid. ML studies can be classified in 2 major divisions: that is supervised and unsupervised learning. The DL investigation is further broken down in optimization, DNN, GAN, RNN and CNN. These all studies either employees CT stan images and or X-Ray images to diagnose coronavirus. In a few studies, coronavirus was predicted using deep learning or machine learning. CNN is the very famous deep learning method. Whereas, Support Vector Mean (SVM) is famous for ML.

(Zou el at., 2020) A review of the literature on the SuEIR model and its application in foretelling confirmed and fatal cases of coronavirus in the US may provide a length of significant conclusions. The SuEIR model is a novel approach for estimating coronavirus cases that were not reported and employs ML techniques to produce short-term estimates at the national and state levels. The model has regularly demonstrated that it is capable of predicting both mortality cases and confirmed cases. Second, the SuEIR model predicts a rapid increase in COVID-19 confirmed cases and fatalities within a month, with estimates of 2 million confirmed cases and 120,000 fatalities by the end of June. Next, The SuEIR model uses training data from March 22, 2020, when the majority of states had already enacted stay-at-home orders, and makes the assumption that the contact rate will remain constant while the training period and prediction. But during the month of May, a lot of states began lifting their restrictions on establishments and public spaces, which might have an effect on the virus's spread and contact rate. The present algorithm doesn’t take that in count. Fourth, the SuEIR model found that discovered rate is below 0.1 for the majority of states, indicating substantial fragment of few people will be saved or pass away even without testing and reporting. It is in line with study from the University of Southern California that estimated CI: [2.8%, 5.6%] (4.65% )of LA citizens had diminished coronavirus, this is almost 23rd times greater than the numbers that were turned in (Sood et al., 2020). This means that despite official reports to the contrary, there may be be a much higher number of infected people in the US. The SuEIR model can be used to anticipate the spread of COVID-19 and identify likely unreported cases in general, but more digging is necessary to completely get around the impact of reporting orders for the virus's dissemination and the accuracy of the model's predictions.

(Int J Biol Sci. el at., 2021) The COVID-19 pandemic study discusses a number of significant topics. The first is that AI is being used in a wide range of COVID-19-related domains. Diagnosis of coronavirus’s +ve cases in particular may be accelerated using AI. Second, to review the literature on AI and COVID-19, preprint services and many other academic databases. This evaluation concentrated on medical application of ML and DL, such as the diagnosis of COVID-19 utilising clinical features, including X Ray, CT scan etc. Third, paper discusses AI's promise as well as the challenges it faces today in battling coronavirus pandemic. These observations can help as roadmap for most effective use of AI technologies during pandemics. In conclusion, applying AI to coronavirus has the promise to advance policy activities as well as research while also enhancing diagnosis and treatment. More study is required to properly understand the potential and constraints of AI in this context.

(Ardabili el at., 2020) Numerous significant observations are made in this study on the use of computing models such as machine learning forecast for coronavirus pandemic. First off, global SARS-CoV-2 pandemic has elevated to a serious national security problem for many countries, and it is crucial to comprehend the transmission and consequences of the disease by creating accurate prediction models. Traditional epidemiological models have had only sporadic success in making long-term projections since there is so much uncertainty and a lack of data. Second, this work compares soft computing and machine learning for predicting coronavirus. On the basis of these findings and the COVID-19 outbreak's complex character, which varies from country to country, the study suggests that ML may be a suitable tool for modelling the time series of the outbreak. According to the findings, more study is needed to assess various ML models for certain countries because there may not be a way to create global models with the potential for generalisation. The mortality rate (n(deaths) / n(infected)) must be accurately estimated in order to determine need for care beds depending on the number of patients. Fourth, the study asserts that future research should focus on combining ML with regular epidemiological models in order to increase the accuracy and lead time of present models.

(Alazab el at., 2020) In the paper, it is described how AI applications should be used for forecast and diagnose coronavirus. As prediction models, the probabilistic autoregressive integrated moving average (PA) algorithm, the autoregressive moving average (ARIMA) algorithm, LSTM approach are used. Diagnosis model uses CNNs created with VGG16 architecture. This study found that the PA algorithm performed best in estimating coronavirus patients, recoveries and fatalities in both Jordan and Australia. CNN model was 99% accurate at detecting COVID-19 using an improved dataset. Prediction models such LSTM, ARIMA and PA algorithms were used for forecasting number of coronavirus confirmations, recoveries and deaths. PA delivered performance of the day. It predicted coronavirus patients, recoveries and fatalities with forecast accuracies of 99.94%, 90.29%, and 94.18% in Australia. Coronavirus patients, recoveries and fatalities in Jordan had projected accuracy of 99.08%, 79.39%, and 86.82%, respectively. Digging into more sophisticated prediction and forecasting techniques will be covered in a subsequent publication. For the purpose of locating corona virus in X ray images of chest, it was suggested to use VGG16 in a diagnosis model. The model can detect COVID-19 rapidly and correctly, enabling it to achieve an F-measure of 99% with the aid of an improved dataset. In a subsequent study, we'll investigate using VGG XX versions to detect coronavirus in CT scan images of chest and evaluate their performance using larger datasets. Another addition of this work is the investigation of spread of corona virus and its statistical dataset based on its global distributions. Due to which, using our AI based research, we arrived at the following two key conclusions: Similar characteristics can be found in the most severely affected areas, and the disease spreads far more swiftly in coastal areas than in other non-coastal ones. Cities by the seaside must therefore take extra care and consideration. We will investigate the effects of topography, humidity, temperature corona virus distribution in cities and the countries in our next research.

(Ziong el at., 2021) Of order to make predictions about the increase in COVID-19 cases and deaths, this study employs Bayesian time series modelling and random forest approaches. It is challenging to predict how corona virus will unfold because of distinct characteristics of virus, the dearth of data, and the dynamic nature of society and political responses. The study also incorporated some models to create predictions about COVID-19 that were supported by real data. By using such models case trajectories are calculated which also incorporates previous knowledge. The compartmental model predicts deaths and generates daily predictions in the US using this distribution and RF algorithms that has been trained using coronavirus data factors. Third, the model's performance was evaluated by looking at its accuracy over 21-day projections and training it on epidemics with progressively longer durations. The model was able to capture the considerable variance between states by comparing the expected trajectories and related uncertainty for three separate locations viz. NY, Colorado, and West Virginia. Overall, use random forest algorithms and Bayesian time series modelling offers a sophisticated and precise strategy for predicting corona virus in the US. It provides precise assessments of the pandemic's current trajectory's level of uncertainty as well as a basis for future forecasts as governmental and social reactions change.

(IEEE access, 2021) The HHO-FKNN model described in this review of literature seems to be a ML model that forecasts severity of coronavirus based on a range of input factors, including patient-specific information, underlying diseases, symptoms, and immunological index. The previously unresearched use of the immunological index as a predictor of COVID-19 severity is one of the innovative features of this model. Another cutting-edge method is the HHO algorithm's simultaneous feature and parameter screening of FKNN. HHO FKNN model outperforms other machine learning algorithms in terms of prediction accuracy and performance stability when it comes to forecasting the severity of COVID-19, according to the experimental findings provided in the review. The authors of the review plan to try using the HHO FKNN model to solve problem of pre diagnosis of corona virus as well as other disease problems in their work.

