UNCOVERING INSIGHTS AND TRENDS: A COMPREHENSIVE ANALYSIS OF COVID-19 PANDEMIC IN MEXICO

Project Report submitted to the

SDM Post Graduate Centre, Ujire

in partial fulfilment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

by

Navya Kamath

Under the supervision of

Asst. Prof. Ms. Supriya Shivadasan Padmavati

Department of PG Studies in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE - 574240

Karnataka, INDIA

AUGUST 2023

SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE
(AUTONOMOUS)
UJIRE - 574240

DEPARTMENT OF STATISTICS

CERTIFICATE

Certified that this is the bonafide record of project work done by Ms. Navya Kamath during the year 2023 as a part of her M.Sc (Statistics) fourth semester course work.

Reg. No.	**2**	**1**	**4**	**0**	**0**	**5**

Project Guide **Head of the Department**

Examiner

1.

2.

Date:
Place: Ujire

DECLARATION

I, Navya Kamath, hereby declare that the matter embodied in this report entitled '**Uncovering Insights and Trends: A Comprehensive Analysis of COVID-19 pandemic in Mexico**' is a bonafide record of project work carried out by me under the guidance and supervision of **Asst. Prof. Ms. Supriya Shivadasan Padmavati**, Department of Statistics, SDM College, Ujire - 574240, Karnataka, India. I further declare that no part of the work contained in the report has previously been formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or any other similar title or recognition of any other university.

Date:
Place: Ujire

(NAVYA KAMATH)
E-mail: maynavya@gmail.com

CERTIFICATE

This is to certify that the project report entitled '**Uncovering Insights and Trends: A Comprehensive Analysis of COVID-19 pandemic in Mexico**' is a bonafide record of an authentic work carried out by **Navya Kamath**, under my guidance and supervision in the Department of Post Graduate Studies and Research in Statistics, SDM College, Ujire, in partial fulfilment of the requirements for the award of the degree of Master of Science in Statistics, under Mangalore University, Mangalagangothri. I further certify that this report or part thereof has not previously been presented or submitted elsewhere for the award of any Degree, Diploma, Associateship, Fellowhsip or any other similar title of any other institution or university.

Date:
Place: Ujire

(Supriya Shivadasan Padmavati)
E-mail: supriyasp@sdmcujire.in

ACKNOWLEDGEMENTS

Firstly, I would like to thank our Principal ***Dr. B.A. Kumara Hegde*** *for providing the necessary facilities for the completion of this project work in our college.*

I would also thank our Dean ***Dr. Vishwanatha P.*** *for his support.*

It is my privilege to thank our HOD ***Prof. Shanthiprakash*** *for his suggestions and support.*

I am very grateful to my Research Supervisor, ***Asst. Prof. Ms. Supriya Shivadasan Padmavati****, Department of Statistics, SDM College, Ujire, for her kind help and encouragement throughout my project work.*

I gratefully acknowledge my teachers at the Department of Statistics, SDM College, Ujire, ***Asst. Prof. Ms. Shwetha Kumari*** *and* ***Asst. Prof. Mr. Pradeep K*** *for their support during my project work.*

I am also thankful to all my family members and friends for their constant encouragement and help in each step.

My sincere thanks also goes to the students of SDM College, Ujire, who have helped me directly or indirectly during my project work.

*Finally to all who helped me in many ways, I say, '****Thank You****!'.*

(Navya Kamath)

Contents

1 Chapter 1 **1**

Introduction **1**
1.1 Motivation 5
1.2 Literature Review 6
1.3 Objectives 8
1.4 Scope of the study 8

2 Chapter 2 **9**

Methodology **9**
2.1 Introduction 9
2.1.1 About the data 9
2.2 Statistical Techniques used for Data Analysis 10
2.2.1 Chi-square test of Independence 10
2.2.2 Cramer V 11
2.2.3 Welch t-test 12
2.2.4 Kruskal Wallis test 13
2.2.5 Dunn test 14
2.2.6 Two sample Z-test for Proportions 14
2.2.7 Logistic Regression 15
2.2.8 Decision Tree 16
2.2.9 Random Forest 17
2.2.10 Extreme Gradient Boosting(XGBoost) 17
2.2.11 Multinomial Regression 18
2.2.12 ARIMA Model 19
2.2.13 Augmented Dickey-Fuller test 20
2.2.14 Ljung Box test 20

3 Chapter 3 **21**

Results and Discussion: Univariate Analysis **21**
3.1 Introduction 21
3.2 Characteristics of Sample Respondents 21
3.2.1 Gender 21

3.2.2 Age . . . 22
3.2.3 Classification . . . 23
3.2.4 Treatment Type . . . 24
3.2.5 Consumption of Tobacco . . . 24
3.2.6 Pneumonia, Diabetes, COPD, Asthma, Immunosuppression, Hypertension, Cardiovascular, Obesity, Renal chronic disease . 24
3.2.7 Death . . . 25
3.3 Conclusion . . . 27

4 Chapter 4 **28**

Results and Discussion: Bivariate Analysis **28**
4.1 Introduction . . . 28
4.2 Bivariate Analysis of factors related to Death . . . 28
4.2.1 Testing the association between Health conditions and Death using Chi-square test of independence . . . 28
4.2.2 Testing the association between Treatment type and Death using Chi-square test of independence . . . 37
4.2.3 Testing the association between consumption of Tobacco and Death using Chi-square test of independence . . . 39
4.2.4 Analysis of factors affecting the Death of COVID-19 patients using Proportionality test . . . 40
4.3 Bivariate Analysis of factors related to Age . . . 45
4.3.1 Analysis of factors affecting the Age of individuals with and without Health Conditions using Welch t-test . . . 45
4.3.2 Analysis of factors affecting the Age of individuals with different degrees of symptoms using Kruskal Wallis test . . . 47
4.4 Conclusion . . . 49

5 Chapter 5 **50**

Results and Discussion: Multivariate Analysis **50**
5.1 Introduction . . . 50
5.2 Multivariate Model for Classifying Severity Levels of Symptoms . . . 50
5.2.1 Multinomial Regression . . . 50
5.3 Analysis of factors influencing the death of COVID-19 patients . . . 54
5.3.1 Logistic Regression . . . 54
5.3.2 Decision Tree . . . 58

5.3.3 Random Forest 62
5.3.4 XGBoost 66
5.3.5 Model Evaluation 69
5.4 Forecasting Death 70
5.4.1 ARIMA Model 70
5.5 Conclusion 72

6 Chapter 6 **73**

Conclusion **73**
6.1 Conclusion 73
6.2 Summary 74

7 Bibliography **75**

8 Appendix **77**

List of Tables

1 Frequency distribution of COVID-19 cases by gender 21
2 Frequency distribution of age 22
3 Frequency distribution of Classification of severity level of symptoms 23
4 Frequency distribution of Intubation and ICU 24
5 Frequency distribution of Smoking habit 24
6 Frequency distribution of various health conditions 25
7 Frequency distribution of Death 25
8 Contingency table for Pneumonia Outcome: Death vs. Survival . . . 28
9 Contingency table for Diabetes Outcome: Death vs. Survival 29
10 Contingency table for COPD Outcome: Death vs. Survival 30
11 Contingency table for Pregnancy Outcome: Death vs. Survival 31
12 Contingency table for Asthma Outcome: Death vs. Survival 31
13 Contingency table for Immunosuppression Outcome: Death vs. Survival 32
14 Contingency table for Hypertension Outcome: Death vs. Survival . . 33
15 Contingency table for Cardiovascular Outcome: Death vs. Survival . 34
16 Contingency table for Renal chronic Outcome: Death vs. Survival . . 35
17 Contingency table for Obesity Outcome: Death vs. Survival 36
18 Contingency table for Other disease Outcome: Death vs. Survival . . 36
19 Contingency table for Intubed Outcome: Death vs. Survival 37
20 Contingency table for ICU Outcome: Death vs. Survival 38
21 Contingency table for Tobacco Outcome: Death vs. Survival 39
22 Table showing results of Levene's homogeneity of variance test 45
23 Table showing results of Welch t-test 46
24 Table showing the results of Dunn test 48
25 Frequency distribution of Classification of severity level of symptoms 50
26 Table showing the covariates of multinomial regression 51
27 Table showing the covariates of Logistic regression 54
28 Frequency distribution of Death 54
29 Table showing the covariates of Decision Tree 58
30 Table showing the covariates of Random Forest 62
31 Table showing the covariates of XGBoost 66
32 Table showing outputs of classification models 69
33 Comparing ARIMA model results 71

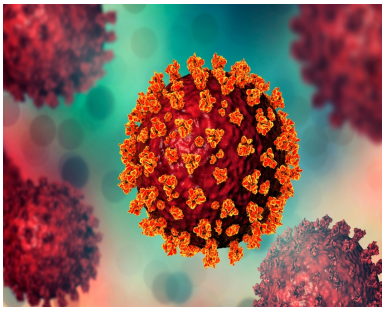
List of Figures

1 Gender of COVID-19 patients 21
2 Histogram of age of COVID-19 patients 22
3 Count plot of Severity level of symptoms 23
4 Count plot of Severity level the COVID-19 patients 26
5 Count plot of Pneumonia under Deceased and Survived 40
6 Count plot of Diabetes under Deceased and Survived 41
7 Count plot of Hypertension under Deceased and Survived 42
8 Count plot of Intubed in Deceased and Survived 43
9 Count plot of ICU in Deceased and Survived 44
10 Figure showing balanced response variable of Multinomial regression . 51
11 Figure showing the confusion matrix of multinomial regression 52
12 Figure showing permutation feature importance of Multinomial regression 53
13 Figure showing the balanced response variable of logistic regression . 55
14 Figure showing the confusion matrix of logistic regression 56
15 Figure showing permutation feature importance of Logistic regression 57
16 Figure showing the visual representation of Decision Tree 59
17 Figure showing the confusion matrix of decision tree classifier 60
18 Figure showing permutation feature importance of Decision tree classifier 61
19 Figure showing the visual representation of random forest 63
20 Figure showing the confusion matrix of random forest classifier 64
21 Figure showing permutation feature importance of Random forest classifier 65
22 Figure showing the confusion matrix of XGBoost classifier 67
23 Figure showing permutation feature importance of XGBoost classifier 68
24 Figure showing COVID-19 death over time 70
25 Figure showing COVID-19 death over time 71

1 Chapter 1

Introduction

COVID-19



COVID-19, short for Coronavirus Disease 2019, was first identified in December 2019 in Wuhan, China, and has since spread rapidly across the globe, leading to a worldwide pandemic. The World Health Organization (WHO) declared COVID-19 a global health emergency and later a pandemic on March 11, 2020.

COVID-19 is a respiratory illness caused by a novel coronavirus known as SARS-CoV-2. Coronaviruses are a family of viruses that cause illness such as respiratory diseases or gastrointestinal diseases. Respiratory diseases can range from the common cold to more severe diseases such as,

- Middle East Respiratory Syndrome (MERS-CoV)
- Severe Acute Respiratory Syndrome (SARS-CoV)

A novel coronavirus is a new strain which have not been identified in humans previously. Corona virus got its name from the way that it look under a microscope. The virus consists of a core of genetic material surrounded by an envelope with protein spikes. This gives it the appearance of a crown. The word Corona means "crown" in Latin. Corona viruses are zoonotic, meaning that the viruses are transmitted between animals and humans. It has been determined that MERS-CoV was transmitted from dromedary camels to humans and SARS-CoV from civet cats to humans. The source of the SARS-CoV-2 (COVID-19) is yet to be determined, but investigations are ongoing to identify the zoonotic source to the outbreak.

Clinical Presentation:

Typically Corona viruses present with respiratory symptoms. Among those who will become infected, some will show no symptoms and some people may have only few symptoms where as some may experience worsened symptoms. Those who do develop symptoms may have a mild to moderate symptoms similar to the seasonal flu. Symptoms may include:

1. Respiratory symptoms
2. Fever
3. Cough
4. Shortness of breath
5. Breathing difficulties
6. Fatigue
7. Sore throat

A minority group of people will have more severe symptoms and need to be hospitalized, most often with pneumonia, and in some instances, the illness can include ARDS, sepsis and septic shock. Emergency warning signs where immediate medical attention should be sought include:

1. Difficulty breathing or shortness of breath
2. Persistent pain or pressure in the chest
3. New confusion or inability to arouse
4. Bluish lips or face

Some people may experience symptoms for more than 4 weeks after they are diagnosed with COVID-19 such health issues are called post-COVID-19 conditions. In children multisystem inflamatory syndrome might affect some of the organs and tissues several weeks after having COVID-19.

High-Risk Populations:

The virus that causes COVID-19 infects people of all ages. However, evidence to date suggests that three groups of people are at a higher risk of getting severe COVID-19 disease:

- People aged above 70.
- People with serious chronic illnesses such as:
    1. Diabetes
    2. Cardiovascular disease
    3. Chronic respiratory disease
    4. Cancer
    5. Hypertension
    6. Chronic liver disease
- People who are physically inactive.

Transmission of COVID-19:

COVID-19 spreads when an infected person breathes out droplets and very small particles that contain the virus. These droplets and particles can be breathed in by other people or land on their eyes, nose, or mouth. In some circumstances, they may contaminate surfaces they touch. People who are closer than 6 feet from the infected person are most likely to get infected by the disease.
The risk of COVID-19 spreading is higher in:

1. Crowded places with many people nearby.
2. Close-contact settings, especially where people have conversations very near each other.
3. Confined and enclosed spaces with poor ventilation.

The incubation period of COVID-19 is between 2 to 14 days. This means that if a person remains well after 14 days after being in contact with a person with confirmed COVID-19, then that person is not infected.

