TRACK:
DATA SCIENCE

CAPSTONE PROJECT:
PREDICTIVE MODELLING FOR COVID-19 IN PUBLIC HEALTH

BY

NAME:
CHESTO EMMANUEL IBRAHIM
ID: FE/23/84581344

NOVEMBER, 2024.

CAPSTONE PROJECT:
PREDICTIVE MODELLING FOR COVID-19 IN PUBLIC HEALTH

BY

NAME:
CHESTO EMMANUEL IBRAHIM
ID: FE/23/84581344

NOVEMBER, 2024.

A CAPSTONE PROJECT REPORT SUBMITTED TO THREE MILLION TECHNICAL TALENT PROGRAM OF NATIONAL INFORMATION TECHNOLOGY DEVELOPMENT AGENCY, FEDERAL MINISTRY OF COMMUNICATIONS, INNOVATION AND DIGITAL ECONOMY ABUJA. IN PARTIAL FULFILMENT TO THE REQUIREMENT FOR THE AWARD OF CERTIFICATE IN DATA SCIENCE.

DECLARATION

I CHESTO, Emmanuel Ibrahim declare that this report is my own original work and that it has not been presented anywhere for any certification award.

---------- Chesto Emmanuel Ibrahim ----------

E.Signature

----------November 2024----------

Date

APPROVAL PAGE

We declare that this Capstone Project on Predictive Modelling for COVID-19 in Public Health is written by Chesto Emmanuel Ibrahim and carried out under the supervision of Sunnah Tech Lab, Sunnah Private School Unguwar Rimi Jos. We have read, examined and found it acceptable for the award of Certificate.

Approved by

Mr. Ibrahim Khaleel	November, 2024
Instructor	Date
Mr. Tech Abdulhafeez	November, 2024
3MTT Fellow Guide	Date
Mr. Ibrahim Nasir	November, 2024
CEO Sunnah Tech Lab	Date

DEDICATION

I would like to dedicate this work to God Guidance, my siblings for care and love that strengthened me through these tough weeks and to my closed friends for their support, prayers and care.

ACKNOWLEDGEMENT

First and famous. My appreciation and sincere thanksgiving to God Almighty for helping me to conduct my training and be able to write this report because without Him it is impossible to do anything.

Secondly, I would like to express my appreciation to my Instructor Mr. Ibrahim Khaleel for his guidance throughout these twelve (12) weeks. It was not possible to produce this report alone but through his guidance all work was accomplished.

Lastly, I would like to appreciate the NITDA, FMCI&DC, Sunnah Tech Lab and Staff members. My friends and 3MTT Fellows during the whole period of my Training. I really appreciate their support they offered.

TABLE OF CONTENTS

ABSTRACT

The Capstone Project focuses on creating a Predictive Modelling for COVID-19 in Public Health on how predictive modeling will be use in predicting the spread of the virus and understanding key factors that influence transmission and patient outcomes. This study was conducted in Sunnah Tec Lab Anguwar Rimi Jos Plateau State. In this work both **Data Collection (Dataset**: Access the COVID-19 Open Research Dataset (CORD-19) on Kaggle, which includes COVID-19 case counts, demographic data, and various health metrics and **Dataset Source**: CORD-19 on Kaggle). **Data Preprocessing (Cleaning**: Address missing values, remove duplicates, and standardize date and location formats, **Transformation**: Normalize data for machine learning models, ensuring consistency across all numerical features, and **Feature Engineering**: Create derived variables, such as daily growth rates, mortality ratios, and cases per population to enrich the dataset and strengthen model insights.) **Exploratory Data Analysis** (**Objective**: Conduct EDA to uncover trends, correlations, and outliers in the data, **Visualizations**: Use charts like line plots, bar charts, and scatter plots to analyze COVID-19 trends, such as case and mortality rates over time, and **Key Insights**: Focus on identifying demographic and environmental factors that could influence the spread and severity of COVID-19 cases.) **Model Development** (**Machine Learning Models**: Apply predictive models such as: Time-Series Models, Classification Models, **Evaluation**: Assess model performance with accuracy, precision, recall, F1-score, or RMSE, as applicable) and **Data Visualization and Reporting** (**Visualization**: Create clear, informative visualizations of model predictions and EDA findings, using libraries like Matplotlib and Seaborn. and **Reporting**: Present results in a structured report, with narratives and visuals that effectively communicate insights to non-technical stakeholders. The results revealed that The model should be coordinated, integrated, organized, systematic, clear and comprehensive to accomplish optimal result. As it is in the computer world-garbage in garbage out – so it is with Modelling. It forecast are inaccurate so will be making projections and predictions resulting in bad management decision to the detriment of the firm. A manager must be cautious when analyzing past experiences. It is recommended that, as discussed, surveyed and considered the work in this capstone project above. It seemed that necessary solutions will be needed to help brings out clean pictures that will fasten Predictive Modelling System in an organization are as follows:

