VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

COVID DATA ANALYSIS

A Course Application Report

submitted as part of the course
Foundations of Data Analytics
CSE3505
School Of Computer Science and Engineering
VIT Chennai

FALL 2021-2022

Course Faculty : Dr. B Radhika Selvamani

Submitted By
Shivam Kumar (19BCE1098)
Kunal Kumar Jha (19BCE1212)
Tejas Vaichole (19BCE1295)
Vedang Sawarkar (19BCE1303)

Abstract

The emerging novel coronavirus (2019-nCoV) caused by a respiratory syndrome coronavirus 2 (SARS-CoV-2) is the lead cause of threat to life worldwide today. It is important to analyze the worldwide pandemic spread so that certain guide strategies can be set for complete situational awareness and application of conventional methodologies to control the impacts caused by it globally. This paper is composed of the visual exploratory data analysis of Mexico based on the number of confirmed, recovered and death cases. We found that men died more compared to women and they showed higher use of tobacco. Similarly, we looked for pandemic's impact on pregnant women, which was also found to be negligible. Next we checked if pandemic age plays a role in getting infected, which we found to be true. From analysis of data we can find the age group to target for prevention or the age group that needs more attention. Similarly, the analysis of factors causing difference in mortality between genders, can be used by government to generate guidelines to prevent such issues.

Contents

1 Introduction

Corona virus disease (Covid-19), a global pandemic has a great effect on human health worldwide since its discovery in late 2019 in Wuhan, China. Until now, many countries got effected by severe social and economic crises due to this disease. As of the end of Nov 2021, more than 200 million cases of Covid 19 have been recorded worldwide, and more than 4 million confirmed deaths. A clear understanding of the structure of the available Covid-19 datasets gives the health care provider a better understanding of identifying some of the features that can adversely affect patient at an early stage. In this article, we will be looking in to a Covid-19 Mexican Patients' Dataset (Covid19MPD) that was publically provided by the Government of Mexico, more details about the dataset will be presented in the next section.We will apply concepts learnt during our course period to perform analysis on Mexican covid dataset. The analysis will look into various aspect of the infection such gender bias, diseases effect on covid patients and factors causing the death due to covid. Such a result can be useful to the health care providers to take procative measures for the forthcoming covid-19 cases in a better efficient manner.

2 Related Works

Analysis on impact of COVID-19 plays an important role to identify the the impact and the nature of disease on people. Various works related to analysis has been carried out. Some of the notable related works are explained below.

Khaled et al. worked on the Covid19 Mexican Patients' Dataset (Covid109MPD) to select the best possible classification algorithm for the death and survived cases in Mexico [1]. The study of performance enhancement of the specified classifiers in terms of their feature selection in order to be able to predict sever, and or death, cases from the available dataset. Guillermo et al. applied a survival analysis on to investigate the impact of COVID-19 on the Mexican population [3]. From the analysis done, a plot of Kaplan-Meier curves was made, and constructed a Cox proportional hazard model. It was analyzed that the risk of dying at any time during follow-up was clearly higher for men, individuals in older age groups, people with chronic kidney disease, and people hospitalized in public health services.

A visual data analysis of the number of confirmed, recovered and death cases along with the comparative analysis of the mortality and recovery rate for nearly 222 nations worldwide was done in [2]. A k-means clustering was also used to cluster the countries according to the number of confirmed and death cases. An analysis of Factors influencing the spatial distribution of provincial cumulative confirmed count of COVID-19 in China was done in [4]. The analysis results showed that the number of COVID-19 patients diagnosed in each province in China was significantly positively correlated with the number of elderly people.

In short COVID-19 is a fast spreading contagious disease which affects a lot. Depending on the age, gender, diseases, etc the severity of COVID-19 varies from simple cold to death also.