Diagnostic Procedures:

A COVID-19 diagnostic testing kit are available in clinical testing labs. The gold standard for testing for COVID-19 is Reverse Transcription Polymerase Chain Reaction (RT-PCR). However, current data suggest that RT-PCR is only 30-70% effective for acute infection, this may be due to incorrect use of lab kits or not enough virus in the blood at the early stages of testing. Plus, the availability of testing will vary from country to country.
The Centers for Disease Control and Prevention(CDC) recommends that any person who may have had contact with a person who is suspected of having COVID-19 and develops a fever and respiratory symptoms listed above are advised to call their healthcare practitioner to determine the best of course of action. The main criteria for testing are:

- Location
- Age
- Medical history and risk factors
- Exposure to the virus and contact history
- Duration of symptoms

If the above criteria are met it is advised that the following testing procedure is followed:

- Collect and test upper respiratory tract specimens, using a nasopharyngeal swab.
- If available testing of lower respiratory tract specimens.
- If a productive cough is evident then a sputum specimen should be collected.

For patients who are receiving invasive mechanical ventilation, a lower respiratory tract aspirate or broncho-alveolar lavage sample should be collected.

1.1 Motivation

The global pandemic caused by the outbreak of Coronavirus disease (COVID-19) in late 2019 has had a devastating impact on global public health. The widespread transmission of the virus has resulted in a significant number of cases, hospitalizations, and fatalities. In just three years, more than 760 million cases have been recorded, with a staggering death toll of 6.9 million. Alarming predictions suggest that COVID-19 could be one to two times more deadly than seasonal influenza by 2023, considering the emergence of new variants such as Omicron.

Given the varying levels of herd immunity and vaccination rates across different countries, the severity of COVID-19 may differ in the coming years. This prediction, based on statistical analysis, serves as a valuable tool for decision-making and can set a precedent for forecasting infectious diseases. Understanding the potential outcomes of future outbreaks is crucial for developing effective strategies to protect and manage public health.

With COVID-19 presenting a wide range of symptoms that often resemble those of seasonal flu, accurate diagnosis can be challenging. This project aims to provide insights that can enhance our ability to safeguard public health and improve management strategies during similar outbreaks in the future.

1.2 Literature Review

In 2020, Kolla Bhanu Prakash and S.Sagar Imambi carried out a Analysis, Prediction and Evaluation of COVID-19 using Machine Learning algorithms. The main objective of this study was to understand the affect of COVID-19 on people, its confirmations and recovery predictions. The classification techniques such as Decision Tree, Gaussian Naive Bayes, Multilinear Regression, Logistic Regression, XGB Classifier, Support Vector Machine, KNN+NCA and Random Forest are used predict the recovery of COVID-19 patients more accurately. According to the study, Random forest classifier has outperformed other model with 99 percent Coefficient of Determination and with Accuracy of 97 percent.

In 2020, Malik Khizar Hayat and Ali Daud carried out Age and Gender based Analysis of Surveillance variables of COVID-19. The main aim of this study was to understand the significant relationship between age and gender, and COVID-19 surveillance variables. The graphical techniques and correlation analysis has been used under this study. The results of this study show that age has strong correlation with death and people aged more than 45 cover the larger portion of fatalities. Also the correlation between age group and number of deaths is statistically significant that can be used for death prediction based on number of people in particular age group.

In 2020, M. Rubaiyat Hossain Mondal and Subrato Bharati carried out Data Analysis for novel coronavirus disease. The main aim of this study is to automate the Machine learning algorithm for the diagnosis of COVID-19. The Machine Learning algorithm used are Support Vector Machine, KNN, Multilayer perceptron, Logistic Regression, Decision Tree, Random Forest and XGBoost. The results show that the MLP, XGBoost and LR can reliably classify COVID-19 patients.

In 2021, Subhranil Das and Rashmi Kumari carried out Statistical Analysis of COVID cases in India. The main objective of study is to provide whole summary of total case India. The GLM Poisson's Ratio as well as Exponential Ratio which are used in analyzing the results obtained from the scatter plot and histogram. The results show that the data acquired from cases concerned with India has played a prominent analysis by using measures such as mean, median, standard deviation.

In 2022, Ananthu James, Jyoti Dalal and Timokleia Kousi carried out an in-depth statistical analysis of COVID-19 pandemic's initial spread in the WHO Africa region. The main objective of this study whether the geographic, demographic and

socioeconomic factors have impact on the spread and severity of COVID-19. Under this study Principal Component Analysis and Regression Analysis techniques has been used. The study showed that the robust surveillance and testing capacities are needed to ensure that public health decisions are based on the data that depict the epidemiological situation accurately.

In 2020, Fatemeh Javanmardi carried out a Systematic review and meta analysis on the underlying diseses in died cases of COVID-19. The aim of the study is evaluating the prevalence of underlying disease in died people with COVID-19. Under this study varinace weighted method and Higgins I^2 and Cochrane Q statistics method are used. The results of this study showed that most prevalent comorbiditis were hypertension, diabetes, lung disease, liver disease, cardiovascular disease, malignancy, cerebrovascular disease, COPD and asthma.

In 2021, Sefer Elezkurtaj and Selina Greuel carried out a analysis on causes of death and comorbities in hospitalized patients with COVID-19. The main objective of this study is to check whether the preexisting health condition contribute to the mechanism of death. The analysis was done using IBM SPSS statistic. Under this study Kaplan-Meier method is used. The results showed that majority of cases with severe and fatal COVID-19 patients had died of this disease, although in the presence of multiple preexisting health conditions.

In 2021, Krishnan Bhaskaran carried out a population based cohor analysis of UK primary care data. The main aim of the study is investigating how specific factors are differently associated with COVID-19 mortality as compared to mortality from causes other than COVID-19. Under this study multinomial logistic regression, binary logistic regressiona and Cox regression model are used. The results of this study showed that COVID-19 largely multiplies existing risk faced by patients, with some notable exceptions.

1.3 Objectives

1. To check association between different health conditions in COVID-19 patients.
2. To compare the mean age of COVID-19 patients across different health conditions.
3. To compare the mean age of COVID-19 patients across various severity levels of symptoms.
4. To compare the proportion of COVID-19 patients with health conditions in the groups that survived and those that died.
5. To predict the severity level of patients based on their health conditions and treatment type.
6. To predict the death based on various health conditions and treatment type in COVID-19 patients.
7. To forecast the number of death related to COVID-19.

1.4 Scope of the study

The results of this project can be used as a source of basic information regarding COVID-19 disease, which will be helpful in medical researches to the doctors, researchers etc. Also it will be helpful for the Health Departments. And also these results will be helpful to assess the risk of COVID-19 transmission, predict disease trends and death toll in the coming years.

2 Chapter 2

Methodology

2.1 Introduction

This chapter contains details about the data and the methods used for collecting the data on COVID-19. It also includes information about the statistical tools and methods used for analyzing the data. In Section 2.2, the variables under study are described, while Section 2.3 provides details about the various statistical techniques used for data analysis.

2.1.1 About the data

A secondary data has been collected from the website "Gobierno De Mexico". This data is collected in the year 2023. The data contains 1338268 observations and 21 variables. Each variable corresponds to an individual COVID-19 patient. The description about the variables considered in the analysis are given as follows:

- **sex -** Male or Female.
- **admission date -** Date of admission of the patient to the care unit.
- **date symptoms -** Date on which patient started experiencing the symptoms.
- **date died -** Date on which the patient died.
- **intubed -** Patient required intubation or not(yes/no).
- **pneumonia -** Patient having pneumonia or not(yes/no).
- **age -** Age of an individual patient.
- **pregnant -** Patient is pregnant or not(yes/no).
- **diabetes -** Patient having diabetes or not(yes/no).
- **copd -** Patient having chronic obstructive pulmonary disease or not(yes/no).
- **asthma -** Patient having asthma or not(yes/no).

- **inmsurp -** Patient having immunosuppression or not(yes/no).
- **hipertension -** Patient having hypertension or not(yes/no).
- **other disease -** Patient having other diseases or not(yes/no).
- **cardiovascular -** Patient having cardiovascular disease or not(yes/no).
- **obesity -** Patient having obesity or not(yes/no).
- **renal chronic -** Patient having chronic renal failure or not(yes/no).
- **tobacco -** Patient having smoking habit or not(yes/no).
- **classification final -** Patient diagnosed with COVID-19 in different degrees (1-Mild, 2-Moderate, 3-Severe).
- **icu -** Patient required admission to an intensive care unit not(yes/no).
- **death -** Patient died or survived(yes/no).

2.2 Statistical Techniques used for Data Analysis

Tools Used:

The open source software and programming languages 'R' and 'Python' has been used to carry out the analysis of the data. The statistical methods considered in order to carry out the analysis are given as follows.

2.2.1 Chi-square test of Independence

The Chi-square test of independence is a statistical test used to determine whether there is a significant association between two categorical variables. It examines if the distribution of one variable is independent of the distribution of the other variable.

The hypothesis of this test are as follows,

H_0(Null Hypothesis): There is no association between the two categorical variables.
H_1(Alternative Hypothesis): There is an association between the two categorical variables.

Assumptions for the Chi-square test of independence:

1. Independent random samples: The data should be collected independently and randomly from the population.

2. Sample size: Each cell of the contingency table should have an expected frequency of at least 5.

3. Mutually exclusive categories: Each observation should belong to one and only one category for each variable.

Formula for the Chi-square test statistic:

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

Where χ^2 is Chi-square test statistic, O is Observed frequency (the actual count in each cell of the contingency table) and E is Expected frequency (the count that would be expected in each cell if the variables were independent)

The Chi-square test statistic follows a Chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table.

To perform the Chi-square test, you calculate the test statistic using the formula above, and then compare it to the critical value from the Chi-square distribution with the appropriate degrees of freedom. If the calculated test statistic is greater than the critical value, you reject the null hypothesis, indicating that there is evidence of an association between the variables. If the test statistic is not greater than the critical value, you fail to reject the null hypothesis, suggesting no significant association between the variables.

2.2.2 Cramer V

Cramer's V is a statistical measure used to assess the strength of association between two categorical variables. It is an extension of the chi-square test and is used to determine the degree of dependence between two variables.

Calculate Cramer's V using the formula:

$$V = \sqrt{\frac{\chi^2}{n \times \min(c-1, r-1)}}$$

The value of Cramer's V lies between 0 and 1. Here's how to interpret the strength of association:

- $V \geq 0.5$: Indicates a high association between the variables.
- $V < 0.5$: Implies a weak association between the variables.
- $V = 0$: Indicates no association between the variables.

2.2.3 Welch t-test

The Welch t-test is a statistical test used to compare the means of two independent groups when the assumptions of equal variances are violated. It is an alternative to the traditional Student's t-test, which assumes equal variances between the groups.

The hypothesis of this test are as follows,
H_0(Null Hypothesis): There is no significant difference between the means of the two groups.
H_1(Alternative Hypothesis): There is a significant difference between the means of the two groups.

Assumptions of Welch t-test are:

1. Independence: The observations in each group are independent of each other.
2. Normality: The data within each group are normally distributed.
3. Homogeneity of variance: The variances of the two groups are equal. This assumption is violated in the Welch t-test.

The formula for the Welch t-test is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}}$$

where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups.
- s_1^2 and s_2^2 are the sample variances of the two groups.
- n_1 and n_2 are the sample sizes of the two groups.

The t-value is calculated by taking the difference between the sample means and dividing it by the standard error, which accounts for the different variances of the two groups.

To determine the degrees of freedom for the Welch t-test, the Welch-Satterthwaite equation is used:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 /(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2 /(n_2 - 1)}$$

If the absolute value of the t-value exceeds the critical value, you can reject the null hypothesis and conclude that there is a significant difference between the means of the two groups.

2.2.4 Kruskal Wallis test

The Kruskal-Wallis test is a non-parametric statistical test used to compare the medians of three or more independent groups in a sample. The Kruskal-Wallis test is appropriate when the assumptions of parametric tests (e.g., ANOVA) are not met, such as when the data is not normally distributed or when the data is ordinal.

The hypothesis of this test are as follows,

H_0(Null hypothesis) : There is no significant difference between the medians of the groups.

H_1(Alternative hypothesis) : There is a significant difference between the medians of the groups.

The test statistic of the Kruskal-Wallis test is denoted as H and is calculated using the formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1),$$

where:

- N is the total number of observations (sum of all sample sizes),
- k is the number of groups being compared,
- R_i is the sum of ranks in group i,
- n_i is the sample size of group i.

If the test statistic exceeds the critical value, you reject the null hypothesis, indicating that there is a significant difference in the medians among the groups.

2.2.5 Dunn test

The Dunn test, also known as the Dunn's post hoc test, is a non-parametric statistical test used to perform multiple pairwise comparisons after rejecting the null hypothesis in an ANOVA(Analysis of variance) or Kruskal-Wallis test. It is employed when dealing with three or more groups to identify which specific groups have significantly different medians or central tendencies. The Dunn test accounts for the problem of multiple comparisons and helps to pinpoint where the significant differences lie among the groups.

2.2.6 Two sample Z-test for Proportions

In statistics, a two-sample z-test for proportions is a method used to determine whether two samples are drawn from the same population. This test is used when the population proportion is unknown and there is not enough information to use the chi-squared distribution. The test uses the standard normal distribution to calculate the test statistic. While performing the test, Z-statistic is computed from two independent samples.

The hypothesis of this test are as follows,

H_0(Null hypothesis) : The two proportions are equal.
H_1(Alternative hypothesis) : The two proportions are not equal.

In order to be able to use the two-sample z-test, the following conditions must be met:

- The two populations must be normal or approximately normal.
- The two samples must be randomly sampled from the two populations.

- The two proportions must be independent.

If any of the above conditions are not met, the two-sample z-test cannot be used and another test must be selected. The two-sample z-test is advantageous because it does not require any knowledge of the proportion standard deviation.

There are two steps in conducting a two-sample z-test for proportions.

- The first step is to calculate the standard error of the difference between the two population proportion.
- The second step is to calculate the z-test statistic. This is done by taking the difference between two population proportions and dividing it by the standard error of the difference.

Once the z-test statistic is calculated, the Z-table can be used to determine whether the two population proportion are different. If the z-statistic is greater than or equal to the critical value or level of significance, then it cam be concluded that there is enough evidence that there exists a difference between the two population proportions. And, the null hypothesis can thus be rejected.

2.2.7 Logistic Regression

Logistic regression is a statistical method used to model the relationship between a categorical dependent variable and one or more independent variables. It is commonly used for binary classification problems, where the dependent variable can take one of two possible outcomes.