(Bo Wang el at., 2021) This literature review discusses the development of a system that makes use of artificial intelligence (AI) to help with the detection of coronavirus in computed tomography images of CT scan. The method is designed to assist radiologists, particularly in areas where COVID-19 is prevalent and there is a shortage of radiologists, by delivering early CT data to make the process quick detecting corona virus patients. This method can also help radiologists with less experience distinguish COVID-19 from ordinary pneumonia in CT images thanks to its highly distinctive features. The method for building the AI model, which included data collection, annotation, testing, user interaction design and clinical deployment is discussed in the study. A model library and instructional resources were also included of the technique. The report asserts that the model was developed and put into use in just 7 days and that it was capable of doing way greater than 1300 screenings per day. The authors note that the way the model performs can be improvised for greater accuracy by updating it frequently with new data and they are also working on developing a multi-modal model that improves screening accuracy by taking into account.

(Peipei Wang el at., 2021) This review of the literature highlights a study that using LSTM networks and rolling updating method to predict the long-term epidemic pattern of COVID-19 in Russia, Peru, and Iran. The study focused on predicting the number of daily new instances of COVID-19 in order to determine the entire curve of all verified cases over the duration of the next 150 days. The study was also the first to evaluate the effectiveness of government-implemented preventive treatments in each country using a metric called the "Diminishing Index" (DI). According to the analysis, the COVID-19 outbreak in Peru would peak in the first few days of December and result in a total of 398,991 infections. It is projected that by mid-November, after several small increases from July to September, Iran's daily positive case count will be below 1000. There should be more than 2000 new cases in Russia by the beginning of December. The results of the study show that COVID-19 development can be significantly slowed by strict control techniques. The assumption that present policies won't change and that people will follow safety precautions are just two of the limitations and assumptions the authors recognise exist in their study. Furthermore, neither the spatial impacts between nations nor the impact of imported cases are taken into account in the study.

# Chapter 3: Contributions

*The following chapter will be discussing the research technique we followed in this paper. It also discusses the problem statement, aims and objectives that are followed. Some methods which are briefly suggested are discussed further.*

## Suitability and Justification of Artifact

There are a number of tools and techniques that have been developed to predict the spread and impact of COVID-19, including statistical models, machine learning algorithms, and data visualization tools. The suitability of a particular artefact for COVID-19 prediction will depend on a number of factors, including the type of data it can analyse, the accuracy and precision of its predictions, and its ability to be easily understood and used by non-technical users. One important consideration when evaluating the suitability of an artefact for COVID-19 prediction is the quality and reliability of the data it is based on. Accurate predictions depend on having access to high-quality, up-to-date data on factors such as the number of confirmed cases, the rate of transmission, and the effectiveness of containment measures. Another important factor is the ability of the artefact to handle the complexity and uncertainty inherent in the COVID-19 pandemic. This may include its ability to account for factors such as changing infection rates, the potential for outbreaks in new areas, and the impact of interventions such as vaccination programs. It is important to carefully evaluate the suitability of any artefact for COVID-19 prediction, and to consider a range of different tools and techniques in order to get a comprehensive view of the situation.

A model that relies on accurate and current data is more likely to produce accurate predictions. The quality of COVID-19 data can vary depending on a number of factors. Some of the key factors that can impact the quality of COVID-19 data include:

**Data collection methods**: The accuracy of COVID-19 data depends on the methods used to collect it. For example, if data is collected through self-reported surveys or online reporting platforms, there may be a higher risk of errors or biases.

**Timeliness**: The quality of COVID-19 data can be impacted by how quickly it is collected and made available to the public. Data that is collected and reported in real-time is generally considered to be of higher quality than data that is delayed.

**Data accuracy**: The accuracy of COVID-19 data is also impacted by the accuracy of the testing methods used. If testing methods are not accurate, then the data collected may not be reliable.

**Data completeness**: The quality of COVID-19 data is also impacted by how complete it is. Data that is missing key information or is incomplete may not be as useful for decision-making purposes.

**Data accessibility**: The quality of COVID-19 data can also be impacted by how easily it is accessed and understood by the public. If

## MODEL COMPLEXITY

Model complexity is also a major factor. A model that is too complex may overfit the data, leading to poor generalization to new cases. On the other hand, a model that is too simple may not capture the necessary nuances and relationships in the data. There are many different models that have been developed to predict the spread and impact of COVID-19. Some of these models are quite simple, while others are more complex. The complexity of a model can depend on a number of factors, including the number of variables it takes into account, the level of detail it is able to capture, and the algorithms or techniques it uses. More complex models may be more accurate but also more difficult to understand and interpret. They may also be more prone to overfitting, which is when a model fits the training data very well but does not generalize well to new data. It is important to carefully evaluate the trade-offs between model complexity and accuracy when developing models for predicting COVID-19 spread and impact. It may be necessary to use a more complex model in some cases, but a simpler model may be sufficient in others.

The choice of features can greatly impact the accuracy of the model. It is important to carefully select relevant and informative features. Some common features used for COVID-19 predictions include demographic factors such as age, gender, and ethnicity, behavioural factors such as social distancing and mask wearing, environmental factors such as population density and air pollution, health care system factors such as hospital capacity and testing rates, economic factors such as unemployment rates and GDP, political factors such as government policies and actions, geographical factors such as proximity to major cities and transportation networks, genetic factors such as susceptibility to severe illness, historical data on previous outbreaks and pandemics, machine learning algorithms and statistical modeling techniques, etc.

The model should be trained and evaluated on a diverse set of data to ensure its robustness and generalizability. Training and evaluation for COVID-19 predictions involve the use of machine learning algorithms to analyze data on the spread and impact of the virus. The goal is to create a model that can accurately predict the spread and impact of COVID-19 in different regions or populations. To train the model, data scientists first gather a large dataset on COVID-19 cases, including information on factors such as population density, healthcare infrastructure, and social distancing measures. This data is then fed into the machine learning algorithm, which uses statistical techniques to identify patterns and trends in the data. Once the model has been trained, it is evaluated using a separate dataset to see how well it performs in predicting the spread and impact of COVID-19. This evaluation process helps to identify any weaknesses or biases in the model, and allows data scientists to fine-tune the model to improve its accuracy. Overall, training and evaluation for COVID-19 predictions is a critical process that helps policymakers and public health officials make informed decisions about how to respond to the pandemic. By accurately predicting the spread and impact of COVID-19, these models can help guide efforts to slow the spread of the virus and protect public health. A model that takes into account these factors and is regularly updated with new data is likely to be more suitable for predicting COVID-19 cases.