1. The model would help in predicting the spread of COVID-19 which will entail how the organization will deploy preventive measures to curtail the spread.
2. The model would help in identifying the key factors that influences COVID-19 which will help how the organization will mitigate the diseases spread.
3. The model would also help in patient outcomes of COVID-19 which the organization will be well informed of the prevalence from the disease spread and make the necessary preventive measures to curtail the spread and improve health resource allocation. With above recommendation, the organization has adequate information that requires actionable insights to inform policies, anticipate future outbreaks, and improve health resource allocation.

CHAPTER ONE

CASE SCENARIO

1.0 Introduction

1.1 Background to the Study

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by the coronavirus SARS-CoV-2. The first known case was identified in Wuhan, China, in December 2019. Most scientists believe the SARS-CoV-2 virus entered into human populations through natural zoonosis, similar to the SARS-CoV-1 and MERS-CoV outbreaks, and consistent with other pandemics in human history. Social and environmental factors including climate change, natural ecosystem destruction and wildlife trade increased the likelihood of such zoonotic spillover. The disease quickly spread worldwide, resulting in the COVID-19 pandemic.

The symptoms of COVID-19 are variable but often include fever, fatigue, cough, breathing difficulties, loss of smell, and loss of taste. Symptoms may begin one to fourteen days after exposure to the virus. At least a third of people who are infected do not develop noticeable symptoms. Of those who develop symptoms noticeable enough to be classified as patients, most (81%) develop mild to moderate symptoms (up to mild pneumonia), while 14% develop severe symptoms (dyspnea, hypoxia, or more than 50% lung involvement on imaging), and 5% develop critical symptoms (respiratory failure, shock, or multiorgan dysfunction). Older people are at a higher risk of developing severe symptoms. Some complications result in death. Some people continue to experience a range of effects (long COVID) for months or years after infection, and damage to organs has been observed. Multi-year studies are underway to further investigate the long-term effects of the disease.

COVID-19 transmission occurs when infectious particles are breathed in or come into contact with the eyes, nose, or mouth. The risk is highest when people are in close proximity, but small airborne particles containing the virus can remain suspended in the

air and travel over longer distances, particularly indoors. Transmission can also occur when people touch their eyes, nose or mouth after touching surfaces or objects that have been contaminated by the virus. People remain contagious for up to 20 days and can spread the virus even if they do not develop symptoms. Testing methods for COVID-19 to detect the virus's nucleic acid include real-time reverse transcription polymerase chain reaction (RT-PCR), transcription-mediated amplification, and reverse transcription loop-mediated isothermal amplification (RT-LAMP) from a nasopharyngeal swab. Several COVID-19 vaccines have been approved and distributed in various countries, many of which have initiated mass vaccination campaigns. Other preventive measures include physical or social distancing, quarantining, ventilation of indoor spaces, use of face masks or coverings in public, covering coughs and sneezes, hand washing, and keeping unwashed hands away from the face. While drugs have been developed to inhibit the virus, the primary treatment is still symptomatic, managing the disease through supportive care, isolation, and experimental measures.