The logistic regression model uses the logistic function (also called the sigmoid function) to model the relationship between the independent variables and the probability of the outcome. The logistic function is defined as:

$$P = \frac{1}{1 + e^{-z}}$$

where P is the probability of the dependent variable taking the value 1, e is the base of the natural logarithm (approximately 2.71828), and Y is the linear combination of the independent variables and their coefficients, given by:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and ci$x_1, x_2, \ldots, x_n$ are the values of the independent variables.

The logistic regression model estimates the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ using maximum likelihood estimation (MLE) and determines their statistical significance using hypothesis testing, typically with a significance level (alpha) of 0.05. The significance of each coefficient indicates whether the corresponding independent variable has a significant impact on the probability of the outcome.

2.2.8 Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

Types of decision trees are based on the type of target variable we have. It can be of two types:

- Categorical Variable Decision Tree Decision: Tree which has a categorical target variable then it is called a Categorical variable decision tree.
- Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

The common problem with Decision trees, especially having a table full of columns, they fit a lot. Sometimes it looks like the tree memorized the training data set. If there is no limit set on a decision tree, it will give you 100% accuracy on the training data set because in the worse case it will end up making 1 leaf for each observation. Thus this affects the accuracy when predicting samples that are not part of the training set. Pruning Decision Tree is used to remove overfitting.

Pruning Decision Trees:

The splitting process results in fully grown tree until the stopping criteria are reached. But, the fully grown tree is likely to overfit the data, leading to poor accuracy on unseen data. In pruning, you trim off the branches of the tree, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets:

training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set,V.

2.2.9 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and problem of overfitting.

Random forest algorithms have mainly three hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve the regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag(oob) sample. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task the most frequent categorical variable will yield the predicted class.

2.2.10 Extreme Gradient Boosting(XGBoost)

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm known for its performance and effectiveness in solving both regression and

classification problems. It is an ensemble learning method that combines the predictions of multiple weak prediction models, typically decision trees, to create a strong predictive model.

XGBoost builds an ensemble of decision trees, often referred to as "boosted trees." Each tree is constructed sequentially, and the subsequent trees are built to correct the mistakes of the previous ones. This iterative process helps in improving the overall predictive power.

The algorithm utilizes gradient boosting to train the individual decision trees. Gradient boosting involves minimizing a loss function by optimizing the parameters of each tree. The loss function represents the difference between the predicted and actual values. The algorithm calculates the gradient of the loss function with respect to the predicted values and uses this information to update the model parameters.

XGBoost applies regularization techniques to control the complexity of the model and avoid overfitting. It incorporates two types of regularization: "L1 regularization" (Lasso regularization) and "L2 regularization" (Ridge regularization). These regularization terms are added to the loss function and help in reducing the impact of individual trees on the final prediction.

XGBoost provides a measure of feature importance, indicating the relative importance of each feature in the model. It calculates this by evaluating the contribution of each feature across all the decision trees in the ensemble. Feature importance can help in feature selection and understanding the underlying patterns in the data.

2.2.11 Multinomial Regression

Multinomial regression, also known as multinomial logistic regression, is a statistical method used to model and analyze relationships between multiple categorical dependent variables and one or more independent variables. It is an extension of binary logistic regression, which deals with two categories.

In multinomial regression, the dependent variable can have three or more categories. The goal is to estimate the probability of each category occurring given the values of the independent variables. The categories are mutually exclusive, meaning an observation can only belong to one category.

The model assumes a linear relationship between the independent variables and the logits (log-odds) of the probabilities for each category. The logits are then transformed using the softmax function to obtain the probabilities.

$$\log\left(\frac{p_i}{p_k}\right) = \beta_{0i} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where $\log\left(\frac{p_i}{p_k}\right)$ represents the log-odds ratio between the probability of category i and a reference category k. β_{0i} represents the intercept term specific to category i. $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients corresponding to the independent variables $X_1, X_2, \ldots, X_p$. $X_1, X_2, \ldots, X_p$ are the independent variables. p_i represents the probability of category i.

To estimate the coefficients, maximum likelihood estimation is typically used. This involves finding the coefficients that maximize the likelihood of observing the given data, assuming the model is correct.

2.2.12 ARIMA Model

Time series analysis is a statistical technique that deals with time series data, or trend analysis, in this analysis we just have one variable that is time. It is used to predict future values based on previous observed values.

Component of time series are as follow.

- Trend: The trend shows the general tendency of the data to increase or decrease during a long period of time.
- Seasonal Variations: These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year.
- Cyclic Variations: The variations in a time series which operate themselves over a span of more than one year are the cyclic variations.
- Random or Irregular variations: These fluctuations are unforeseen, uncontrollable, unpredictable and are erratic.

The ARIMA model is a widely used time series forecasting technique that combines autoregression, differencing, and moving averages to model and predict future values in a time series. It is a powerful and flexible approach that can handle a wide range of time series data.

Components of ARIMA Model

1. Autoregression (AR): This component involves using past observations in the time series to predict future values. It assumes that the future values of the series are linearly dependent on its own past values. The order of autoregression, denoted as "p" represents the number of lagged observations used in the model.

2. Integration (I): This component deals with the differencing of the time series to make it stationary. Differencing is done to remove trends and seasonality from the data. The order of differencing, denoted as "d," indicates the number of times the series needs to be differenced to achieve stationarity.

3. Moving Average (MA): This component uses the past forecast errors in the time series to predict future values. It assumes that the future values are related to the errors of the model's previous forecasts. The order of the moving average, denoted as "q," represents the number of lagged forecast errors used in the model.

2.2.13 Augmented Dickey-Fuller test

The Augmented Dickey-Fuller (ADF) test is a statistical test commonly used to determine whether a time series data has non-stationarity. Non-stationarity occurs when a time series has a trend or seasonality that affects its statistical properties over time. The ADF test is widely used in time series analysis, particularly in the context of testing for stationarity.

H_0(Null Hypothesis): The time series is non-stationary.
H_1(Alternative Hypothesis): The time series is stationary.

2.2.14 Ljung Box test

The Ljung-Box test is a statistical test used to assess the presence of autocorrelation in a time series data. The test helps to determine if the data exhibits significant autocorrelation, which is the correlation between a time series and its lagged values at different time points.

H_0(Null Hypothesis): The null hypothesis of the Ljung-Box test states that there is no autocorrelation in the time series data.
H_1(Alternative Hypothesis): The alternative hypothesis of the Ljung-Box test, sometimes referred to as the working hypothesis, states that there is significant autocorrelation in the time series data.

3 Chapter 3

Results and Discussion: Univariate Analysis

3.1 Introduction

This chapter focuses on the exploratory data analysis of the variables considered in the dataset. In Section 3.2, the characteristics of the sample respondents are provided. Section 3.3 presents the conclusions drawn from the univariate analysis.

3.2 Characteristics of Sample Respondents

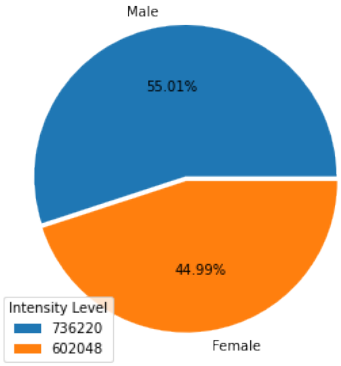
3.2.1 Gender

The following table shows the frequency distribution of Gender of respondents.

Table 1: Frequency distribution of COVID-19 cases by gender

Gender	Frequency	Percentage (%)
Male	7,36,220	55.01
Female	6,02,048	44.99
Total	13,38,268	100

Figure 1: Gender of COVID-19 patients



We observe that 55% of the Males have COVID-19 and about 45% of the Females have COVID-19.

3.2.2 Age

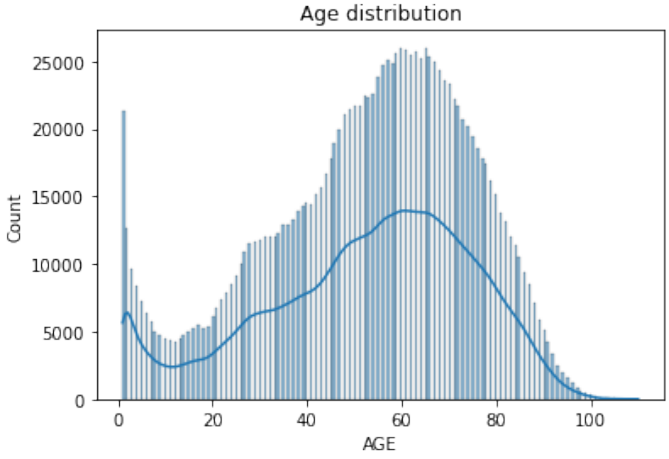
The following table shows the frequency distribution of the age (in years) of the respondents.

Table 2: Frequency distribution of age

Age group (yrs)	Frequency	Percentage (%)
0-14 (Children)	103358	7.7
15-24 (Youths)	63086	4.7
25-64 (Adults)	719884	53.8
65 & Above (Elders)	451940	33.8
Total	1338268	100

The following histogram shows the distribution of age of respondents.

Figure 2: Histogram of age of COVID-19 patients



We observe that 54% fall within the age range of 25-64 years old, 34% of the respondents are 65 years and above, 8% of the respondents fall into the age group of 0-14 years old and 4% of the respondents belong to the age group of 15-24 years old,.

3.2.3 Classification

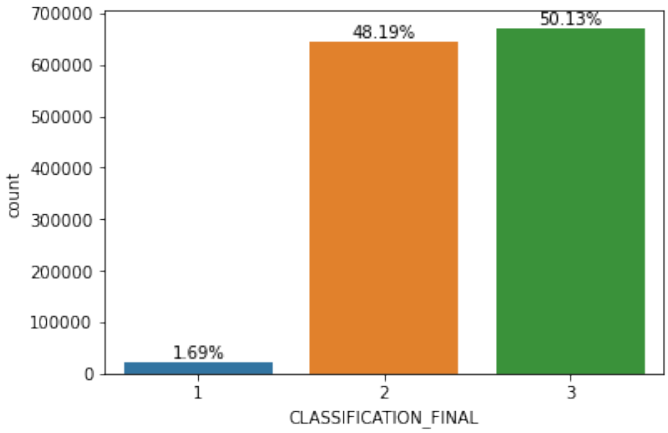
The following table shows the frequency distribution of Classification of respondents based on severity level of their symptoms.

Table 3: Frequency distribution of Classification of severity level of symptoms

Classification	Frequency	Percentage (%)
Mild (1)	22,591	1.69
Moderate (2)	6,44,863	48.19
Severe (3)	6,70,814	50.13

The following countplot shows the classification of respondents based on different degrees of symptoms i.e., Mild, Moderate and Severe.

Figure 3: Count plot of Severity level of symptoms



We observe that around 50% of the respondents are classified with Severe symptoms, 48% are classified with Moderate symptoms and only 2% are classified with Mild symptoms.

3.2.4 Treatment Type

The following table shows the frequency distribution of treatment type of respondents i.e., Intubed and ICU.

Table 4: Frequency distribution of Intubation and ICU

Treatment Type	Yes		No	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Intubed	1,17,539	8.78	12,20,729	91.22
ICU	80,880	6.04	12,57,388	93.96

We observe that 91% of the respondents didn't require intubation and 9% of the respondents required intubation. Also we observe that 94% didn't require treatment in the ICU treatment and 6% of the respondents required treatment in the ICU.

3.2.5 Consumption of Tobacco

The following table shows the frequency distribution of the respondents with smoking habit.

Table 5: Frequency distribution of Smoking habit

	Yes		No	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Tobacco	96,073	7.18	12,42,195	92.82

We observe that 93% of the respondents do not consume tobacco and 7% of the respondents consume tobacco.

3.2.6 Pneumonia, Diabetes, COPD, Asthma, Immunosuppression, Hypertension, Cardiovascular, Obesity, Renal chronic disease

The following table shows the frequency of the respondents having pneumonia, diabetes, copd, asthma, immunosuppression, hypertension, cardiovascular, obesity, renal chronic disease considered in this study.

Table 6: Frequency distribution of various health conditions

Health Conditions	Yes		No	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Pneumonia	6,52,316	48.74	6,85,958	51.26
Diabetes	3,75,549	28.06	9,62,719	71.94
COPD	54,150	4.05	12,84,118	95.95
Pregnancy	24,014	1.79	13,14,254	98.21
Asthma	30,763	2.30	13,07,505	97.70
Immunosurppression	41,289	3.09	12,96,979	96.91
Hypertension	4,40,950	32.95	8,97,318	67.05
Cardiovascular	70,017	5.23	12,68,251	94.77
Renal chronic	90,854	6.79	12,47,414	93.21
Obesity	2,04,906	15.31	11,33,362	84.69
Other disease	70,699	5.28	12,67,569	94.72

Here we observe that, 49% of the respondents are having pneumonia, 33% of the respondents are having hypertension, 28% of the respondents are having diabetes, 15% of the respondents are having obesity, 7% of the respondents are having renal chronic disease, 5% of the respondents are having cardiovascular disease, 5% of the respondents have other diseases, 4% of the respondents are having copd, 3% of the respondents are having immunosuppression, 2% of the respondents are pregnant and 2% of the respondents are having asthma considered in this study.

3.2.7 Death

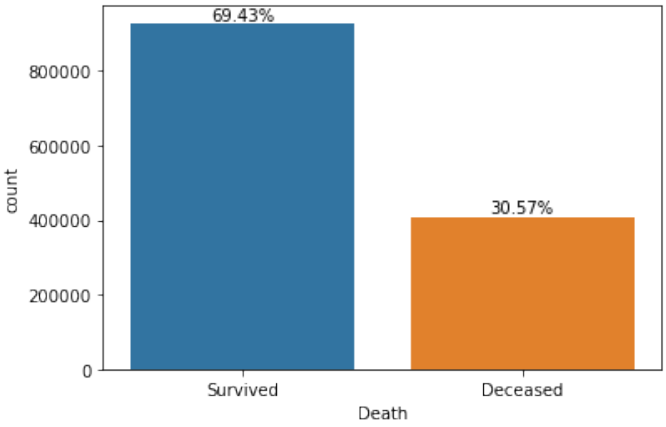
The following table shows the frequency distribution of death of the respondents.