## Proposed Method

### Overview of Machine Learning Algorithms

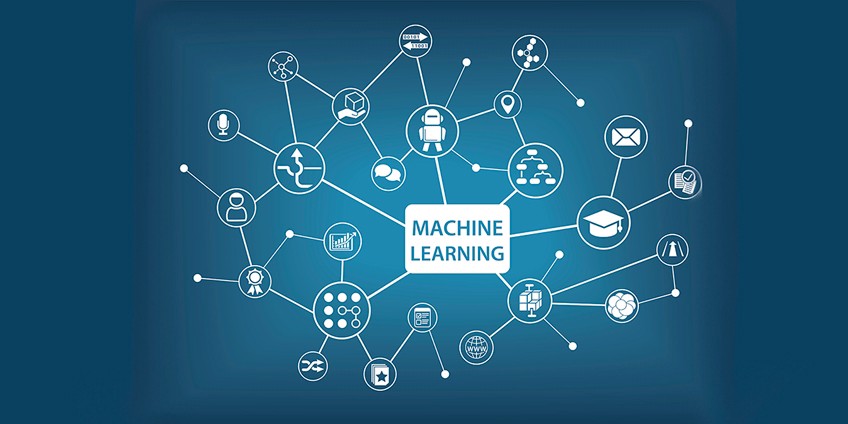
****Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence that is concerned with the design and development of algorithms that can learn from and make predictions on data. The main goal of machine learning is to enable computers to learn automatically without being explicitly programmed.

Fig. 3.1 Machine Learning

There are many different types of machine learning, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In machine learning, an algorithm is trained on a data set. The algorithm then makes predictions or decisions without being explicitly told how to perform the task. For example, a machine learning algorithm trained on a data set of customer data could make predictions about which customers are likely to churn (stop being a customer).

### Polynomial regression

Polynomial regression is a type of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial. Polynomial regression can be used to model relationships between variables that are not linear. For example, if you are trying to predict the amount of rainfall in a particular region based on the temperature, you might find that there is a polynomial relationship between these two variables. In this case, you could use polynomial regression to model the relationship and make predictions about rainfall based on temperature. Polynomial regression is similar to linear regression, but instead of fitting a straight line to the data, it fits a polynomial curve. The degree of the polynomial (i.e., the number of terms in the polynomial) determines the flexibility of the model and the complexity of the relationship between the variables. A high degree polynomial can model complex relationships, but it is more prone to overfitting than a simpler model.

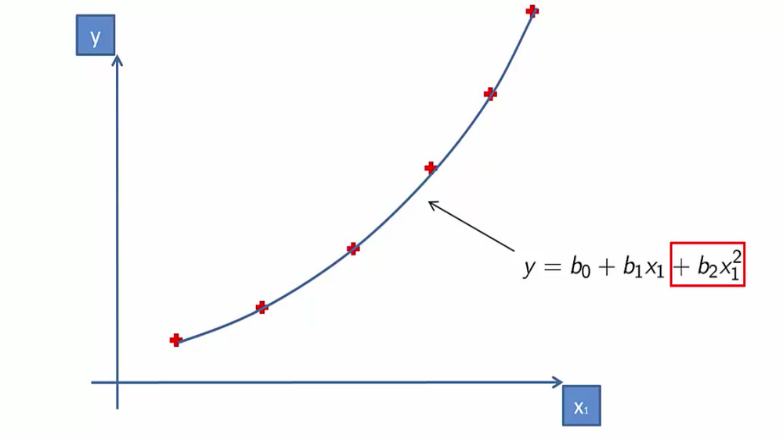
In polynomial regression, we try to fit a polynomial function to the data. A polynomial function is defined as:

Fig. 3.2 Polynomial Regression (Selvi T., 2021)

f(x) = b0 + b1x+ b2x2 + ... + bnxn

where b0, b1, ... bn are the coefficients and x is the independent variable. The degree of the polynomial function is n.

### SVM

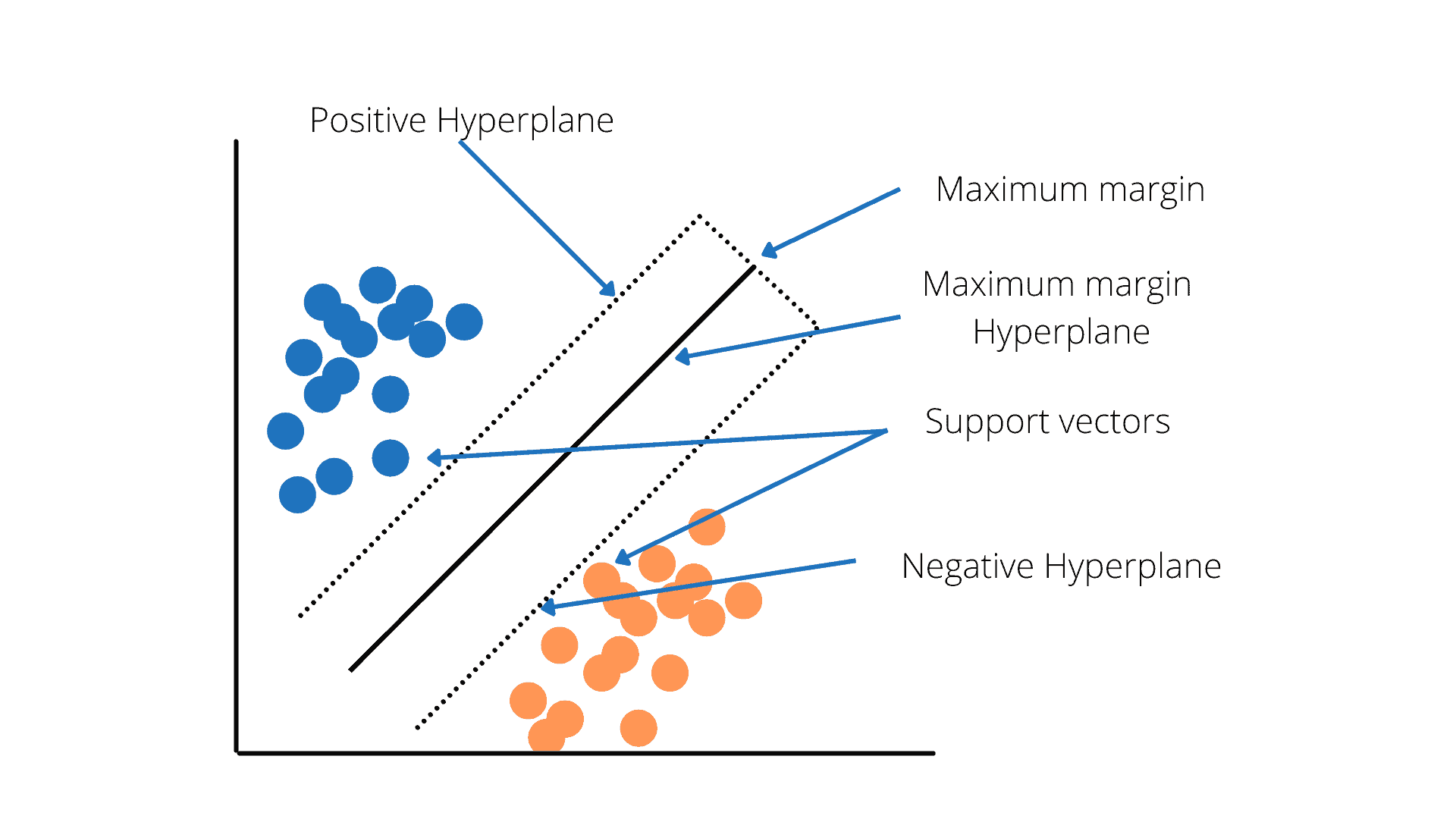
Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. The goal of an SVM is to find the hyperplane in a high-dimensional space that maximally separates the different classes. In the case of classification, the algorithm finds the hyperplane that maximally separates the different classes, so that the data points belonging to different classes are as far apart as possible. In the case of regression, the algorithm finds the hyperplane that is as close as possible to all of the data points, so that the errors between the predicted values and the true values are minimized. SVMs are particularly effective in cases where the number of dimensions (features) is much greater than the number of samples. They are also effective in cases where the classes are highly imbalanced (e.g., there are many more negative examples than positive examples). SVMs are based on the idea of finding a hyperplane that maximally separates different classes. In the case of binary classification, this hyperplane is a line that separates the two classes. In the case of multi-class classification, there are multiple hyperplanes that can be used to separate the different classes. The distance from the hyperplane to the nearest data points is known as the margin. The goal is to find the hyperplane that has the maximum margin. This is known as the maximum margin classifier. SVMs are a powerful tool for classification, but they can also be used for regression. In this case, the goal is to find the hyperplane that best fits the data. SVMs have a number of advantages, including the ability to handle high-dimensional data and the ability to perform well even when the data is not linearly separable. However, they can be sensitive to the choice of hyperparameters and can be computationally expensive to train.