2.1 History

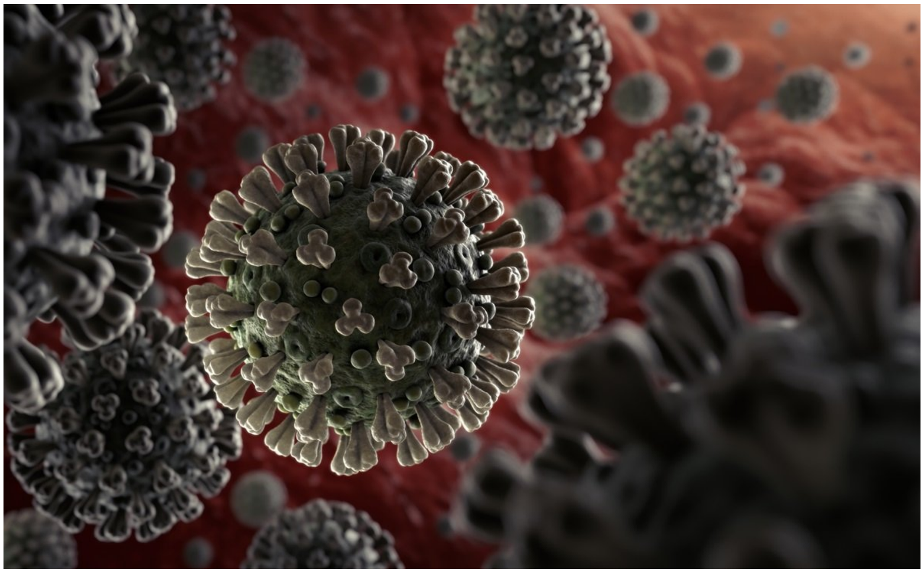


Figure 1: Image of COVID-19 Virus

3 Design

Data Description

- **sex** - [1, 2]Gender of the individual. 1 denotes Female 2 denotes Male
- **patienttype** - [1, 2] PATIENT TYPE identifies the type of care received by the patient in the unit.
- **intubed** - [1, 2, 97] INTUBED identifies if the patient required intubation.
- **pneumonia** - [1, 2, 99] PNEUMONIA identifies if the patient was diagnosed with pneumonia
- **age** - range[0:110] AGE of the tested group
- **pregnancy** - [1, 2, 98, 97] PREGNANCY identifies if the patient is pregnant
- **diabetes** - [1, 2, 98] DIABETES identifies if the patient has a diagnosis of diabetes

- **copd** - [1, 2, 98] COPD identifies if the patient has a diagnosis of COPD.
- **asthma** - [1, 2, 98] ASMA identifies if the patient has a diagnosis of asthma,
- **inmsupr** - [1, 2, 98] INMUSUPR identifies if the patient has immunosuppression.
- **hypertension** - [1, 2, 98] HYPERTENSION identifies if the patient has a diagnosis of hypertension.
- **other_disease** - [1, 2, 98] OTRAS COM identifies if the patient has a diagnosis of other diseases.
- **cardiovascular** - [1, 2, 98] CARDIOVASCULAR identifies if the patient has a diagnosis of cardiovascular diseases.
- **obesity** - [1, 2, 98] OBESITY identifies if the patient is diagnosed by obesity.
- **renal_chronic** - [1, 2, 98] RENAL CHRONIC identifies if the patient has a diagnoses of chronic kidney failure.
- **tobacco** - [1, 2, 98] TOBACCO identifies if the patient has a smoking habit.
- **contact_other_covid** - [1, 2, 99] OTHER_CASE identifies if the patient had contact with any other cases.
- **covid_res** - [1, 2, 3] RESULT identifies the result of the analysis of the sample reported by the laboratory.
- **icu** - [1, 2, 97] ICU identifies if the patient required to enter an Intensive Care Unit.
- **class** - [1, 2] Indicating if the patient passed away or survived the covid19.