Table 7: Frequency distribution of Death

	Frequency	Percentage (%)
Deceased	4,09,140	30.57
Survived	9,29,128	69.43
Total	13,38,268	100

The following countplot shows the death of the respondents.

Figure 4: Count plot of Severity level the COVID-19 patients



We observe that 31% died due to COVID-19 and 69% survived. This indicate that the survival is slight higher than death.

3.3 Conclusion

1. From the univariate analysis of the gender factor, we observe that the majority of males(55%) are infected by COVID-19.

2. From the univariate analysis of the age factor, we observe that adults(54%) and elders(34%) are at a higher risk of getting infected by the coronavirus disease.

3. Regarding the classification factor, the univariate analysis shows that the majority of respondents were classified under severe(50%) and moderate(48%) levels of symptoms.

4. Analyzing the treatment type factor, we observe that only 9% and 6% of respondents required treatment under ICU and intubation, respectively.

5. The univariate analysis of the consumption of tobacco factor reveals that only 7% of respondents consume tobacco.

6. Regarding health conditions, the univariate analysis indicates that pneumonia(49%), diabetes(28%), and hypertension(33%) are more prevalent among the respondents.

7. Finally, from the univariate analysis of the death factor, we observe that the 31% respondents died due to the novel disease.

4 Chapter 4

Results and Discussion: Bivariate Analysis

4.1 Introduction

In this chapter, bivariate analysis of the factors related to COVID-19 is presented. Section 4.2 provides details of the association between factors related to COVID-19 death. Section 4.3 elaborates on the different factors affecting the age of individuals with and without health conditions. Section 4.4 outlines the conclusions drawn from the bivariate analysis.

4.2 Bivariate Analysis of factors related to Death

4.2.1 Testing the association between Health conditions and Death using Chi-square test of independence

Analysis of the Association between Pneumonia and Death

The following table shows the distribution of survival and death among individuals with and without pneumonia.

Table 8: Contingency table for Pneumonia Outcome: Death vs. Survival

Pneumonia	Survived	Deceased
Yes	371481	280835
No	557647	128305

The Hypothesis are as follows:

H_0 : There is no association between pneumonia and death.
H_1 : There is an association between pneumonia and death.

The obtained values are as follows,

- Chi-square statistic = 93377.9965
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable pneumonia and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.2641

From Cramer V value, we conclude that the strength of association between pneumonia and death is 0.2641. Hence, we conclude that there is weak association between pneumonia and death.

Analysis of the Association between Diabetes and Death

The following table shows the distribution of survival and death among individuals with and without diabetes.

Table 9: Contingency table for Diabetes Outcome: Death vs. Survival

Diabetes	Survived	Deceased
Yes	151309	280835
No	704888	257831

The Hypothesis are as follows:

H_0 : There is no association between diabetes and death.
H_1 : There is an association between diabetes and death.

The obtained values are as follows,

- Chi-square statistic = 23225.5837
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable diabetes and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.3628

From Cramer V value, we conclude that the strength of association between diabetes and death is 0.3628. Hence, we conclude that there is weak association between diabetes and death.

Analysis of the Association between COPD and Death

The following table shows the distribution of survival and death among individuals with and without copd.

Table 10: Contingency table for COPD Outcome: Death vs. Survival

COPD	Survived	Deceased
Yes	33511	20639
No	895617	388501

The Hypothesis are as follows:

H_0 : There is no association between COPD and death.
H_1 : There is an association between COPD and death.

The obtained values are as follows,

- Chi-square statistic = 1512.0242
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable COPD and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0336

From Cramer V value, we conclude that the strength of association between COPD and death is 0.0336. Hence, we conclude that there is weak association between COPD and death.

Analysis of the Association between Pregnancy and Death

The following table shows the distribution of survival and death among individuals with and without pregnancy.

Table 11: Contingency table for Pregnancy Outcome: Death vs. Survival

Pregnancy	Survived	Deceased
Yes	23536	478
No	905592	408662

The Hypothesis are as follows:

H_0 : There is no association between Pregnancy and death.
H_1 : There is an association between Pregnancy and death.

The obtained values are as follows,

- Chi-square statistic = 9409.8839
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable pregnancy and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0838

From Cramer V value, we conclude that the strength of association between pregnancy and death is 0.0838. Hence, we conclude that there is weak association between pregnancy and death.

Analysis of the Association between Asthma and Death

The following table shows the distribution of survival and death among individuals with and without asthma.

Table 12: Contingency table for Asthma Outcome: Death vs. Survival

Asthma	Survived	Deceased
Yes	23748	7015
No	905380	402125

The Hypothesis are as follows:

H_0 : There is no association between Asthma and death.
H_1 : There is an association between Asthma and death.

The obtained values are as follows,

- Chi-square statistic = 894.9810
- Degrees of freedom = 1
- p-value = 1.2103e-196

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable asthma and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0258

From Cramer V value, we conclude that the strength of association between asthma and death is 0.0258. Hence, we conclude that there is weak association between asthma and death.

Analysis of the Association between Immunosuppression and Death

The following table shows the distribution of survival and death among individuals with and without immunosuppression.

Table 13: Contingency table for Immunosuppression Outcome: Death vs. Survival

Immunosuppression	Survived	Deceased
Yes	29409	11880
No	899719	397260

The Hypothesis are as follows:

H_0 : There is no association between Immunosuppression and death.
H_1 : There is an association between Immunosuppression and death.

The obtained values are as follows,

- Chi-square statistic = 64.9125

- Degrees of freedom = 1
- p-value = 7.8296e-16

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable immunosupression and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0069

From Cramer V value, we conclude that the strength of association between immunosuppression and death is 0.0069. Hence, we conclude that there is weak association between immunosuppression and death.

Analysis of the Association between Hypertension and Death

The following table shows the distribution of survival and death among individuals with and without hypertension.

Table 14: Contingency table for Hypertension Outcome: Death vs. Survival

Hypertension	Survived	Deceased
Yes	260443	180507
No	668685	228633

The Hypothesis are as follows:

H_0 : There is no association between Hypertension and death.
H_1 : There is an association between Hypertension and death.

The obtained values are as follows,

- Chi-square statistic = 33276.3003
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable hypertension and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.1577

From Cramer V value, we conclude that the strength of association between hypertension and death is 0.1577. Hence, we conclude that there is weak association between hypertension and death.

Analysis of the Association between Cardiovascular and Death

The following table shows the distribution of survival and death among individuals with and without cardiovascular condition.

Table 15: Contingency table for Cardiovascular Outcome: Death vs. Survival

Cardiovascular	Survived	Deceased
Yes	46231	23786
No	882897	385354

The Hypothesis are as follows:

H_0 : There is no association between Cardiovascular and death.
H_1 : There is an association between Cardiovascular and death.

The obtained values are as follows,

- Chi-square statistic = 402.0699
- Degrees of freedom = 1
- p-value = 1.9513e-89

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable cardiovascular and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0173

From Cramer V value, we conclude that the strength of association between cardiovascular and death is 0.0173. Hence, we conclude that there is weak association between cardiovascular and death.

Analysis of the Association between Renal chronic and Death

The following table shows the distribution of survival and death among individuals with and without renal chronic disease.

Table 16: Contingency table for Renal chronic Outcome: Death vs. Survival

Renal chronic	Survived	Deceased
Yes	56715	34139
No	872413	375001

The Hypothesis are as follows:

H_0 : There is no association between Renal chronic and death.
H_1 : There is an association between Renal chronic and death.

The obtained values are as follows,

- Chi-square statistic = 2251.9296
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable renal chronic and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0410

From Cramer V value, we conclude that the strength of association between renal chronic and death is 0.0410. Hence, we conclude that there is weak association between renal chronic and death.

Analysis of the Association between Obesity and Death

The following table shows the distribution of survival and death among individuals with and without obesity.

Table 17: Contingency table for Obesity Outcome: Death vs. Survival

Obesity	Survived	Deceased
Yes	127538	77368
No	801590	331772

The Hypothesis are as follows:

H_0 : There is no association between Obesity and death.
H_1 : There is an association between Obesity and death.

The obtained values are as follows,

- Chi-square statistic = 5884.9963
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable obesity and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0663

From Cramer V value, we conclude that the strength of association between obesity and death is 0.0663. Hence, we conclude that there is weak association between obesity and death.

Analysis of the Association between Other disease and Death

The following table shows the distribution of survival and death among individuals with and without other disease.

Table 18: Contingency table for Other disease Outcome: Death vs. Survival

Other disease	Survived	Deceased
Yes	46294	24405
No	882834	384735

The Hypothesis are as follows:

H_0 : There is no association between Other disease and death.
H_1 : There is an association between Other disease and death.

The obtained values are as follows,

- Chi-square statistic = 547.7129
- Degrees of freedom = 1
- p-value = 1.9367e-180

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable other disease and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0248

From Cramer V value, we conclude that the strength of association between other disease and death is 0.0248. Hence, we conclude that there is weak association between other disease and death.

4.2.2 Testing the association between Treatment type and Death using Chi-square test of independence

Analysis of the Association between Intubed and Death

The following table shows the distribution of survival and death among individuals with and without intubation.

Table 19: Contingency table for Intubed Outcome: Death vs. Survival

Intubed	Survived	Deceased
Yes	28725	88814
No	900403	320326

The Hypothesis are as follows:

H_0 : There is no association between intubed and death.
H_1 : There is an association between intubed and death.

The obtained values are as follows,

- Chi-square statistic = 122870.6196
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable intubed and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.3030

From Cramer V value, we conclude that the strength of association between intubed and death is 0.3030. Hence, we conclude that there is weak association between intubed and death.

Analysis of the Association between ICU and Death

The following table shows the distribution of survival and death among individuals with and without icu.

Table 20: Contingency table for ICU Outcome: Death vs. Survival

ICU	Survived	Deceased
Yes	41221	39659
No	887907	369481

The Hypothesis are as follows:

H_0 : There is no association between icu and death.
H_1 : There is an association between icu and death.

The obtained values are as follows,

- Chi-square statistic = 13822.3848
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable icu and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.1016

From Cramer V value, we conclude that the strength of association between icu and death is 0.1016. Hence, we conclude that there is weak association between icu and death.

4.2.3 Testing the association between consumption of Tobacco and Death using Chi-square test of independence

Analysis of the Association between consumption of Tobacco and Death

The following table shows the distribution of survival and death among individuals with and without consumption of tobacco.

Table 21: Contingency table for Tobacco Outcome: Death vs. Survival

Tobacco	Survived	Deceased
Yes	63694	32379
No	865434	376761

The Hypothesis are as follows:

H_0 : There is no association between tobacco and death.
H_1 : There is an association between tobacco and death.

The obtained values are as follows,

- Chi-square statistic = 477.6150
- Degrees of freedom = 1
- p-value = 0

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that the variable tobacco and death are dependent.

To measure the association we use Cramer's V and value is,

Cramer V = 0.0189

From Cramer V value, we conclude that the strength of association between tobacco and death is 0.0189. Hence, we conclude that there is weak association between tobacco and death.

4.2.4 Analysis of factors affecting the Death of COVID-19 patients using Proportionality test

Analysis of proportionality between Pneumonia and Death

The hypothesis of proportionality test are as follows:

H_0 = The proportion of patients having Pneumonia under death and survival are same.
H_1 = The proportion of patients having Pneumonia under death and survival are not same.

The following plot shows the proportion of patients having Pneumonia under Death and Survival.

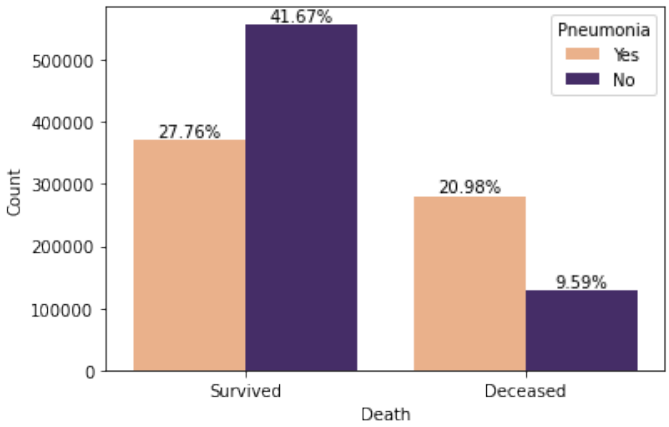


Figure 5: Count plot of Pneumonia under Deceased and Survived

The result obtained are as follows,

- Z-score = 305.580
- p-value = 0

From the above result, the obtained Z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportion are not same and there is a significant difference. Hence we can conclude that, proportion of COVID-19 patients having pneumonia under death and survived are not same.

From the graph we can observe that proportion of pneumonia patient under survival is greater than the proportion of pneumonia patients under deceased.

Analysis of proportionality between Diabetes and Death

The hypothesis of proportionality test are as follows:

H_0 = The proportion of patients having Diabetes under death and survival are same.
H_1 = The proportion of patients having Diabetes under death and survival are not same.

The following plot shows the proportion of patients having Diabetes under Death and Survival.

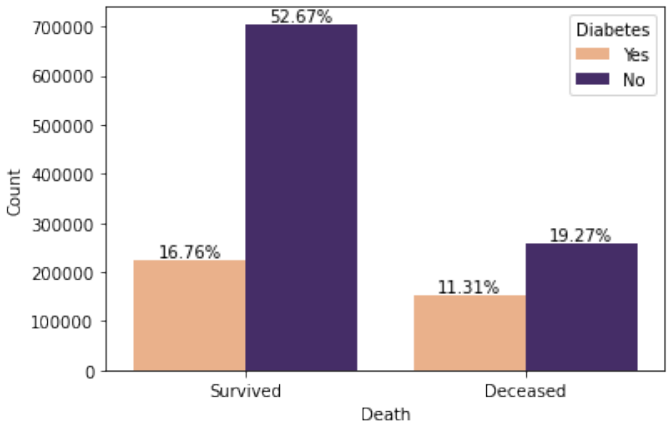


Figure 6: Count plot of Diabetes under Deceased and Survived

The result obtained are as follows,

- Z-score = 152.4015
- p-value = 0

From the above result, the obtained Z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportion are not same and there

is a significant difference. Hence we can conclude that, proportion of COVID-19 patients having diabetes under death and survived are not same.