1.2 Statement of Problem

In response to the COVID-19 pandemic, public health organizations have faced immense challenges in predicting the spread of the virus and understanding key factors that influence transmission and patient outcomes. Imagine you have been hired as a data scientist by a public health organization, "

1.3 Research Questions

Using historical COVID-19 data, I will conduct data cleaning, perform exploratory data analysis (EDA), and develop predictive models to forecast COVID-19 trends.

I will present my findings through visualizations and provide a final report summarizing insights and recommendations for public health responses.

1.4 Objective of the Study

1. To build a predictive modeling system. The organization requires actionable insights

2. To inform policies, anticipate future outbreaks, and improve health resource allocation.

1.5 Significance of the Study

The present work has numerous importance's some of which are considered in this context are:

a) To build a predictive modeling system that requires actionable insights
b) It has the significances of performance tracking, measures progress toward outbreak etc.
c) To inform policies, anticipate future outbreaks, and improve health resource allocation.

1.6 Scope of the Study

The scope of the study will cover COVID-19. The study is limited to only Sunnah Tech Lab Anguwn Rimi Jos Plateau State out of the remaining the entire world and "HealthGuard Analytics.

Sunnah Tech Lab Anguwn Rimi Jos Plateau State is very important to the study because is very close to the researcher which will help in gathering information and materials as well as save me time, energy and money at the cause of this report.

1.7 Definition of Terms

Data: information, especially in a scientific or computational context or with the implication that it is organized.

Forecast: to estimate how something will be in the future.

Health: the state of being free from physical or psychological disease, illness or malfunction, wellness.

Modelling: Model a simulation used to analyze the workings of real world system or event.

Predictive: expressing the expected accuracy of statistical measure or of diagnostic test.

Public: open to all members of a community, as opposed to only a segment of it; especially, provided by national or local authorities and supported by money from taxes.

Trends: a line drawn on a graph that approximates the trend of a number of disparate points.

CHAPTER TWO

METHODOLOGY

2.0 Introduction

The methodology is the study of methods used in field. Also is a collection of methods, practices, procedures and rules used by those who work in some field.

2.1 Data Collection

- **Dataset**: to Access the COVID-19 Open Research Dataset (CORD-19) on Kaggle, which includes COVID-19 case counts, demographic data, and various health metrics.
- **Dataset Source**: CORD-19 on Kaggle

2.2 Data Preprocessing

- **Cleaning**: Address missing values, remove duplicates, and standardize date and location formats.

● **Transformation**: Normalize data for machine learning models, ensuring consistency across all numerical features.

● **Feature Engineering**: Create derived variables, such as daily growth rates, mortality ratios, and cases per population to enrich the dataset and strengthen model insights.

2.3 Exploratory Data Analysis

- **Objective**: Conduct EDA to uncover trends, correlations, and outliers in the data.
- **Visualizations**: Use charts like line plots, bar charts, and scatter plots to analyze COVID-19 trends, such as case and mortality rates over time.
- **Key Insights**: Focus on identifying demographic and environmental factors that could influence the spread and severity of COVID-19 cases.

2.4 Model Development

- **Machine Learning Models**: Apply predictive models such as:
  - Time-Series Models.
  - Classification Models
- **Evaluation**: Assess model performance with accuracy, precision, recall, F1-score, or RMSE, as applicable.

2.5 Data Visualization and Reporting

- **Visualization**: Create clear, informative visualizations of model predictions and EDA findings, using libraries like Matplotlib and Seaborn.
- **Reporting**: Present results in a structured report, with narratives and visuals that effectively communicate insights to non-technical stakeholders.

CHAPTER THREE

ASSESSMENT CRITERIA

3.0 Selected Sub-Headings

3.0.1 Data Preparation

Proficiency in data cleaning, transformation, and feature engineering.