Fig. 3.3 Support Vector Machines (Alam, B., 2022)

In Support Vector Machines (SVMs), the goal is to find the hyperplane in an N-dimensional space that maximally separates the classes. The hyperplane is represented by the equation:

w1x1 + w2x2 + ... + wnxn + b = 0

where x1, x2, ..., xn are the features and w1, w2,...,wn are the coefficients. The value of b is the bias. This equation defines the decision boundary of the SVM.

To find the optimal hyperplane, we need to maximize the margin, which is the distance between the hyperplane and the nearest data point from either class. This can be represented mathematically as the following optimization problem:

maximize (2/||w||)

subject to: yi(w.xi + b) >= 1, for i = 1, ..., n

Here, xi is a vector of the features for the ith data point and yi is the label of the ith data point (-1 or 1). The term ||w|| represents the Euclidean norm of the vector w.

This optimization problem can be solved using quadratic programming techniques. Once the optimal hyperplane is found, it can be used to classify new data points by plugging in the feature values into the equation of the hyperplane and checking the sign of the result. If it is positive, the point is classified as belonging to one class, and if it is negative, it is classified as belonging to the other class.

### Bayesian Ridge

Bayesian Ridge Polynomial Regression is a non-linear extension of Bayesian Ridge Regression. It is a probabilistic model that uses Bayesian inference to estimate the model parameters. It is a combination of Ridge Regression and Polynomial Regression.

In Bayesian Ridge Regression, the regularization term is a L2 penalty on the parameters, which helps to prevent overfitting. Bayesian Ridge Polynomial Regression extends this approach by treating the regularization term as a random variable that is assigned a prior distribution.

The main difference is that Bayesian Ridge Polynomial Regression uses polynomial functions as the basis function. The polynomials are used to fit a non-linear relationship. The degree of polynomials can be chosen to fit the complexity of the relationship in the data. This can be a powerful tool to model non-linear relationships in the data, but it can also increase the risk of overfitting if the degree of polynomials is too high.

In Bayesian Ridge Polynomial Regression, the prior distribution for the weight vector is assumed to be a normal distribution with mean 0 and precision (inverse of variance) alpha. The precision of the noise is also assumed to be gamma. The posterior distribution is calculated by using bayes rule and it is also a normal distribution. The posterior mean and covariance are used to calculate the weight vector and variance of the error term.

The formula is:

w = (XTX + (alpha-1)I)-1 XTY

where I is the identity matrix.

In this way, Bayesian Ridge Polynomial Regression estimates the model parameters in a probabilistic manner, which can provide uncertainty estimates for the predictions.

Top of Form

## Outline for the research paper

**Step 1:** Gather the datasets from a trusted source like JHU.

**Step 2:**  Perform data pre-processing to annotate the data.

**Step 3:** Perform EDA to explore, understand and analyse the data better.

**Step 4:** Splitting the dataset into training and testing sets.

**Step 5:** Applying model parameter settings for different models (Polynomial, Bayesian Ridge, SVM)

**Step 6:** Evaluating the model on the testing data.

**Step 7:** Comparing MAE and MSE for different models.

**Step 8:** Using the trained model to predict the cases in future.

# Chapter 4: Dataset

## Data collection

The required dataset is fetched from JHU’s (John Hopkins University) website. The Johns Hopkins University (JHU) COVID-19 data is generally considered to be reliable. The data is compiled from multiple sources, including the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), and other national and regional health agencies. The data is also cross-referenced with media reports and other sources to ensure accuracy.

It's important to note that the COVID-19 pandemic is an evolving situation, and the data on the JHU website may not always be up-to-date or complete. However, the JHU team works to ensure that the data is as accurate as possible and is regularly updated.

From John Hopkins University we have imported a total of 4 data sets in the form of CSV files. The first data set contains all the confirmed cases around the globe for a particular country, the second data set shows the deaths around the world for each country, the 3rd data set contains information about the total recovery done so far, fourth data set is off the latest data which contains information about the confirmed cases, deaths and recoveries of a particular country in total.

Dataset for confirmed cases, deaths and recoveries done so far also contains the following information:

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Features | Data Type | Description |
| 1 | Province/ State | object | Names for the province/ State of a particular country |
| 2 | Country/ Region | object | Names of a particular Country/ Region |
| 3 | Latitude | float64 | Latitude of a particular country |
| 4 | Longitude | float 64 | Longitude of a particular country |
| 5 | 1/22/2020 till 12/31/2022 | Int64 | Number of respective cases till each date for a particular country |

Table 4.1 Description of dataset

Latest dataset contains the following information where the rows with important information are highlighted:

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Features | Data Type | Description |
| 1 | FIPS | float64 | Federal Information Processing System (FIPS) code |
| 2 | Admin2 | object | NaN |
| 3 | Province\_State | object | Names for the province/ State of a particular country |
| 4 | Country\_Region | object | Names of a particular Country/ Region |
| 5 | Last Update | object | Date and time on which the data was last updated |
| 6 | Lat | float64 | Latitude of a particular country |
| 7 | Long\_ | float64 | Longitude of a particular country |
| 8 | Confirmed | int64 | Total number of confirmed cases for that particular country |
| 9 | Deaths | int64 | Total number of deaths for that particular country |
| 10 | Recovered | float64 | Total number of recovered cases for that particular country |
| 11 | Active | float64 | NaN |
| 12 | Combined\_Key | object | Names of a particular Country/ Region |
| 13 | Incident\_Rate | float64 | Incident rate for each country |
| 14 | Case\_Fatality\_Ratio | float64 | Case Fatality Ratio for each country |

Table 4.2 Description of Latest dataset

# Chapter 5: Exploratory Data Analysis

## Software environment

Python is a high-level, interpreted programming language. It was created in the late 1980s by Guido van Rossum. Python is known for its simplicity and readability, making it a popular choice for beginners and experienced programmers alike. Python has a large and comprehensive standard library and supports many programming paradigms, including object-oriented, imperative, functional, and procedural. It is often used for web development, scientific computing, data analysis, and artificial intelligence.