From the graph we can observe that proportion of diabetes patient under survival is greater than the proportion of diabetes patients under deceased.

Analysis of proportionality between Hypertension and Death

The hypothesis of proportionality test are as follows:

H_0 = The proportion of patients having Hypertension under death and survival are same.
H_1 = The proportion of patients having Hypertension under death and survival are not same.

The following plot shows the proportion of patients having Hypertension under Death and Survival.

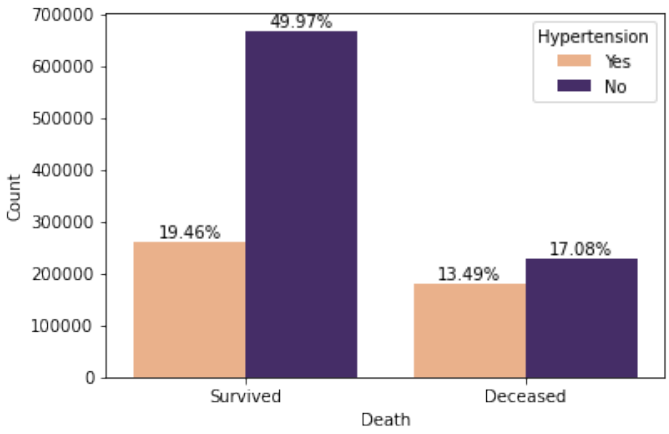


Figure 7: Count plot of Hypertension under Deceased and Survived

The result obtained are as follows,

- Z-score = 182.4199
- p-value = 0

From the above result, the obtained Z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportion are not same and there

is a significant difference. Hence we can conclude that, proportion of COVID-19 patients having hypertension under death and survived are not same.

From the graph we can observe that proportion of hypertension patient under survival is greater than the proportion of hypertension patients under deceased.

Analysis of proportionality between Intubed and Death

The hypothesis of two proportionality test are as follows:

H_0 = The proportion of patients Intubed under death and survival are same.
H_1 = The proportion of patients Intubed under death and survival are not same.

The following plot shows the proportion of patients Intubed under Death and Survival.

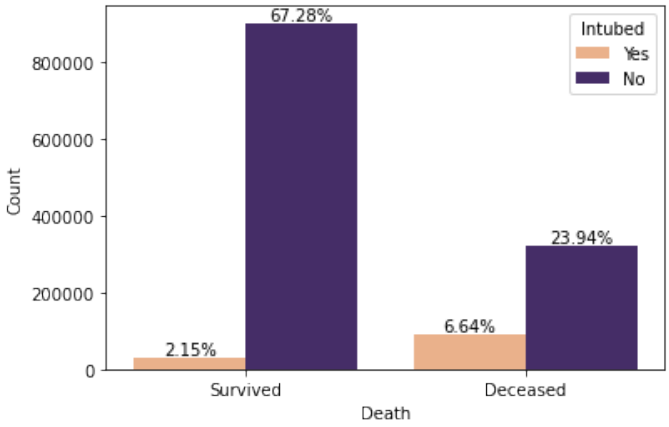


Figure 8: Count plot of Intubed in Deceased and Survived

The result obtained are as follows,

- Z-score = 350.5324
- p-value = 0

From the above result, the obtained Z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportion are not same and there is a significant difference. Hence we can conclude that, proportion of COVID-19 patients intubed under death and survived are not same.

From the graph we can observe that proportion of intubed patient under deceased is greater than the proportion of intubed patients under survival.

Analysis of proportionality between ICU and Death

The hypothesis of proportionality test are as follows:

H_0 = The proportion of patients having ICU under death and survival are same
H_1 = The proportion of patients having ICU under death and survival are not same

The following plot shows the proportion of patients having ICU under Death and Survival.

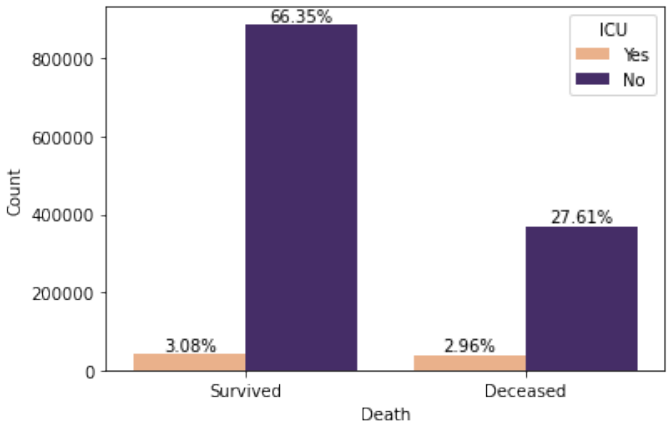


Figure 9: Count plot of ICU in Deceased and Survived

The result obtained are as follows,

- Z-score = 117.5726
- p-value = 0

From the above result, the obtained Z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportion are not same and there is a significant difference. Hence we can conclude that, proportion of COVID-19 patients icu under death and survived are not same.

From the graph we can observe that proportion of icu patient under survival is greater than the proportion of icu patients under deceased.

4.3 Bivariate Analysis of factors related to Age

4.3.1 Analysis of factors affecting the Age of individuals with and without Health Conditions using Welch t-test

Since the sample size is large assuming that data follows normality. To check homogeneity of variance assumption, Levene's test is applied.

The hypothesis are as follows:

H_0 = There is homogeneity of variances.
H_1 = There is no homogeneity of variances.

The result obtained are as follows:

Table 22: Table showing results of Levene's homogeneity of variance test

Factors	Test statistic	p-value	Decision
Pneumonia	40090.72	0	Reject H_0
Pregnancy	18342.76	0	Reject H_0
Diabetes	126523.94	0	Reject H_0
Hypertension	107546.14	0	Reject H_0
Cardiovascular	4839.19	0	Reject H_0
Renal chronic	8202.78	0	Reject H_0
COPD	15346.92	0	Reject H_0
Asthma	829.02	3.008e-180	Reject H_0
Immunosuppression	1966.32	0	Reject H_0
Obesity	31921.13	0	Reject H_0
Other disease	1019.03	1.590e-223	Reject H_0
Intubed	9415.86	0	Reject H_0
ICU	1689.30	0	Reject H_0
Tobacco	6210.98	0	Reject H_0

The assumption of homogeneity of variance does not hold. The p-value of pneumonia, pregnancy, diabetes, hypertension, cardiovascular, renal chronic, copd, asthma, immunosuppression, obesity, other disease, intubed, icu and tobacco are less than 0.05. Hence we reject the null hypothesis. Therefore we can conclude that the data does not follow homogeneity condition.

Hence we use alternative parametric test, Welch t-test to carry out the analysis.

The hypothesis of Welch t-test are as follows:

H_0 : The mean age of patients with health condition is same as that of mean age of patients without health condition.
H_1 : The mean age of patients with health condition is greater than that of mean age of patients without health condition.

The following table shows the result of Welch t-test

Table 23: Table showing results of Welch t-test

Factors	Test statistic	p-value	Decision
Pneumonia	221.79	0	Reject H_0
Pregnancy	-535.08	1	Do not reject H_0
Diabetes	423.70	0	Reject H_0
Hypertension	547.59	0	Reject H_0
Cardiovascular	167.24	0	Reject H_0
Renal chronic	89.26	0	Reject H_0
COPD	318.56	0	Reject H_0
Asthma	-44.53	1	Do not reject H_0
Immunosuppression	-52.93	1	Do not reject H_0
Obesity	89.87	0	Reject H_0
Other disease	2.31	0.0105	Reject H_0
Intubed	82.89	0	Reject H_0
ICU	27.26	6.9403e-163	Reject H_0
Tobacco	110.32	0	Reject H_0

From the above results of Welch t-test, we observe that the obtained p-value is greater than the level of significance ($\alpha = 0.05$) for the factors Pregnancy, Asthma and Immunosuppression. Hence we do not reject H_0. Hence we conclude that mean age of patients with Pregnancy, Asthma and Immunosuppression are same as that of mean age of patients without Pregnancy, Asthma and Immunosuppression.

Also we observe that the obtained p-value is less than the level of significance ($\alpha = 0.05$) for the factors Pneumonia, Diabetes, Hypertension, Cardiovascular, Renal chronic, COPD, Obesity, Other disease, Intubed, ICU and Tobacco. Hence we reject H_0. Hence we conclude that the mean age of patients with Pneumonia, Diabetes, Hypertension, Cardiovascular, Renal chronic, COPD, Obesity, Other disease,

Intubed, ICU and Tobacco are greater than mean age of patients without Pneumonia, Diabetes, Hypertension, Cardiovascular, Renal chronic, COPD, Obesity, Other disease, Intubed, ICU and Tobacco.

4.3.2 Analysis of factors affecting the Age of individuals with different degrees of symptoms using Kruskal Wallis test

Since the sample size is large we assuming that data follows normality. To check homogeneity of variance assumption, Levene's test is applied.

The hypothesis are as follows:

H_0 = There is homogeneity of variances.
H_1 = There is no homogeneity of variances.

The result obtained are as follows:

- F statistic = 42282.3889
- p-value = 0

We observe that the p-value of the test if less than the level of significance ($\alpha = 0.05$). Thus, we reject the null hypothesis and conclude that there is no homogeneity of variances.

Hence we use alternative non parametric test, Kruskal Wallis to carry out the analysis. The hypothesis are as follows:

H_0 : There is no significant difference in the median age among different severity levels.
H_1 : There is a significant difference in the median age among different severity levels.

The results obtained are as follows:

- Kruskal-Wallis test statistic = 39595.0629
- p-value = 0

From the above results of Kruskal Wallis test, we observe that the obtained p-value is less than the level of significance ($\alpha = 0.05$). Hence we reject H_0. Hence we

conclude there is a significant difference in the median age among different severity levels.

Further we go for post hoc Dunn test to find significant difference among the severity levels.

The following table gives the significant difference between the severity levels.

Table 24: Table showing the results of Dunn test

Severity levels	p-value
Level 1 - Level 2	0
Level 1 - Level 3	7.966e-16
Level 2 - Level 3	0

Using post hoc Dunn test, we observe that there is significant difference between median age among severity level 1 and 2, severity level 1 and 3 and severity level 2 and 3.

4.4 Conclusion

1. From bivariate analysis using chi-square test of independence, We find a weak association between pneumonia, diabetes, COPD, pregnancy, asthma, immunosuppression, hypertension, cardiovascular disease, renal chronic disease, obesity, other disease, intubed, icu and the consumption of tobacco with death. The strength of the association between each factor and death is relatively low, below 0.5.

2. From bivariate analysis using the proportionality test, we observe that the proportion of patients with pneumonia, diabetes, hypertension and icu is higher among survivors, while the proportion of intubated patients is higher among the deceased.

3. From bivariate analysis using the Welch t-test indicates that the average age of patients with pneumonia, diabetes, hypertension, cardiovascular disease, renal chronic disease, COPD, obesity, other diseases, intubation, ICU and tobacco consumption is higher than the average age of patients without these conditions and treatments.

4. From bivariate analysis using the Kruskal-Wallis test, we observe that there is a significant difference in symptom levels, with mild symptoms differing from moderate and severe symptom levels, and moderate symptom levels differing from severe symptom levels.

5 Chapter 5

Results and Discussion: Multivariate Analysis

5.1 Introduction

In this chapter, the multivariate analysis of factors affecting death and the classification of severity levels is presented. Section 5.2 and Section 5.3 attempt to study various factors that influence death and the severity classification of COVID-19 patients. Section 5.4 provides the forecast model for death. Finally, Section 5.5 outlines the conclusions drawn from the multivariate analysis.

5.2 Multivariate Model for Classifying Severity Levels of Symptoms

5.2.1 Multinomial Regression

Let severity level of symptoms of respondents be considered as the response variable. Here, the response variable has three categories : Mild, Moderate and Severe.

The frequency distribution of the response variable is as follows.

Table 25: Frequency distribution of Classification of severity level of symptoms

Classification	Percentage frequency
Mild (1)	1.69
Moderate (2)	48.19
Severe (3)	50.13

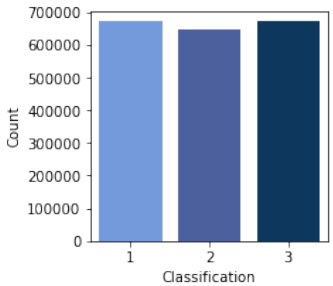
The independent variables taken into consideration are as follows.

Table 26: Table showing the covariates of multinomial regression

Covariates	Values	Condition	Covariates	Values	Condition
Pneumonia	1	Yes	Insmsurp	1	Yes
	0	No		0	No
Diabetes	1	Yes	Obesity	1	Yes
	0	No		0	No
Hypertension	1	Yes	Other disease	1	Yes
	0	No		0	No
Cardiovascular	1	Yes	Intubed	1	Yes
	0	No		0	No
Renal chronic	1	Yes	ICU	1	Yes
	0	No		0	No
COPD	1	Yes	Age	Continuous	Age of respondents
	0	No			
Asthma	1	Yes			
	0	No			

We observe that the data is highly imbalanced. So, before building the model we must balance the data. Here, upsampling method is used to balance the data. The train-test split is done with size = 80%. Then the train data is up sampled in order to balance the response variable. After balancing, we observe the following classification in the train set.