Our dataset deals with CVOID-19 patients. initial part of our design includes data pre-processing. In this part we have converted the string dates to date format, converting the encoded gender(0 and 1) to gender labels(”male” and ”female”), converting the encoded patient type(0 and 1) to patient labels(”inpatient” and ”outpatient”), creating a mortality column based on the date died attribute, converting the encoded presence and absence of disease (0 and 1) to labelled presence and absence of disease (”yes” and ”no”), converting encoded COVID result(1,2 and 3) to labelled result (1=COVID positive, 2 = COVID negative and 3 - awaited result) and finally converting the encoded ICU admission(0 and 1) to labelled ICU admission(1 = admitted in ICU and 0 = not admitted in ICU).

After pre-processing of the data-set, we have used the pre-processed data to solve the questions related to COVID-19 data-set analysis.

Our question analysis includes :

1. Is COVID related death biased towards a specific gender. If yes, suggest some feature that result in this bias?

2. Is this time of pandemic a bad period for couples to consider about pregnancy option, specific to the effect of COVID-19 on pregnant women?

3. Impact of COVID on respiratory diseases.

4. Correlation between the different types of diseases in COVID time.

5. One of the most challenging problem plaguing the world is Obesity and Hypertension, does these factors have any impact on the COVID-19 test result?

6. What is the impact of age on getting infected with COVID-19 and the chances of making a recovery?

7. Find out percentage of people with "other factors"= 13 who were put in ICU.

8. Get relation between (date of first symptom, date of entry in hospital) with intubed, icu. You may use arithmetic operations, graphs, regression, conditional statements, etc.

9. Does the time period between appearance of first symptoms to the time of hospitalization have any effect on the seriousness of the COVID-19 on the patient.

3.1 sample

4 Results & Discussion

. A detailed analysis/interpretation of the above should be given by explaining figure**??**

1. COVID related death biased towards a specific gender. Tobacco consuming people suffered more from COVID-19 deaths. Men consumed more tobacco than female. Females suffered more from hypertension.See figure 2 & figure 3

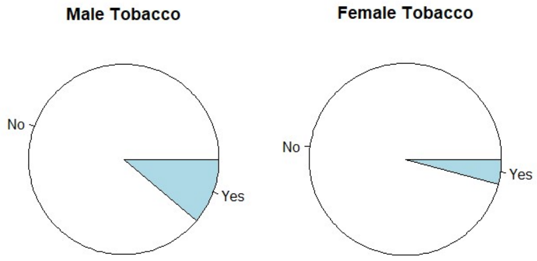


Figure 2: Tobacco

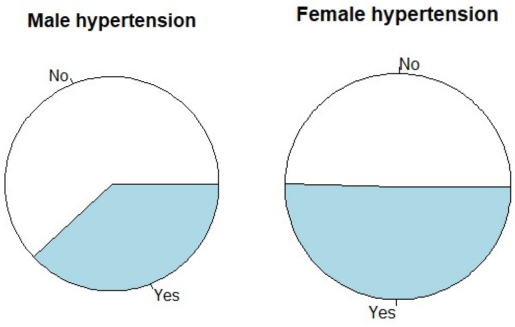


Figure 3: hypertension

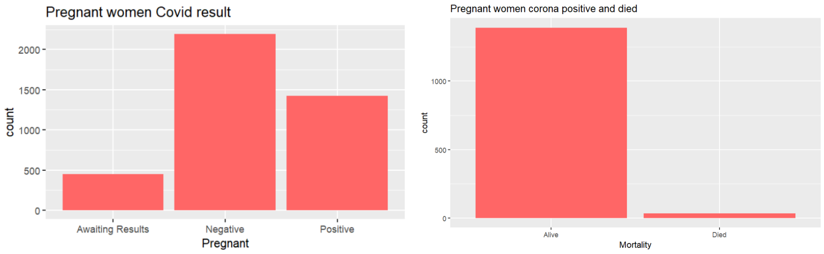


Figure 4: Effect of Covid on pregnancy

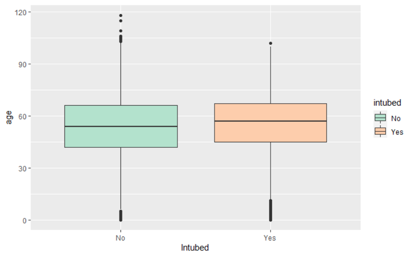
2. This time of pandemic was not a bad period for couples to consider about pregnancy option, specific to the effect of COVID-19 on pregnant women.