**NumPy**: NumPy is a Python library for working with large, multi-dimensional arrays and matrices of numerical data. It provides tools for performing mathematical operations on these objects, including linear algebra, statistical functions, and random number generation.

**Pandas**: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. pandas is built on top of NumPy and is intended to integrate well with the larger scientific computing and data analysis ecosystem in Python.

**Matplotlib**: Matplotlib is a powerful tool for creating visualizations of data in a 2D plot and is often used in combination with other libraries such as NumPy and pandas to plot data stored in arrays and data frames. It is a flexible library that can be used in a wide range of applications, including data visualization for scientific research, visualization of data for business intelligence applications, and creating visualizations for data storytelling.

**Scikit-learn**: Scikit-learn is a Python library for machine learning that provides simple and efficient tools for data mining and data analysis. scikit-learn is widely used in the data science community and is a powerful tool for creating machine learning models and conducting data analysis. A wide range of supervised and unsupervised learning algorithms, including support vector machines, random forests, gradient boosting, k-means, etc.

## MAE and MSE

Mean square error (MSE) and mean absolute error (MAE) are two common measures of the difference between the predicted values of a model and the true values.

Mean square error (MSE) is the average of the square of the differences between the predicted and actual values. It is calculated as:

MSE = (1/n) \* Σ(y\_i - y\_pred\_i)2

Where y\_i is the actual value, y\_pred\_i is the predicted value, and n is the number of observations.

The main advantage of MSE is that it punishes large errors more than smaller errors.

Mean Absolute Error (MAE) is the average of the absolute differences between the predicted and actual values. It is calculated as:

MAE = (1/n) \* Σ |y\_i - y\_pred\_i|

Where y\_i is the actual value, y\_pred\_i is the predicted value, and n is the number of observations.

The main advantage of MAE is that it gives equal weight to all errors regardless of the magnitude.

Both MSE and MAE are commonly used to evaluate the performance of a model, but MSE is more sensitive to outliers than MAE, which means that if the model is sensitive to outliers, it's better to use MAE as a performance metric. Also, MSE is sensitive to the scale of the data, while MAE is not, so if the data has different scales it's better to use MAE.

## Data Exploring

Exploratory data analysis (EDA) is an approach to analyzing and understanding data by summarizing its main characteristics, often through visual methods. It is used to detect patterns, anomalies, and relationships in the data, and to generate hypotheses about the underlying processes that generate the data. EDA is a crucial step in data analysis, as it helps to identify potential issues with the data and to guide the selection of appropriate statistical models.

The following table contains the information of the total number of positive coronavirus cases for top 10 countries.

|  |  |  |
| --- | --- | --- |
| SR. No. | Country | Number of cases |
| 1 | US | 100780723 |
| 2 | India | 44679873 |
| 3 | France | 39498188 |
| 4 | Germany | 37369866 |
| 5 | Brazil | 36331281 |
| 6 | Japan | 29234677 |
| 7 | Korea | 29116800 |
| 8 | Italy | 25143705 |
| 9 | United Kingdom | 24365688 |
| 10 | Russia | 21490515 |

Table 5.1 Country’s Confirmed cases

Table

Description automatically generatedThis is the graph of all the countries plotting it’s total confirmed cases in an ascending order.

Fig. 5.1 Ascending graph of cases for all the countries

Top of Form

The table presented below displays the top 20 countries with the highest confirmed cases, deaths and corresponding mortality rates, arranged in descending order of confirmed cases. The background colour of each entry is representative of the severity of the cases, with darker shades indicating a higher number of confirmed cases.

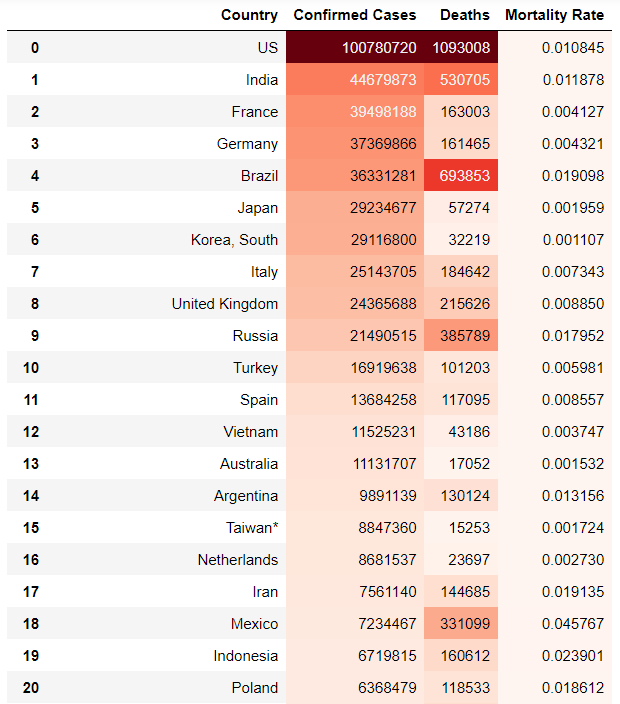
****

Fig. 5.2 Colour gradient for top 20 country

The following graph shows the increase in the number of covid-19 cases worldwide from 1/22/2020 till 12/31/2022.

**Chart, line chart

Description automatically generated**

Fig. 5.3 Worldwide cases

The following graph shows the increase in the number of deaths due to covid-19 worldwide from 1/22/2020 till 12/31/2022.

**Chart, line chart

Description automatically generated**

Fig. 5.4 Worldwide deaths

The following graph plots the everyday increase in the number of coronavirus cases along with the moving average every 7 days over the past 3 years.

**Chart, histogram

Description automatically generated**

Fig. 5.5 Everyday increase with moving average

The following graph plots the everyday increase in the number of deaths worldwide along with the moving average every 7 days over the past 3 years.

Chart, histogram

Description automatically generated

Fig. 5.6 Everyday deaths

India’s everyday increase in confirmed cases and deaths over the past three years.

**Chart, histogram

Description automatically generated**

Fig. 5.7 India's cases

**Chart, histogram

Description automatically generated**

Fig. 5.8 India's deaths

United Kingdom’s everyday increase in confirmed cases and deaths over the past three years.

Chart

Description automatically generated

Fig. 5.9 UK's cases

**Chart, histogram

Description automatically generated**

Fig. 5.10 UK's Deaths

US’s everyday increase in confirmed cases and deaths over the past three years.

**Chart, histogram

Description automatically generated**

Fig. 5.11 US's Cases

**Chart

Description automatically generated**

Fig. 5.12 US's Deaths

The following graphs shows the bar plot for top 10 countries around the globe with the highest number of cases till 12/31/2022.