Figure 10: Figure showing balanced response variable of Multinomial regression



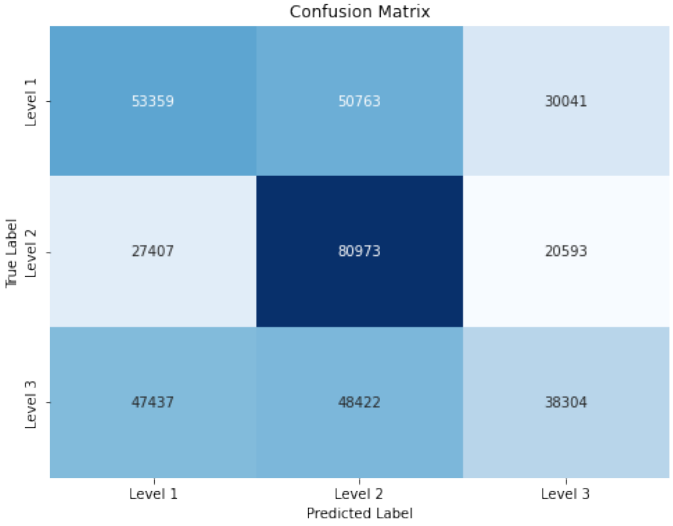
The hyper parameter tuning is done using grid search and the following best pa-

rametes are obtained.

- Best estimator C = 10
- Solver = lbfgs
- Penalty = l2

Hence, the multinomial regression model with optimal hyper parameter is fit to the training set. This model is then used to predict the response values for the test set. The following confusion matrix is obtained. Here, '1′ indicates the severity level of symptoms of COVID-19 patient is 'Mild', '2′ indicates the severity level of symptoms of COVID-19 patient is 'Moderate' and '3′ indicates the severity level of symptoms of COVID-19 patient is 'Severe'.

Figure 11: Figure showing the confusion matrix of multinomial regression



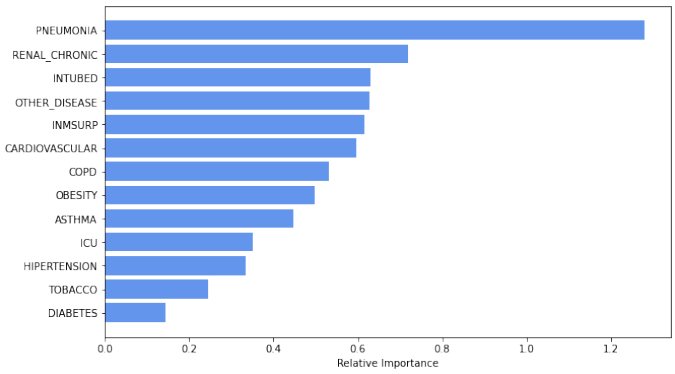
The accuracy of the fitted model for the test set is found to be 43%. This implies that the model has correctly predidcted 43% of the predictions out of all the predictions made. On the other hand, the accuracy of the train set is obtained to be 43%. Thus, the model is not capturing all the variation in the train data and the model accuracy on test is also less. Hence, the model is underfit.

The model performance can also be evaluated using the metrics like accuracy, precision, recall and F1 score.

- Precision = 45% : Out of all the patients predicted to be classified under moderate symptoms 45% did actually were classified under patients with moderate symptoms due to COVID-19.
- Recall = 63% : Out of all the patients who died, 63% were correctly predicted to be classified under Moderate symptoms.
- F1 Score = 52% : This indicates that the model has a poor performance.

The feature importance is calculated using permutation importance to study which factors influence the death. The following plot shows the permutation feature importance.

Figure 12: Figure showing permutation feature importance of Multinomial regression



From the above plot, we observe that the variables 'Pneumonia' have major contribution in determining the classification of severity level of symptoms. The variables 'Renal chronic', 'Intubed', 'Other disease', 'Inmsurp', 'Cardiovascular', 'COPD', 'Obesity', 'Asthma', 'ICU', 'Hypertension', 'Tobacco' and 'Diabetes' also have moderate influence on the classification of severity level of symptoms.

5.3 Analysis of factors influencing the death of COVID-19 patients

5.3.1 Logistic Regression

Let the death of COVID-19 patients be considered as the response variable. Here the response variable has two categories: Deceased and Survived.

The 13 independent variables taken into consideration are as follows.

Table 27: Table showing the covariates of Logistic regression

Covariates	Values	Condition	Covariates	Values	Condition
Pneumonia	1	Yes	Insmsurp	1	Yes
	0	No		0	No
Diabetes	1	Yes	Obesity	1	Yes
	0	No		0	No
Hypertension	1	Yes	Other disease	1	Yes
	0	No		0	No
Cardiovascular	1	Yes	Intubed	1	Yes
	0	No		0	No
Renal chronic	1	Yes	ICU	1	Yes
	0	No		0	No
COPD	1	Yes	Age	Continuous	Age of respondents
	0	No			
Asthma	1	Yes			
	0	No			

The frequency distribution of response variable is as follows. The following table shows the frequency distribution of death of the respondents.

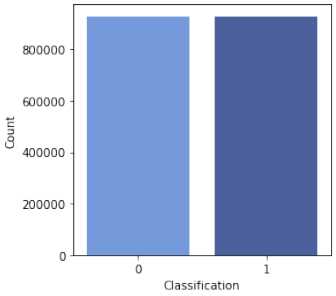
Table 28: Frequency distribution of Death

Status	Percentage frequency
Deceased	30.57
Survived	69.43

We observe that the data is imbalanced. So, before building the logit model, we

must balance the data. Here, up sampling method is used to balance the data. The train-test is done with train size = 80%. Then, the train data is up sampled in order to balance the response variable 'Death'. After balancing, we observe the following classification in the train set.

Figure 13: Figure showing the balanced response variable of logistic regression

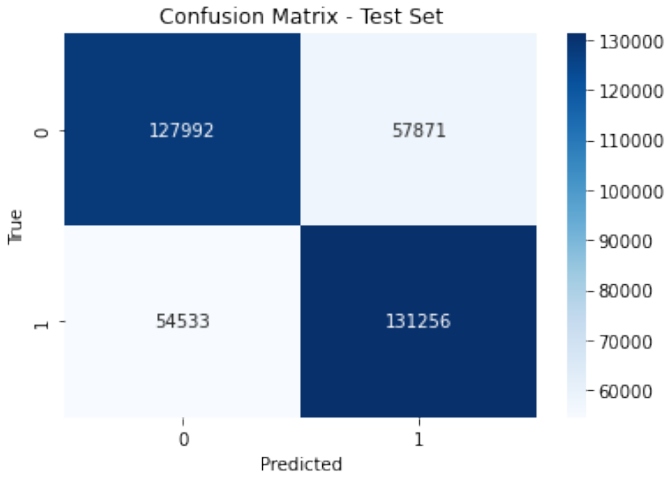


The hyper parameter tuning is done using grid search and the following best parameters are obtained.

- Best estimator C = 0.1
- Solver = lbfgs
- Penalty = l2

Hence, the logistic regression model with optimal hyper parameter is fit to the training set. This model is then used to predict the response values for the test set. The following confusion matrix is obtained. Here, '0' indicates the status of COVID-19 patient is 'Survived' and '1' indicates 'Deceased'.

Figure 14: Figure showing the confusion matrix of logistic regression



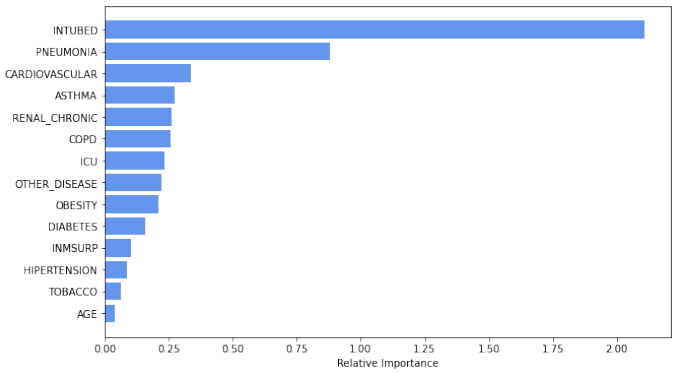
The accuracy of the fitted model for the test set is found to be 70%. This implies that the model has correctly predidcted 70% of the predictions out of all the predictions made. On the other hand, the accuracy of the train set is obtained to be 70%. This implies that the model correctly predicted 70% of the train set samples. Hence, the model is goodfit.

The model performance can also be evaluated using the metrics like accuracy, precision, recall and F1 score.

- Precision = 69% : Out of all the patients predicted to be deceased 69% did actually have died due to COVID-19.
- Recall = 71% : Out of all the patients who died, 71% were correctly predicted to have died due to COVID-19.
- F1 Score = 70% : This indicates that the model performs moderately.

The feature importance is calculated using permutation importance to study which factors influence the death. The following plot shows the permutation feature importance.

Figure 15: Figure showing permutation feature importance of Logistic regression



From the above plot, we observe that the variables 'Intubed' and 'Pneumonia' have major contribution in determining the death. The variables 'Cardiovascular', 'Asthma', 'Renal chronic', 'COPD', 'ICU', 'Other disease' and 'Obesity' also have an influence on the death. The other variable 'Diabetes', 'Inmsurp', 'Hypertesnion', 'Tobacco' and 'Age' have low influence on the death.

5.3.2 Decision Tree

A Decision Tree model is built to classify and predict the death based on following features.

Table 29: Table showing the covariates of Decision Tree

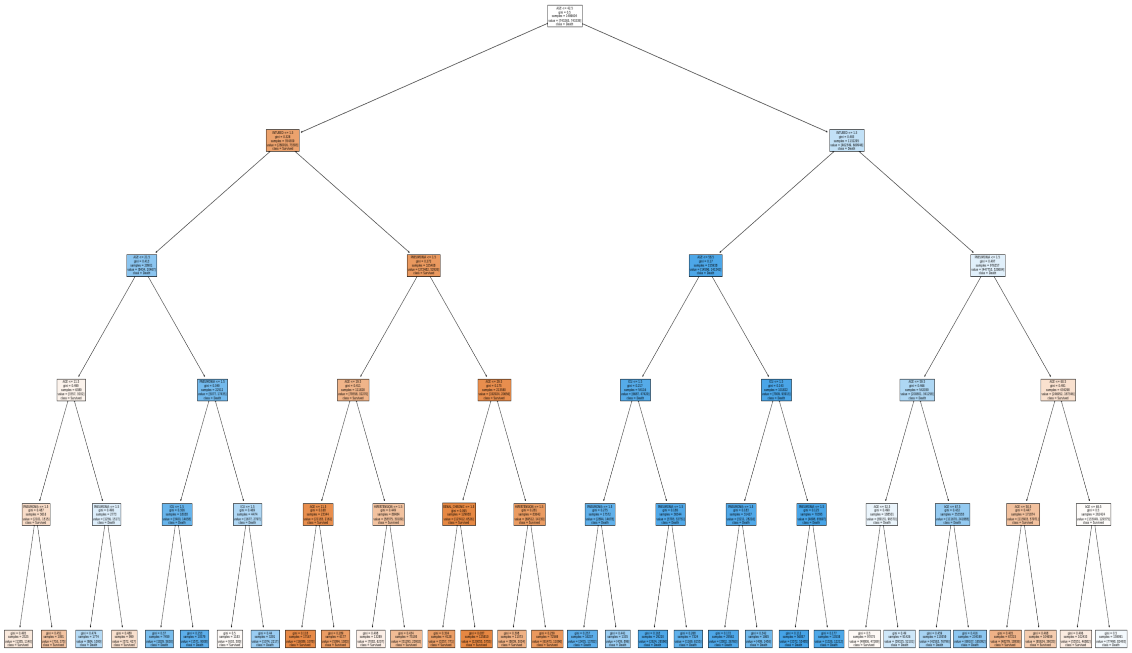
Covariates	Values	Condition	Covariates	Values	Condition
Pneumonia	1	Yes	Insmsurp	1	Yes
	0	No		0	No
Diabetes	1	Yes	Obesity	1	Yes
	0	No		0	No
Hypertension	1	Yes	Other disease	1	Yes
	0	No		0	No
Cardiovascular	1	Yes	Intubed	1	Yes
	0	No		0	No
Renal chronic	1	Yes	ICU	1	Yes
	0	No		0	No
COPD	1	Yes	Age	Continuous	Age of respondents
	0	No			
Asthma	1	Yes			
	0	No			

The train-test split is done with train size = 80%. The hyper parameter tuning is done using grid search and the best parameters obtained are as follows.

- criterion = gini
- min_samples_leaf = 1
- min_samples_split = 2

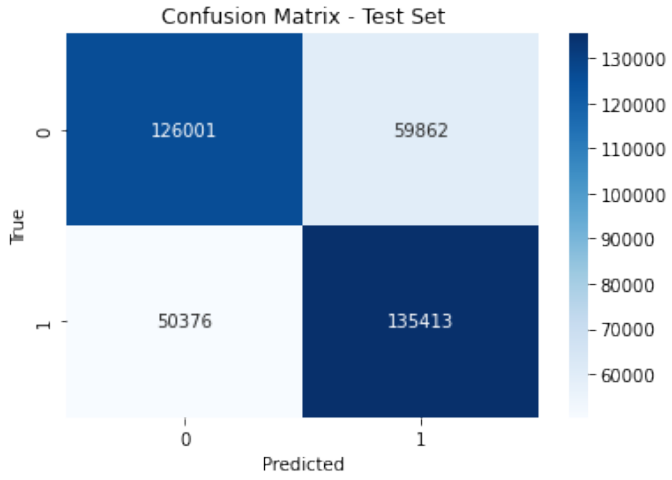
The following plot gives a visual representation of the decision tree classifier.

Figure 16: Figure showing the visual representation of Decision Tree



Hence, the decision tree classifier with optimal hyper parameter is fit to the training set. This model is then used to predict the response values for the test set. The following confusion matrix is obtained.