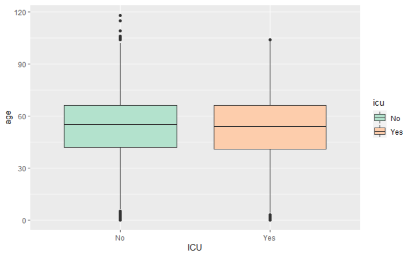
3. COVID had some impact on respiratory diseases. According to our analysis, people suffering from pneumonia suffered more from COVID-19 as its correlation value with COVID results was nearly 19%.

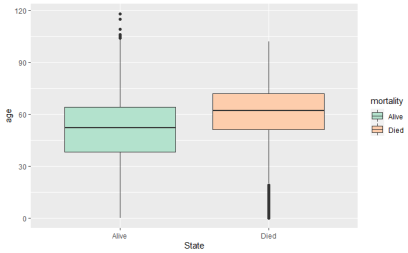
4. We calculated correlation between different types of diseases. According to our analysis, diabetes had high correlation (21%) with asthma, renal chronic disease had high correlation(38%) with asthma.

5. The correlation result of obesity on the covid is low (8%) and for hypertension on covid is also low (7%). It means the impact of obesity and hypertension on covid is negligible.

6. Age is a factor which affects chance of getting affected. From, the data it was observed that people in age group 20-80 were more infected compared to the rest. Mortality is higher in higher age group(60+). For in-tubed & ICU cases age isn't a major factor.







7. This task was to consolidate all the data related to disease in form of an array for every patient.

8. After the above task it was asked to check for percentage of people having all the diseases, it was found that only 0.001% have all the mentioned diseases i.e. hypertension, cardiovascular, obesity, diabetes, asthma, etc.

9. We looked if there is some relation between effect of covid and time taken to admit in hospital(time difference between first symptom and admission to hospital). It is observed that most of the people are admitted to hospital within 5 days within those days we see the most cases where people die or get admitted to ICU or need to use ventilator. This implies that the disease affects the body really quickly and causes very huge damage to patient's body. This also means that the hospitals get frequent critical situations where they have to accommodate such cases. Thus, implying the need to care of healthcare workers and doctors.

5 Conclusion & Future Work

This analysis observes that there is higher risk of death for men and older people, in Mexico .Though it was found that there is higher higher consumption of tobacco in men higher hypertension in women, which could be potential causes for this bias.Hence, further research is necessary to understand the origin of this differential risk clearly. It was also observed that covid-19 was not specifically bad for pregnant women, but the sample size for pregnant women was comparatively lesser, so we can expect some changes in this overtime. We