3.0.2 EDA and Insights

Depth of EDA, relevance of insights, and quality of visualizations used to communicate findings.

3.0.3 Model Performance

Evaluation of model suitability and accuracy using appropriate metrics

3.0.4 Clarity and Presentation

Quality of the final report, visualization clarity, and coherence of findings and insights.

CHAPTER FOUR

SUBMISSION DETAILS

4.0 Deliverables

1. **Technical Report**: A detailed report covering data preparation, EDA, model development, model evaluation, and key insights with visualizations.
2. **Code Repository**: A GitHub repository with organized, well-documented code for all phases of the project.
3. **Presentation**: A slide deck or recorded presentation (10–15 minutes) summarizing findings, model insights, and public health implications.

4.1 Submission Format

- **Report**: Submit as a PDF or Word document.
- **Code**: Provide a GitHub link or a .zip file containing the project code.
- **Presentation**: Submit as PowerPoint, Google Slides, or a video file (MP4).

4.2 Data Resources

All participants will use the COVID-19 Dataset (CORD-19) available on Kaggle.

Here is the URL to the dataset:

https://www.kaggle.com/datasets/imdevskp/corona-virus-report

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.0 Summary

A platform was set in chapter one upon which the project would be focused: Predictive Modelling for COVID-. The introduction, background, statement of the problem all highlighted the issues facing COVID-19 the methodology and assessment criteria were discussed also in chapter two and three of this report. Chapter four gave us a vivid picture of the submission details used to collect, analyze and help present the data. In chapter five, Summary, conclusion and recommendation based on the result will be made in this chapter.

5.1 Conclusion

The model should be coordinated, integrated, organized, systematic, clear and comprehensive to accomplish optimal result.

As it is in the computer world-garbage in garbage out – so it is with Modelling. It forecast are inaccurate so will be with the making of projections and predictions resulting in bad management decision to the detriment of the firm. A manager must be cautious when analyzing past experiences.

5.2 Recommendation

As discussed, surveyed and considered the work in this capstone project above. It seemed that necessary solutions will be needed to help brings out clean pictures that will fasten Predictive Modelling System in an organization are as follows:

1. The model would help in predicting the spread of COVID-19 which will entail how the organization will deploy preventive measures to curtail the spread.

2. The model would help in identifying the key factors that influences COVID-19 which will help how the organization will mitigate the diseases spread.
3. The model would also help in patient outcomes of COVID-19 which the organization will be well informed of the prevalence from the disease spread and make the necessary preventive measures to curtail the spread and improve health resource allocation.

With above recommendation, the organization has adequate information that requires actionable insights to inform policies, anticipate future outbreaks, and improve health resource allocation.