Figure 17: Figure showing the confusion matrix of decision tree classifier



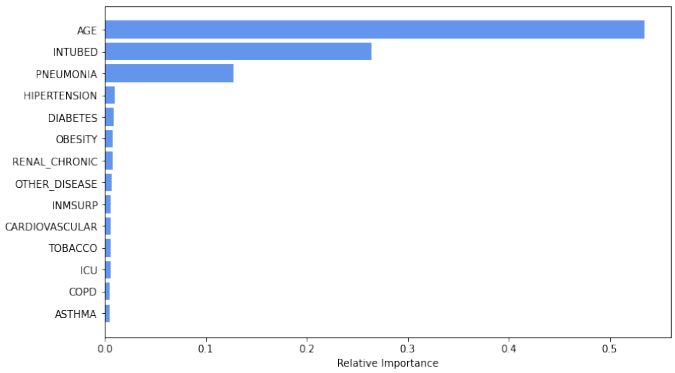
The accuracy of the fitted model for the test set is found to be 70%. This implies that the model has correctly predidcted 70% of the predictions out of all the predictions made. On the other hand, the accuracy of the train set is obtained to be 72%. This implies that the model correctly predicted 72% of the train set samples. Hence, the model is goodfit.

The model performance can also be evaluated using the metrics like accuracy, precision, recall and F1 score.

- Precision = 69% : Out of all the patients predicted to be deceased 69% did actually have died due to COVID-19.
- Recall = 73% : Out of all the patients who died, 73% were correctly predicted to have died due to COVID-19.
- F1 Score = 71% : This indicates that the model performs moderately.

The feature importance is calculated using permutation importance to study which factors influence the death. The following plot shows the permutation feature importance.

Figure 18: Figure showing permutation feature importance of Decision tree classifier



From the above plot, we observe that the variables 'Age' and 'Intubed' have major contribution in determining the death. The variables 'Pneumonia' also have an influence on the death. The other variable have low influence on the death.

5.3.3 Random Forest

A Random Forest model is built to classify and predict the death based on following features.

The frequency distribution of response variable is as follows.

Table 30: Table showing the covariates of Random Forest

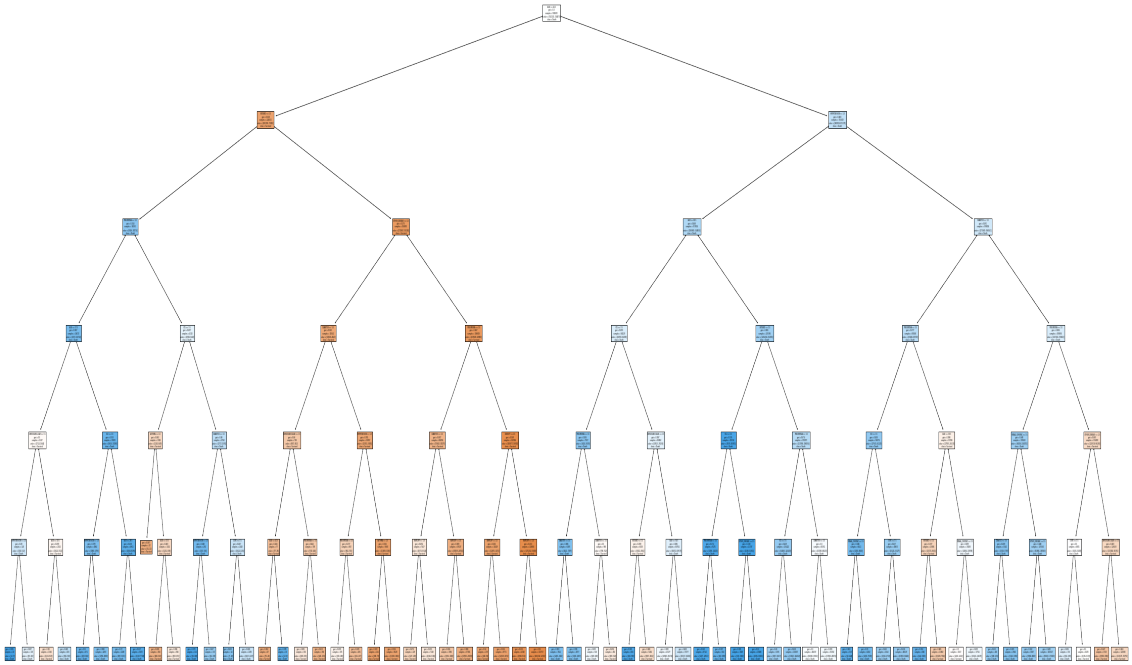
Covariates	Values	Condition	Covariates	Values	Condition
Pneumonia	1	Yes	Insmsurp	1	Yes
	0	No		0	No
Diabetes	1	Yes	Obesity	1	Yes
	0	No		0	No
Hypertension	1	Yes	Other disease	1	Yes
	0	No		0	No
Cardiovascular	1	Yes	Intubed	1	Yes
	0	No		0	No
Renal chronic	1	Yes	ICU	1	Yes
	0	No		0	No
COPD	1	Yes	Age	Continuous	Age of respondents
	0	No			
Asthma	1	Yes			
	0	No			

The train-test split is done with train size = 80%. The hyper parameter tuning is done using grid search and the best parameters obtained are as follows.

- min_samples_leaf = 1
- min_sample_split = 2
- n_estimators = 3

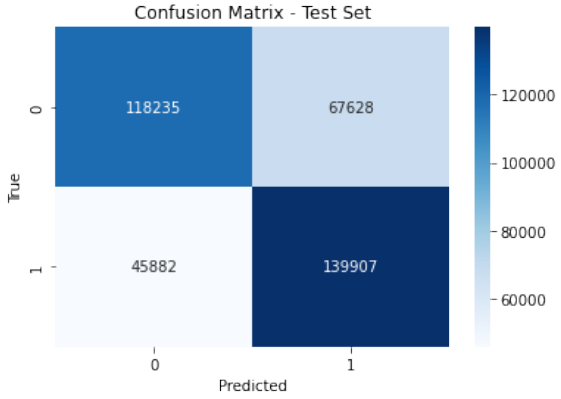
The following plot gives a visual representation of the random forest classifier.

Figure 19: Figure showing the visual representation of random forest



Hence, the random forest classifier with optimal hyper parameter is fit to the training set. This model is then used to predict the response values for the test set. The following confusion matrix is obtained.

Figure 20: Figure showing the confusion matrix of random forest classifier



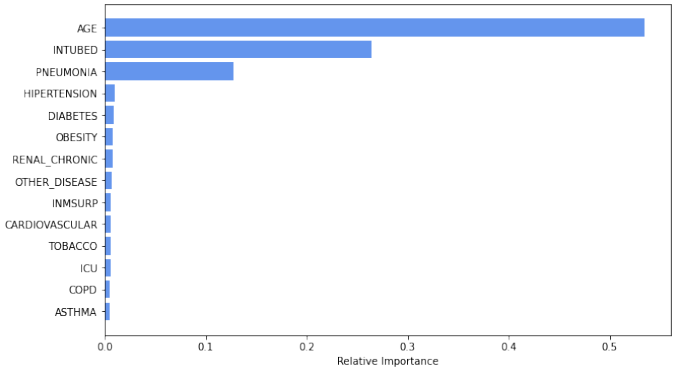
The accuracy of the fitted model for the test set is found to be 70%. This implies that the model has correctly predidcted 70% of the predictions out of all the predictions made. On the other hand, the accuracy of the train set is obtained to be 71%. This implies that the model correctly predicted 71% of the train set samples. Hence, the model is goodfit.

The model performance can also be evaluated using the metrics like accuracy, precision, recall and F1 score.

- Precision = 69% : Out of all the patients predicted to be deceased 69% did actually have died due to COVID-19.
- Recall = 71% : Out of all the patients who died, 71% were correctly predicted to have died due to COVID-19.
- F1 Score = 70% : This indicates that the model performs moderately.

The feature importance is calculated using permutation importance to study which factors influence the death. The following plot shows the permutation feature importance.

Figure 21: Figure showing permutation feature importance of Random forest classifier



From the above plot, we observe that the variables 'Age' and 'Intubed' have major contribution in determining the death. The variable 'Pneumonia' also have an influence on the death. The other variable have low influence on the death.

5.3.4 XGBoost

A Extreme Gradient Boost model is built to classify and predict the death based on following features.
The frequency distribution of response variable is as follows.

Table 31: Table showing the covariates of XGBoost

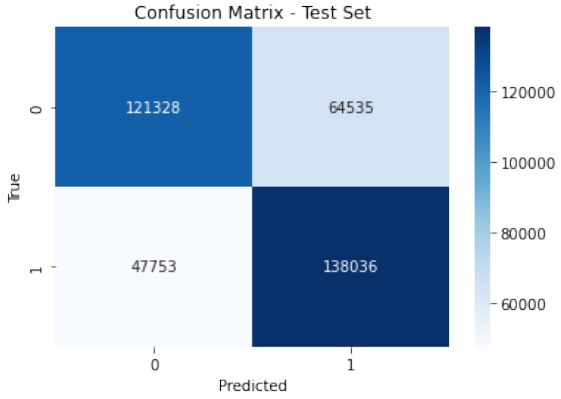
Covariates	Values	Condition	Covariates	Values	Condition
Pneumonia	1	Yes	Insmsurp	1	Yes
	0	No		0	No
Diabetes	1	Yes	Obesity	1	Yes
	0	No		0	No
Hypertension	1	Yes	Other disease	1	Yes
	0	No		0	No
Cardiovascular	1	Yes	Intubed	1	Yes
	0	No		0	No
Renal chronic	1	Yes	ICU	1	Yes
	0	No		0	No
COPD	1	Yes	Age	Continuous	Age of respondents
	0	No			
Asthma	1	Yes			
	0	No			

The train-test split is done with train size = 80%. The hyper parameter tuning is done using grid search and the best parameters obtained are as follows.

- max_depth = 0.2
- learning_rate = 0.1
- n_estimators = 7
- gamma = 3

Hence, the XGBoost classifier with optimal hyper parameter is fit to the training set. This model is then used to predict the response values for the test set. The following confusion matrix is obtained.

Figure 22: Figure showing the confusion matrix of XGBoost classifier



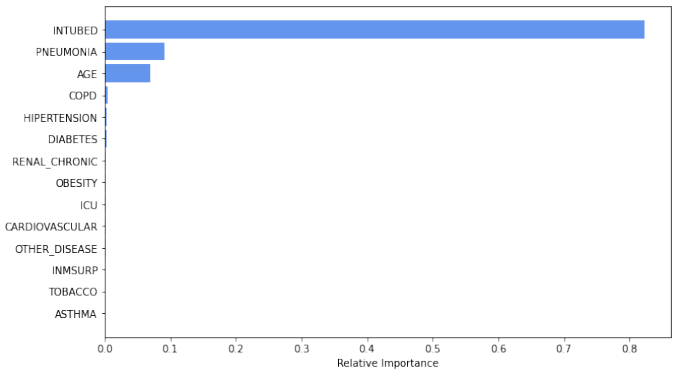
The accuracy of the fitted model for the test set is found to be 70%. This implies that the model has correctly predicted 70% of the predictions out of all the predictions made. On the other hand, the accuracy of the train set is obtained to be 70%. This implies that the model correctly predicted 70% of the train set samples. Hence, the model is goodfit.

The model performance can also be evaluated using the metrics like accuracy, precision, recall and F1 score.

- Precision = 69% : Out of all the patients predicted to be deceased 69% did actually have died due to COVID-19.
- Recall = 71% : Out of all the patients who died, 71% were correctly predicted to have died due to COVID-19.
- F1 Score = 70% : This indicates that the model performs moderately.

The feature importance is calculated using permutation importance to study which factors influence the death. The following plot shows the permutation feature importance.

Figure 23: Figure showing permutation feature importance of XGBoost classifier



From the above plot, we observe that the variables 'Intubed' have major contribution in determining the death. The variables 'Pneumonia' and 'Age' have low influence on the death. The other variable have no influence on the death.

5.3.5 Model Evaluation

The following table shows the different classification models considered and the corresponding outputs.

Table 32: Table showing outputs of classification models

Model	Best Parametes	Precision	Recall	F1 Score	Train Accuracy	Test Accuracy
Logistic Regression	'C' : 0.1 'solver' : 'lbfgs' 'penalty' : 'l2'	69%	71%	70%	69.65%	69.76%
Decision Tree	'criterion' : 'gini' 'min_samples_leaf ': 1 'min_samples_split' : 2	69%	73%	71%	71.77%	70.34%
Random Forest	'min_samples_leaf' : 1 'min_sample_split' : 2 'n_estimators' : 3	69%	71%	70%	71.52%	70.21%
XGBoost	'max_depth' : 0.2 'learning_rate' : 0.1 'n_estimators' : 7 'gamma' : 3	69%	71%	70%	69.68%	69.79%

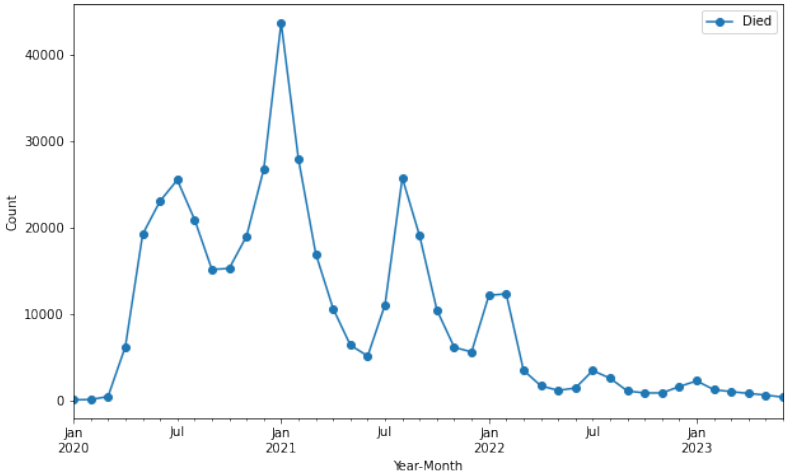
From the above table, it can be observed that the Logistic Regression and XGBoost models have lower performance on the test sets compared to the Decision Tree and Random Forest models. The Decision Tree classifier exhibits better recall and F1 score however, it shows a notable difference in train and test accuracy compared to the other models. Based on the evaluation metrics, we can conclude that among the four models considered, the Decision Tree classifier performs the best in predicting death.

5.4 Forecasting Death

5.4.1 ARIMA Model

The following time plot shows the death due to COVID-19 from the year 2020 till June, 2023.