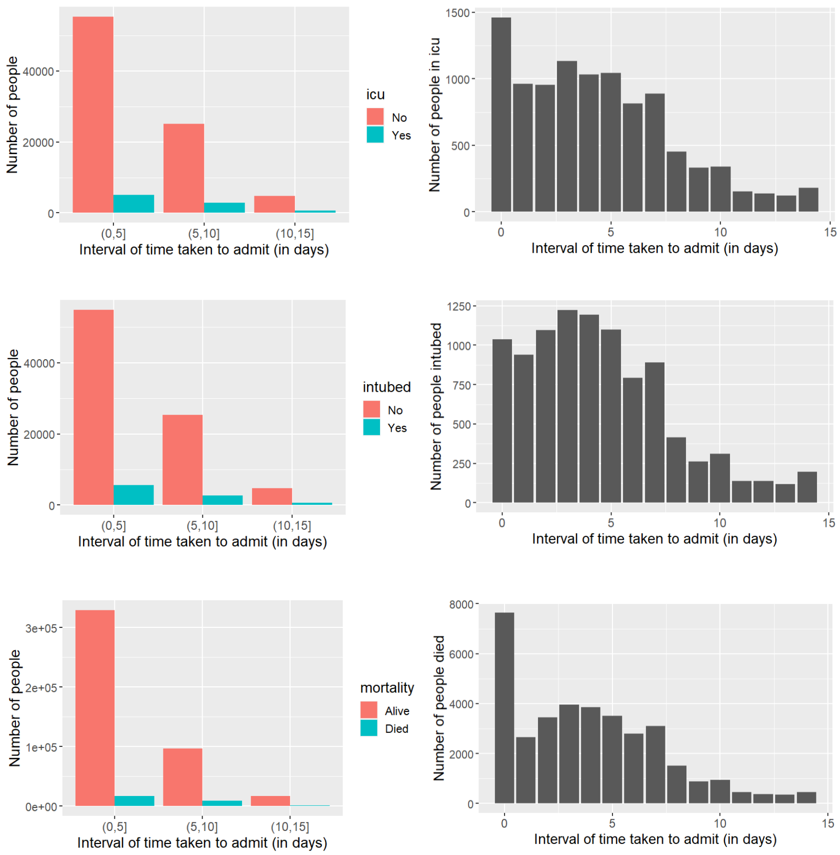


Figure 6: Effect of time to admit in hospital on covid

also found that there is correlation between covid and respiratory diseases, this result further adds to the previous studies done on this topic. When looking at difference between time of appearance symptoms and getting admitted to hospital with conjunction to mortality, ICU admission and getting in-tubed, it was found that these things escalate really quickly after the show of symptoms. Hence, there is a need to follow guidelines properly. It also means that the hospitals are also met with serious cases immediately after admission which can explain the increasing burden on the healthcare system. Given the persistence of SARS-Cov-2, it is vital to focus on mitigation campaigns on the population at higher risk of fatal outcomes. We also need stronger public health campaigns aimed at reducing the prevalence of obesity, diabetes, hypertension, and their comorbidities. These conditions may increase the susceptibility of the Mexican population to COVID-19 and cause a rapid progression to severe states of the disease and death. For, future work researcher can create prediction model using past history of patient to predict effect of covid on patient. Furthermore, mutation-specific analysis could be done to find potential harbinger to that mutation. This data can be applied to different regions to apply mutation specific guidelines.

Acknowledgments

We sincerely thank you and Team DBZ for giving us the opportunity to work on this project. We would also like thank the the people we have referenced for our work.

References

[1] Khaled Mohamad Almustafa. Covid19-Mexican-Patients' dataset (Covid19MPD) classification and prediction using feature importance. *Concurr. Comput.*, page e6675, 2021.

[2] Sumindar Kaur Saini, Vishal Dhull, Sarbjeet Singh, and Akashdeep Sharma. Visual exploratory data analysis of COVID-19 pandemic. In *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6. IEEE, 2020.

[3] Guillermo Salinas-Escudero, María Fernanda Carrillo-Vega, Víctor Granados-García, Silvia Martínez-Valverde, Filiberto Toledano-Toledano, and Juan Garduño-Espinosa. A survival analysis of COVID-19 in the mexican population. *BMC Public Health*, 20(1):1616, 2020.

[4] Liyun Su and Haiqin Yu. Analysis of factors influencing the spatial distribution of provincial cumulative confirmed count of novel coronavirus pneumonia (COVID-19) in china. In *2020 International Conference on Public Health and Data Science (ICPHDS)*, pages 81–85. IEEE, 2020.