**Chart, bar chart

Description automatically generated**

Fig. 5.13 Top 10 countries

Comparing positive cases in India vs US vs all the other countries overtime.

Chart

Description automatically generated with medium confidence

Fig. 5.14 India vs US vs others

Percentage split of positive cases in India vs US vs other countries

United States 100780720 cases (15.3%)

India 44679873 cases (6.8%)

Other countries 515018586 cases (78.0%):

Total: 660479293 cases

The below pie chart shows the following positive COVID-19 cases around the globe for different countries.

**Chart, sunburst chart

Description automatically generated**

Fig. 5.15 Pie chart for top countries

Comparing number of positive cases till date for countries viz. India, UK, US, China, Austria, Italy and Pakistan.

**Chart, line chart

Description automatically generated**

Fig. 5.16 Country's positive cases comparison

Comparing number of deaths till date for countries viz. India, UK, US, China, Austria, Italy and Pakistan.

**Chart, line chart

Description automatically generated**

Fig. 5.17 Country's deaths comparison

### Mortality Rate

The mortality rate for COVID-19 refers to the number of deaths caused by the disease as a proportion of the total number of confirmed cases. It is typically expressed as a percentage. For example, if there were 100 confirmed cases of COVID-19 and 10 deaths, the mortality rate would be 10%. The COVID-19 mortality rate can also be calculated using the number of deaths per 100,000 population. This allows for comparison of the disease's impact across different populations, regardless of their size.

The following graph represents the mortality rate and the mean of it over the period of last three years.

**Chart, line chart

Description automatically generated**

Fig. 5.18 Mortality Rate

# Chapter 6: Implementation and evaluation

## Train-test split

Train-test split is a technique used in machine learning to divide a dataset into two subsets: a training set and a test set. The training set is used to train a model, while the test set is used to evaluate the performance of the model on unseen data. This technique is used to ensure that the model is not overfitting and is able to generalize well to new data. The typical split ratio is 80% of the data for training and 20% for testing.

**Skipping days:** When using historical data to predict future events, it is common to "skip" or "hold out" a portion of the data for use as a test set. This is done to evaluate the performance of the model on unseen data and ensure that it is not overfitting to the training data. It's also important to consider the fact that the dynamics of the spread of the virus might change with time, so it's important to use recent data and not too old data, to make predictions of the spread of the virus.

Using a train-test split is especially important when working with COVID-19 data because it is a rapidly evolving situation and new information is constantly being added. By using a train-test split, a model can be trained on historical data and then tested on more recent data to evaluate its ability to make accurate predictions. Additionally, it can ensure that the model has not overfit to the training data, which is particularly important when dealing with limited data. It also allows to evaluate the robustness and reliability of the model with unseen data. It helps also to identify any potential biases in the data that could impact the model's performance and helps to make sure that the model is able to generalize well to new data.

## Implementation and evaluation of polynomial regression

pol = PolynomialFeatures(degree=3)

When using polynomial regression, the degree of the polynomial term (i.e., the highest power of x in the equation) is a hyperparameter that must be chosen by the researcher. A higher degree polynomial will be able to fit more complex relationships between the independent and dependent variables, but may also be more prone to overfitting. In general, a lower degree polynomial will be less flexible and may be less accurate, while a higher degree polynomial will be more flexible and may be more accurate but also more prone to overfitting. In practice, the degree of the polynomial term is often chosen through a process of trial and error, where the researcher iteratively fits models with different degrees and selects the one that gives the best performance. In this paper, after trying different degree, 3rd degree fits the best for our dataset.

**Predicting the test data:**

linear\_pred\_test = model.predict(x\_test\_pol)

**Plotting graph to compare the results of polynomial regression with the testing data.**

**Chart, line chart

Description automatically generated**

Fig. 6.1 Polynomial Regression Predictions

**Checking the coefficients for the polynomial regression:**

print(model.coef\_)

When using polynomial regression, it is important to check the coefficients of the polynomial terms in the model. These coefficients indicate the strength and direction of the relationship between the independent variable(s) and the dependent variable.

[[ 0.00000000e+00 7.79654773e+07 -7.69241494e+04 2.54234517e+01]]

**Checking the mean square error and mean absolute error:**

print('Mean Absolute Error:', mean\_absolute\_error(linear\_pred\_test, y\_test))

print('Mean Squared Error:', mean\_squared\_error(linear\_pred\_test, y\_test))

**Results:**

Mean Absolute Error: 451832.86442587717

Mean Squared Error: 294329520918.9113

## Implementation and evaluation of Bayesian Ridge Regression

bayesian\_pol = PolynomialFeatures(degree=3)

**Selecting parameters for Bayesian grid**

The choice of the values for the tolerance (tol), alpha, and lambda hyperparameters in Bayesian Ridge Regression can have a significant impact on the performance of the model.

The tolerance (tol) parameter controls the tolerance for stopping the optimization of the model's parameters. A smaller tolerance will result in more precise estimates of the parameters but will also require more computational resources.

The alpha parameter is a regularization term that controls the balance between fitting the data and preventing overfitting. A small alpha value will result in a model that is more complex and may overfit the data, while a large alpha value will result in a simpler model that may underfit the data.

The lambda\_1 and lambda\_2 are the regularization parameters for the model. lambda\_1 controls the L1 regularization, and lambda\_2 controls the L2 regularization. L1 regularization will produce sparse models, with many coefficients set to zero. L2 regularization will produce models with small but non-zero coefficients.

The best way to select the optimal values for these hyperparameters is through a process of trial and error, using techniques such as cross-validation or a held-out test set. One common way is to use GridSearchCV from sklearn library, which allows to specify a range of values for each hyperparameter and will automatically tune them to get the best performance.

Additionally, you can use techniques like Bayesian Optimization which uses a probabilistic model to guide the search for the best hyperparameters.

It's also important to note that these hyperparameters can be data dependent, so it's important to select the optimal value for each dataset.

baye\_search = RandomizedSearchCV(baye, bay\_grid)

baye\_search.fit(x\_train\_bay, y\_train)

baye\_search.best\_params\_

{'tol': 0.0001,

'lambda\_2': 1e-05,

'lambda\_1': 0.001,

'alpha\_2': 1e-05,

'alpha\_1': 0.0001}

Predicting the test data

baye\_pred\_test = baye\_confirmed.predict(x\_test\_bay)

**Plotting graph to compare the results of Bayesian Ridge regression with the testing data.**

Chart, line chart

Description automatically generated

Fig. 6.2 Bayesian Ridge Predictions

**Checking the MAE and MSE for Bayesian Ridge Polynomial Regression model**

print('Mean Absolute Error:', mean\_absolute\_error(baye\_pred\_test, y\_test))

print('Mean Squared Error:', mean\_squared\_error(baye\_pred\_test, y\_test))

**Results:**

Mean Absolute Error: 9949144.572066076

Mean Squared Error: 100286372600178.38

## Implementation and evaluation of SVM

svm = SVR(kernel = 'poly')

In a Support Vector Machine (SVM) model, the kernel is a function that transforms the input data into a higher-dimensional space. The goal of this transformation is to create a linear boundary that separates the different classes in the transformed space.