Figure 24: Figure showing COVID-19 death over time



Here we can observe a trend where deaths initially increased, reaching their peak in the year 2021, and then gradually declined over time.

Using Augmented Dickey-Fuller test to check the stationarity of the data.

The result obtained are as follows:

- Dickey-Fuller statistic = -0.7756
- p-value = 0.8261

Since p-value is greater than 0.05, we say that the data is non stationary.

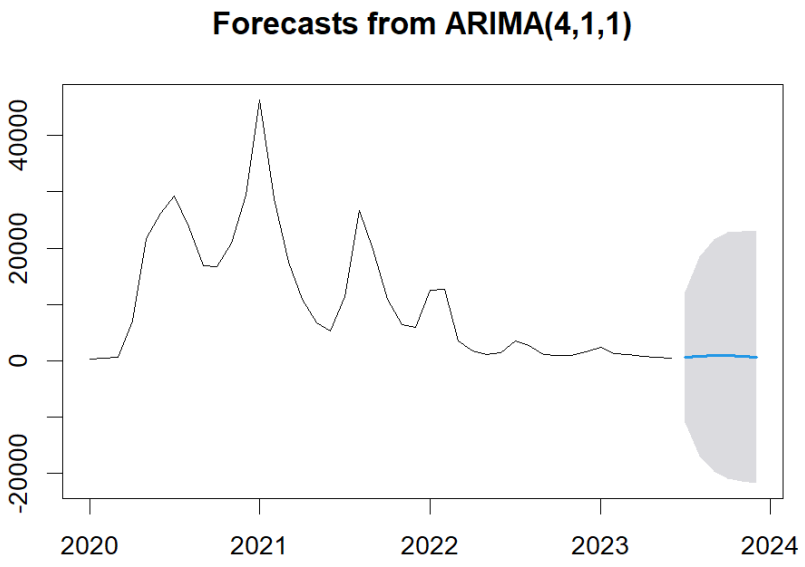
The following table shows result of various ARIMA models.

Table 33: Comparing ARIMA model results

S. No.	Model	RMSE	Ljung box
1	ARIMA(4,1,0)	5487.368	0.7958
2	ARIMA(2,1,1)	5686.754	0.8427
3	ARIMA(3,1,0)	5808.99	0.6775
4	ARIMA(2,1,0)	5997.877	0.6729
5	ARIMA(4,1,1)	5477.292	0.8788

From the diagnostic checking, we say that residuals are IID in all the models. But here we can observe that RMSE is less for the model ARIMA(4,1,1) as compared to other models. So we say that ARIMA(4,1,1) is the best model for forecasting death. The following plot shows the forecast of death from the month of July till December 2023.

Figure 25: Figure showing COVID-19 death over time



5.5 Conclusion

1. The multinomial regression model built to classify the patients based on severity level of symptoms had accuracy 43%. The best estimator obtained is C = 1. Though the F1 score is found to be 52%, the model proved to be an underfit with very low accuracy. The feature importance plot shows that the variable 'Pneumonia' have significant influence on classification of severity level of symptoms.

2. The logistic regression model built to classify the death in COVID-19 patients had an accuracy 69.8%. The best parameter obtained is C = 0.1. Though the F1 score is found to be 70%, the model proved to be a goodfit with moderate accuracy. The feature importance plot shows that the variable 'Intubed' and 'Pneumonia' have significant influence on death of COVID-19 patients.

3. The decision tree classifier built on the data to classify the death in COVID-19 patients had an accuracy 70.3%. Though the F1 score is found to be 70%, the model proved to be an goodfit with moderate accuracy. The feature importance plot shows that the variable 'Age' and 'Intubed' have significant influence on death of COVID-19 patients.

4. The random forest classifier built on the data to classify the death of COVID-19 patients had an accuracy 70.2%. The best parameter obtained is n_estimator = 3. Though the F1 score is found to be 70%, the model proved to be an goodfit with moderate accuracy. The feature importance plot shows that the variable 'Age' and 'Intubed' have significant influence on death of COVID-19 patients.

5. The XGBoost classifier built on the data to classify the death of COVID-19 patients had an accuracy 69.8%. The best parameter obtained is n_estimator = 7. Though the F1 score is found to be 70%, the model proved to be an goodfit with moderate accuracy. The feature importance plot shows that the variable 'Intubed' have significant influence on death of COVID-19 patients.

6. Among the four classification models built to predict the death of COVID-19 patients, the Decision Tree Classifier is the best model with good accuracy and comparatively high F1 score.

7. The ARIMA model built on the data to forecast death of COVID-19 patients had an RMSE value 5477.29. Among all the other ARIMA model ARIMA(4,1,1) is most suited model for forecasting.

6 Chapter 6

Conclusion

6.1 Conclusion

To sum up, the following conclusions can be drawn.

1. The significant proportion of males were infected by COVID-19, indicating a potential gender-based difference in infection rates. Additionally, adults and elders were identified as being at a higher risk of contracting the coronavirus disease.

2. The classification factor revealed that considerable portion of individuals experienced more severe symptoms of the disease. In terms of treatment, it reflected relatively limited need for intensive medical interventions. Additionally, the prevalence of tobacco consumption was relatively low.

3. The individual with health conditions such as pneumonia, diabetes and hypertension may be more susceptible to COVID-19. Additionally it was observed that the majority of respondents survived the novel disease.

4. The health condition, treatment type and tobacco consumption have weak association with death. The proportion of patients having pneumonia, diabetes, hypertension and icu treatment was higher among survivors while the proportion of intubed patients was higher among deceased.

5. The average age of patients having health condition and treatment type is higher compared to patients without health conditions and treatment. Also the classification of mild symptoms differing from moderate and severe symptoms levels, and moderate symptoms differing from severe symptoms levels.

6. The multinomial model for classification of severity levels of symptoms is not recommended due to very low accuracy.

7. The Decision Tree classifier is the most suited model for predicted the death of COVID-19 patients.

8. The ARIMA(4,1,1) is most suited model for predicting the death due to COVID-19.

6.2 Summary

A project entitled "Uncovering Insights and Trends: A Comprehensive Analysis of COVID-19 pandemic in Mexico" has been done. A secondary data has been collected form the website Gobierno De Mexico. The subjects considered in this study are patients aged from 0 to 110 years with problem of coronavirus disease. This data is collected in the year 2023. The dataset consists of 1338268 observations and 21 variables.

The analysis and interpretation of the data is done by using some of the statistical methods like Chi-square test of independence, Welch t-test, Proportionality test, Kruskal Wallis test, Logistic regression, Decision tree, Random forest, XGBoost, Multinomial regression, ARIMA model and some of the visualization techniques.

The results uncovered a significant gender-based difference, with a higher proportion of males being infected. Moreover, adults and elders were identified as more susceptible to the virus. The classification factor indicated a considerable portion of individuals experiencing severe symptoms, but treatment needs, including ICU and intubation, were relatively limited. Tobacco consumption was infrequent among respondents. Further analysis emphasized that health conditions like pneumonia, diabetes, and hypertension increased the susceptibility to COVID-19. Despite the challenges posed by the pandemic, a majority of respondents survived the disease. The association between health conditions, treatment type, and tobacco consumption with death was weak, with differences observed in the proportion of certain conditions among survivors and deceased. The multinomial model for symptom severity classification showed low accuracy, recommending against its use. The Decision Tree classifier was identified as the most suitable model for predicting COVID-19 deaths. Also, the ARIMA(4,1,1) model was identified as suitable model for forecasting COVID-19 deaths. These findings provide valuable insights for understanding and managing COVID-19, contributing to public health interventions and future research efforts.

7 Bibliography

1. https://towardsdatascience.com/building-a-logistic-regression-in-python-step\ \-by-step-becd4d56c9c8

2. https://www.datacamp.com/tutorial/decision-tree-classification-python

3. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

4. https://www.geeksforgeeks.org/xgboost/

5. https://www.simplilearn.com/data-science-tutorial/time-series-forecasting-in-r

6. https://www.geeksforgeeks.org/welchs-t-test-in-python/

7. https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportions_ztest.html

8. https://community.rstudio.com/t/dplyr-way-s-and-base-r-way-s-of-creating-\ \age-group-from-age/89226/4

9. Kolla Bhanu Prakash, S. Sagar Imambi, et al. (2020). Analysis,Prediction and Evaluation of COVID-19 datasets using Machine Learning algorithms-*International Journal of emerging trends in Engineering Research*. Vol. 8, No. 5.

10. Malik Khizar Hayat, Ali Daud, et al. (2020). Coronavirus Disease (COVID-19) Dynamics: Age and Gender-based Analysis of Surveillance Variables-*Preprints 2020*.

11. M. Rubaiyat Hossain Mondal, Subrato Bharati, et al. (2020). Data analytics for novel coronavirus disease. Vol 20.

12. Subhranil Das, Rashmi Kumari (2021). Statistical Analysis of COVID cases in India-*Journal of Physics: Conference Series*. Vol. 1797, No. 1.

13. Ananthu James, Jyoti Dalal, et al. (2022). An in-depth statistical analysis of the COVID-19 pandemic's initial spread in the WHO African region-*BMJ Global Health*.

14. Javanmardi F, Keshavarzi A, et al. (2020). Prevalence of underlying diseases in died cases of COVID-19: A systematic review and meta-analysis. Vol. 15, No. 10.

15. Elezkurtaj, S., Greuel, S., Ihlow, J. et al. (2020). Causes of death and comorbidities in hospitalized patients with COVID-19-*Sci Rep*. Vol. 11, No. 4263.

16. Krishnan Bhaskaran, Sebastian Bacon, et al. (2021). Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform-*The Lancet Regional Health*- Europe. Vol. 6.

8 Appendix

Python and R Codes used for the data analysis

PYTHON CODES:

```
#Importing libraries required for analysis:

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.utils import resample
from sklearn.model_selection import GridSearchCV,
RepeatedStratifiedKFold
from sklearn.metrics import accuracy_score, confusion_matrix,
precision_score
from sklearn.metrics import recall_score, f1_score,
mean_squared_error
from sklearn.metrics import classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.ensemble import RandomForestClassifier
import xgboost as xgb

#Splitting data to train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, stratify=y, random_state=42)
```

```
#Logistics Regression
model = LogisticRegression(solver='lbfgs', penalty='l2',
max_iter=1000)
model.fit(X_train,y_train)

param_grid = {'C': [0.0001,0.001,0.1, 1, 10]}
grid_search = GridSearchCV(model, param_grid, cv=5,
scoring='roc_auc')
grid_search.fit(X_train, y_train)
y_train_pred=model.predict(X_train)
y_test_pred=model.predict(X_test)
train_accuracy = accuracy_score(y_train, y_train_pred)
test_accuracy = accuracy_score(y_test, y_test_pred)
print(classification_report(y_test,y_test_pred))

# Plot the confusion matrix for test set
cm = confusion_matrix(y_test, y_test_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix - Test Set')
plt.show()

#Decision Tree
tree_clf = DecisionTreeClassifier()

param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [ 5, 10, 15],
    'min_samples_split': [2, 5, 10],
```

```
    'min_samples_leaf': [1, 2, 4]}
grid_search = GridSearchCV(tree_clf, param_grid, cv=5)
grid_search.fit(X_train, y_train)
y_pred=best_tree_clf.predict(X_test)
test_accuracy = best_tree_clf.score(X_test, y_test)
train_accuracy = best_tree_clf.score(X_train, y_train)
print(classification_report(y_test,y_pred))

# Plot the confusion matrix for test set
cm = confusion_matrix(y_test, y_test_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix - Test Set')
plt.show()

#Decision tree plot
dtree1=DecisionTreeClassifier(criterion='gini',max_depth=5,
min_samples_leaf=1,min_samples_split=2,random_state=42)
dtree1.fit(X_train,y_train)
feature_names = X.columns.tolist()
plt.figure(figsize=(40,20))
plot_tree(dtree1, feature_names=feature_names, class_names=[
'Survived', 'Death'], filled=True)
plt.show()

#Random Forest
model=RandomForestClassifier()

param_grid={'n_estimators':[1,2,3],
```

```
            'max_depth':[None,5,10],
            'min_samples_split':[2,5,10],
            'min_samples_leaf':[1,2,4]}
gs=GridSearchCV(model,param_grid,cv=5)
gs.fit(X_train,y_train)
test_accuracy = best_rf.score(X_test, y_test)
train_accuracy = best_rf.score(X_train, y_train)
print(classification_report(y_test,y_pred))

# Plot the confusion matrix for test set
cm = confusion_matrix(y_test, y_test_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix - Test Set')
plt.show()

#Random forest tree plot
rf1=RandomForestClassifier(max_depth=6,min_samples_leaf=4,
min_samples_split=2,n_estimators=3,random_state=42)
rf1.fit(X_train,y_train)
feature_names = X.columns.tolist()
plt.figure(figsize=(40,20))
plot_tree(rf1.estimators_[0], feature_names=
feature_names, class_names=['Survived', 'Death'],
filled=True)
plt.show()

#XGBoost
xgb_model = xgb.XGBClassifier()
```

```
param_grid = {
    'max_depth': [3, 5, 7],
    'learning_rate': [0.1, 0.01, 0.001],
    'n_estimators': [1, 2, 3],
    'gamma': [0, 0.1, 0.2]}
grid_search = GridSearchCV(estimator=xgb_model,
param_grid=param_grid, scoring='roc_auc', cv=5)
grid_search.fit(X_train, y_train)
test_accuracy = best_rf.score(X_test, y_test)
train_accuracy = best_rf.score(X_train, y_train)
print(classification_report(y_test,y_pred))

#Plot the confusion matrix for test set
cm = confusion_matrix(y_test, y_test_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix - Test Set')
plt.show()
```

R CODES:

```
#Importing libraries required for analysis:

library(tseries)
library(forecast)
ts.plot(X,start=2020,frequency=12)
model<- Arima(data, order = c(4, 1, 1))
residuals<- residuals(model)
data_fitted<- data - residuals
```

```
ggtsdisplay(residuals)
dataforecast=forecast(model,level=c(95),h=6)
plot(dataforecast)
Box.test(residuals,type="Ljung-Box")
accuracy(model)
```