REFERENCE

1. "Covid-19". Oxford English Dictionary (Online ed.). Oxford University Press. April 2020. Retrieved 15 April 2020. (Subscription or participating institution membership required.)
2. "Symptoms of Coronavirus". U.S. Centers for Disease Control and Prevention (CDC). 13 May 2020. Archived from the original on 17 June 2020. Retrieved 18 June 2020.
3. "Q&A on coronaviruses (COVID-19)". World Health Organization (WHO). 17 April 2020. Archived from the original on 14 May 2020. Retrieved 14 May 2020.
4. Jump up to:[a] [b] Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, et al. (2020–2024). "Coronavirus Pandemic (COVID-19)". Our World in Data. Retrieved 17 November 2024.
5. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, et al. (5 March 2020). "Coronavirus Pandemic (COVID-19)". Our World in Data. Archived from the original on 24 February 2024. Retrieved 24 February 2024.
6. "The pandemic's true death toll". The Economist. 28 August 2023 [2 November 2021]. Retrieved 28 August 2023.
7. Page J, Hinshaw D, McKay B (26 February 2021). "In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market – The man with the first confirmed infection of the new coronavirus told the WHO team that his parents had shopped there". The Wall Street Journal. Retrieved 27 February 2021.
8. Jump up to:[a] [b] [c] Pekar J (26 July 2022). "The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2". Science. **377** (6609): 960–966. Bibcode:2022Sci...377..960P. doi:10.1126/science.abp8337. PMC 9348752. PMID 35881005.
9. Jump up to:[a] [b] [c] Jiang X, Wang R (25 August 2022). "Wildlife trade is likely the source of SARS-CoV-2". Science. **377** (6609): 925–926. Bibcode:2022Sci...377..925J. doi:10.1126/science.add8384. PMID 36007033. S2CID 251843410. Retrieved 20 November 2022.
10. Jump up to:[a] [b] Terrestrial and Freshwater Ecosystems and Their Services. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (PDF). IPCC. 2022. pp. 233–235. Retrieved 14 March 2023.
11. Jump up to:[a] [b] Health, Wellbeing, and the Changing Structure of Communities. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (PDF). IPCC. 2022. pp. 1067–1070. Retrieved 14 March 2023.
12. Jump up to:[a] [b] "Climate change may have driven the emergence of SARS-CoV-2". University of Cambridge. Science of the Total Environment. 5 February 2021. Retrieved 14 March 2023.
13. Jump up to:[a] [b] "Climate change the culprit in the COVID-19 pandemic". European Commission. Retrieved 24 March 2023.
14. Islam MA (April 2021). "Prevalence and characteristics of fever in adult and paediatric patients with coronavirus disease 2019 (COVID-19): A systematic review and meta-analysis of 17515 patients". PLOS ONE. **16** (4):

e0249788. Bibcode:2021PLoSO..1649788I. doi:10.1371/journal.pone.0249788. PMC 80 23501. PMID 33822812.

15. Saniasiaya J, Islam MA (April 2021). "Prevalence of Olfactory Dysfunction in Coronavirus Disease 2019 (COVID-19): A Meta-analysis of 27,492 Patients". The Laryngoscope. **131** (4): 865–878. doi:10.1002/lary.29286. ISSN 0023-852X. PMC 7753439. PMID 33219539.
16. Saniasiaya J, Islam MA (November 2020). "Prevalence and Characteristics of Taste Disorders in Cases of COVID-19: A Meta-analysis of 29,349 Patients" (PDF). Otolaryngology–Head and Neck Surgery. **165** (1): 33–42. doi:10.1177/0194599820981018. PMID 33320033. S2CID 229174644.
17. Agyeman AA, Chin KL, Landersdorfer CB, Liew D, Ofori-Asenso R (August 2020). "Smell and Taste Dysfunction in Patients With COVID-19: A Systematic Review and Meta-analysis". Mayo Clin. Proc. **95** (8): 1621–1631. doi:10.1016/j.mayocp.2020.05.030. PMC 7275152. PMID 32753137.
18. Wang B, Andraweera P, Elliott S, Mohammed H, Lassi Z, Twigger A, et al. (March 2023). "Asymptomatic SARS-CoV-2 Infection by Age: A Global Systematic Review and Meta-analysis". The Pediatric Infectious Disease Journal. **42** (3): 232–239. doi:10.1097/INF.0000000000003791. PMC 9935239. PMID 36730054. Retrieved 15 November 2023.
19. Oran DP, Topol EJ (January 2021). "The Proportion of SARS-CoV-2 Infections That Are Asymptomatic: A Systematic Review". Annals of Internal Medicine. **174** (5): M20-6976. doi:10.7326/M20-6976. PMC 7839426. PMID 33481642.
20. "Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19)". U.S. Centers for Disease Control and Prevention (CDC). 6 April 2020. Archived from the original on 2 March 2020. Retrieved 19 April 2020.
21. Jump up to:[a b c] Davis HE, McCorkell L, Vogel JM, Topol EJ (March 2023). "Long COVID: major findings, mechanisms and recommendations". Nature Reviews. Microbiology. **21** (3): 133–146. doi:10.1038/s41579-022-00846-