There are several different types of kernels that can be used in an SVM model, including:

**Linear kernel:** This is the default kernel and simply performs a dot product between the input data and the model parameters.

**Polynomial kernel:** This kernel transforms the input data into a polynomial space of a specified degree.

**Radial basis function (RBF) kernel:** This kernel transforms the input data into a space where the distance to a specified point (the "center") is used as a feature. This is also known as the Gaussian kernel.

**Sigmoid kernel:** This kernel is similar to the RBF kernel, but uses a sigmoid function instead of a Gaussian function.

The choice of kernel depends on the specific characteristics of the data and the problem being solved.

For this particular paper a polynomial kernel fits the best because it is particularly useful when the data is not linearly separable, such as in the case of COVID-19 spread predictions where the spread of the virus can be affected by many different factors.

**Predicting the test data:**

svm\_test\_predictions = svm.predict(x\_test)

**Plotting graph to compare the results of SVM with the testing data.**

**Chart, line chart

Description automatically generated**

Fig. 6.3 SVM Predictions

**Checking the MAE and MSE for SVM model**

print('Mean Absolute Error', mean\_absolute\_error(svm\_test\_predictions, y\_test))

print('Mean Squared Error', mean\_squared\_error(svm\_test\_predictions, y\_test))

**Results:**

Mean Absolute Error 6519254.72971799

Mean Squared Error 42706579393991.33

## Future predictions

In this part we are going to discuss about the future predictions for the next 20 days done by different models.

The ability to predict the spread of COVID-19 using machine learning models can help public health officials and governments respond more effectively to the pandemic, protect vulnerable populations, and ultimately save lives.

Predicting future events related to COVID-19 using machine learning models can be important for a number of reasons:

**Planning:** Predictive models can be used to forecast the spread of the virus and the number of cases, hospitalizations, and deaths. This can help public health officials and governments plan for future outbreaks and allocate resources more effectively.

**Early warning:** Predictive models can be used to identify potential outbreaks before they occur, allowing for early intervention and prevention measures.

**Identifying at-risk populations:** Predictive models can be used to identify populations that are at a higher risk of contracting the virus, such as older adults or those with underlying health conditions.

**Identifying risk factors:** Predictive models can be used to identify factors that contribute to the spread of the virus, such as travel patterns, population density, and social distancing measures.

**Evaluating interventions:** Predictive models can be used to evaluate the effectiveness of different interventions, such as social distancing measures, travel restrictions, and vaccination programs.

**Prioritizing resources:** Predictive models can be used to prioritize the allocation of resources, such as personal protective equipment and hospital beds, to areas where they are needed most.

**Personalization:** Predictive models can be used to provide personalized recommendations, based on an individual's risk factors, for reducing the risk of contracting the virus.

For this paper, we have done predictions for the next 20 days that is from 01/01/2023 till 01/20/2023.

### Predictions from polynomial regression model

Predicting the number of cases in future for next 20 days

linear\_predictions = model.predict(pol\_future\_forecast)

**Results:**

Table

Description automatically generated

Fig. 6.4 Polynomial Regression Future Predictions

### Predictions from Bayesian Ridge model

Predicting the number of cases in future for next 20 days

baye\_pred = baye\_confirmed.predict(bay\_future\_forecast)

**Results:**

Table

Description automatically generated

Fig. 6.5 Bayesian Ridge Future Predictions

### Predictions from SVM model

Predicting the number of cases in future for next 20 days

svm\_predictions = svm.predict(forecast\_future)

**Results:**

**Table

Description automatically generated**

Fig. 6.6 SVM Future Predictions

## Comparing different models

Comparing different machine learning models is important for several reasons: hey

Model selection: By comparing different models, it's possible to select the best model for a specific problem. This can lead to better performance and more accurate predictions.

Identifying strengths and weaknesses: Comparing different models can help to identify the strengths and weaknesses of each model, allowing for a more informed decision about which model to use for a specific problem.

Understanding the problem: Comparing different models can also help to better understand the problem and the underlying data. It can help to identify patterns and trends in the data, which can be used to improve the performance of the model.

**The following graph shows the comparison between the testing data and the predictions done by different models.**

**Chart, line chart

Description automatically generated**

Fig. 6.7 Predictions Comparison

**In conclusion from the above comparisons, it could be seen that polynomial regression is best suitable for our research. In the prediction graph it could be seen that the polynomial regression predictions are the closest to the testing data. Also for the future predictions for the next 20 days, Polynomial regression’s predictions are the closest to the actual values as compared to the other 2 models. And finally, Polynomial regression has the lowest MAE & MSE values.**

## Discussion

After analyzing the situation, it has been found that using machine learning techniques has been successful in creating models for predicting the COVID-19 cases since its outbreak. However, these models have shown discrepancies due to challenges such as data uncertainty. To improve the accuracy of these models, these issues need to be addressed. One of the major problems is the limited access to medical image and text data sets due to their segregation in different regions of the world. Additionally, machine learning methods require larger data sets for more accurate results. Therefore, controlling data sources is a crucial task to build reliable models. The results of the forecasting algorithms are often inadequate because they are based on information from the internet. To enhance the performance of the models, acquiring real-world data sets is crucial.

# Chapter 7: Conclusion and Future Work

## Conclusion

Investigators are constantly searching for new challenges in various functional areas and are actively working to address them. The COVID-19 pandemic is the most recent topic of study to see significant development in recent years. As of June 15, 2020, COVID-19 had affected around 8.1 million people and caused 436,276 deaths worldwide, according to Worldometer. Due to the high number of reported cases, it is considered the deadliest disease in the world. The pandemic has had a negative impact on social, economic, political and religious development globally. To improve the healthcare and government response, it is necessary to examine various forecasting and prediction techniques. Despite the use of various techniques, research and measures taken by the public and medical professionals, the number of cases continue to rise. These measures have only temporarily slowed the spread of the virus, but have made significant contributions to its suppression. The end of the epidemic is uncertain. The unique nature of the virus, its easy transmission, wide range of symptoms, and global reach have caused concern worldwide. The pandemic may eventually end not due to a decrease in cases, but due to people becoming accustomed to living with it, as has been seen with previous pandemics. Predictions on the progression of COVID-19 are provisional, but the eventual outcome will likely involve a combination of social preventative measures, new medications to alleviate symptoms and a vaccine. While some researchers work alone, most must collaborate across multiple disciplines to improve treatments. A vaccine is the only long-term option for controlling the outbreak and its development may take months or even years. Without access to the vaccine globally, the virus will likely continue to spread to those who have not yet been affected.

The purpose of the study was towards tracing the evaluation public opinion regarding COVID-19 add to predict the number of coronavirus positive cases all around the globe. To reach these goals techniques few techniques were applied like Polynomial Regression, Bayesian Ridge and SVM. No accurate outcomes were obtained but polynomial regression predicted the closest values.

**The following table shows the comparison between the actual values over next 20 days and the values predicted by different models.**

A picture containing table

Description automatically generated

Fig. 7.1 Future Cases Prediction Comparison

From the above figure we could see that even though there is a huge difference between the predicted values and the actual values but the predictions made by polynomial regression model were the closest to the actual values.

**The following table shows the comparison of MAE and MSE values between different models.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Model** | **MAE** | **MSE** |
| 1 | **Polynomial** | 451832.86442587717 | 294329520918.9113 |
| 2 | **Bayesian Ridge** | 9949144.572066076 | 100286372600178.38 |
| 3 | **SVM** | 6519254.72971799 | 42706579393991.33 |

Table 7.1 MAE and MSE Comparison

Lower the MAE and MSE, better than results are. From the above table it is concluded that the Polynomial Regression model has the lowest mean absolute error and mean squared error values as compared to Bayesian Ridge and Support Vector Machine models.

It is important to note that these predictions are not infallible and must be interpreted with caution. It is important to consider these factors when selecting and adjusting algorithms for a specific task, as different models may perform better or worse depending on the characteristics of the data and the task at hand. Additionally, it is important to ensure that the algorithms are not biased and that ethical considerations are considered when using them for decision making.

## Limitations

**Limited understanding of the underlying dynamics:** Machine learning models can only predict the future based on the patterns and relationships found in the historical data. If there is a change in the underlying dynamics of the disease transmission, the model may not be able to capture it and make accurate predictions.

**Lack of interpretability:** Some machine learning models, such as deep learning models, can be difficult to interpret and understand. This can make it difficult to understand why a model is making certain predictions and to identify any errors in the model.

**Complexity of the disease:** COVID-19 is a complex disease with many factors affecting its spread. Including all these factors in a single model can be hard and may lead to overfitting.

**Limited ability to predict future events:** Some important factors that influence the spread of the disease, such as changes in public behavior and government policies, are hard to predict. Thus, the model's ability to predict future cases may be limited.

## Future Work

**Incorporating more data:** More data can be used to train the model, such as data on travel patterns, economic conditions, and demographics. This can help the model to make more accurate predictions.

**Ensemble models:** Using an ensemble of different models can help to improve the overall accuracy of the predictions. This is because different models can capture different patterns and relationships in the data.

**Incorporating causal inference:** Incorporating causal inference techniques can help to understand the underlying mechanisms that drive the spread of the disease, which can lead to better predictions.

**Incorporating domain knowledge:** Incorporating domain knowledge about the disease, such as the biology of the virus and the epidemiology of the disease, can help to improve the accuracy of the predictions.

**Incorporating real-time data:** Incorporating real-time data such as social media data and news articles can help to capture the changing dynamics of the disease spread and improve predictions in near future.

To get the accuracy nearest to 100%, a whole lot of system expertise is needed as the covid data is very tedious to deal with because of the data uncertainty. LSTM models have many hyperparameters that can affect their performance, such as the number of hidden units and the dropout rate. Hyperparameter tuning can help to find the optimal values for these parameters and improve the model's performance. An LSTM model can be trained to make multi-step predictions, such as predicting the number of cases for the next week or month. This can help to provide more accurate predictions and give a better idea of the future trend.

# Chapter 8: REFERENCES

Pal, R., Sekh, A.A., Kar, S. and Prasad, D.K., 2020. Neural network based country wise risk prediction of COVID-19. *Applied Sciences*, *10*(18), p.6448.

Vega, R., Flores, L. and Greiner, R., 2022. SIMLR: Machine Learning inside the SIR model for COVID-19 Forecasting. *Forecasting*, *4*(1), pp.72-94.

Liu, H., Manzoor, A., Wang, C., Zhang, L. and Manzoor, Z., 2020. The COVID-19 outbreak and affected countries stock markets response. *International Journal of Environmental Research and Public Health*, *17*(8), p.2800.

Dong, E., Du, H. and Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), pp.533-534.

Alyasseri, Z. A. A., Al-Betar, M. A., Doush, I. A., Awadallah, M. A., Abasi, A. K., Makhadmeh, S. N., Alomari, O. A., Abdulkareem, K. H., Adam, A., Damasevicius, R., Mohammed, M. A., & Zitar, R. A. (2022). Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. Expert systems, 39(3), e12759. <https://doi.org/10.1111/exsy.12759>

Zou, D., Wang, L., Xu, P., Chen, J., Zhang, W. and Gu, Q., 2020. Epidemic model guided machine learning for COVID-19 forecasts in the United States. *Medrxiv*.

Huang, S., Yang, J., Fong, S. and Zhao, Q., 2021. Artificial intelligence in the diagnosis of COVID-19: Challenges and perspectives. *International Journal of Biological Sciences*, *17*(6), p.1581.

Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T. and Atkinson, P.M., 2020. Covid-19 outbreak prediction with machine learning. *Algorithms*, *13*(10), p.249.

Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V. and Alhyari, S., 2020. COVID-19 prediction and detection using deep learning. *International Journal of Computer Information Systems and Industrial Management Applications*, *12*(June), pp.168-181.

*Alam, B., 2022. Implementation Of Support Vector Machine (SVM) Using Python*.

*Selvi, T., 2021. Why -Polynomial Regression and not Linear Regression?*

Watson, G.L., Xiong, D., Zhang, L., Zoller, J.A., Shamshoian, J., Sundin, P., Bufford, T., Rimoin, A.W., Suchard, M.A. and Ramirez, C.M., 2021. Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLoS computational biology*, *17*(3), p.e1008837.

Ye, H., Wu, P., Zhu, T., Xiao, Z., Zhang, X., Zheng, L., Zheng, R., Sun, Y., Zhou, W., Fu, Q. and Ye, X., 2021. Diagnosing coronavirus disease 2019 (COVID-19): Efficient Harris Hawks-inspired fuzzy K-nearest neighbor prediction methods. *Ieee Access*, *9*, pp.17787-17802.

Wang, B., Jin, S., Yan, Q., Xu, H., Luo, C., Wei, L., Zhao, W., Hou, X., Ma, W., Xu, Z. and Zheng, Z., 2021. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. *Applied Soft Computing*, *98*, p.106897.

Wang, P., Zheng, X., Ai, G., Liu, D. and Zhu, B., 2020. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. *Chaos, Solitons & Fractals*, *140*, p.110214.

Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E. and Bhattacharya, J., 2020. Seroprevalence of SARS-CoV-2–specific antibodies among adults in Los Angeles County, California, on April 10-11, 2020. *Jama*, *323*(23), pp.2425-2427.

Wang, H., 2021. Estimation of Total and Excess Mortality due to COVID-19. *Institute for Health Metrics and Evaluation, April*, *22*.

Clements, J., Workshop on aluminum in vaccines. *Presented by National Vaccine Program Office, Department of Health and Human Services. San Juan, Puerto Rico (May 11-12, 2000). As noted in David Ayoub’s presentation (see Note 14)*.

Berríos-Torres, S.I., Umscheid, C.A., Bratzler, D.W., Leas, B., Stone, E.C., Kelz, R.R., Reinke, C.E., Morgan, S., Solomkin, J.S., Mazuski, J.E. and Dellinger, E.P., 2017. Centers for disease control and prevention guideline for the prevention of surgical site infection, 2017. *JAMA surgery*, *152*(8), pp.784-791.