Coronavirus Mortality Prediction Using a Classification Prediction Model

Albina Cako, Colin Green, Lucy Zhang, Sean Zhang

Abstract abstract goes here

Background

Covid-19 is a viral infection caused by the novel virus called severe acute respiratory syndrome 2 (SARS-COV-2). Incubation period of Covid-19 is 2-14 days, which is the time a person shows symptoms after being exposed to the virus. The main symptoms are fever, cough, runny nose, headache, fatigue and muscle aches. The virus was first recorded in China in Fall 2019, however, around February 2020 it started rapidly spreading across the world. On March 11, 2020, the World Health Organization declared Covid-19 to be a global pandemic. One of the most tragic events of Covid-19 was the first wave of the virus in Italy, where around 44,000 deaths and 250,000 infections were recorded in just 3 months, from March to May 2020. Overwhelmed hospitals, shortage of medical supplies, hospital beds, icu units and medical staff, were a great challenge during the first wave of the virus. However, the fear of these factors is rising as the second wave of the pandemic is currently happening in the world. The virus has caused over 3 million deaths and has spread across 188 countries to date.

Governments all over the world have instructed people who tested positive for Covid-19 to quarantine for 14 days if they have mild symptoms or to visit the nearest hospital if they experience difficulty breathing, cyanosis or chest pain. People are considered at high risk of developing more serious symptoms or death if they have pre-existing medical conditions such as, but not limited to, chronic obstructive pulmonary disease (COPD), chronic kidney disease, obesity, smoking, immunocompromised, heart conditions, type two diabetes and cancer. Older age population is also at higher risk, as they are more likely to have a pre-existing medical condition and lower immunity.

As Covid-19 is still a threat to public health worldwide and the second wave of the virus is now growing globally, information is needed by medical professionals and people tested positive with Covid-19 when making decisions on whether hospitalization is needed. The use of machine learning is growing in the fight against Covid-19.

A death prediction app can assist medical professionals to make decisions on which patients are at high risk and require hospitalization. In addition, the app can also assist people infected with Covid-19 in deciding whether to self-quarantine or go to a hospital. We expect that our application, when used in conjunction with expert medical advice, will help guide resource allocation during the second wave.

Objective

The objective of this project is to create a supervised machine learning model that predicts probability of death due to Covid-19 based on age and pre-existing medical conditions. This app could be used by medical professionals to guide resource allocation for the treatment of Covid-19 patients. Having an app that predicts death probability can help hospitals decide which patients need hospitalization. In addition, the app can also be used by people who test positive with covid-19 as a guide to decide whether they should visit the hospital. Having this app can reduce unnecessary surges to the hospital of people who are at low risk. The app is simple to use and would be easily accessible by anyone.

Disclaimer: This application should only be used as a guide and is not a replacement for expert advice or government recommendations.

Data Analysis

The dataset was made publicly available from the Mexican government. It has over 560,000 anonymous records of patients who attended a hospital from March – July 2020 regarding Covid-19. Over 200,000 records were confirmed positive for Covid-19.

Data Dictionary

The table below defines each column in the dataset.

Label	Description
id	Case identifying number
sex	Sex of patient, where 1 means female, 2 means male, 99 means non-specified
patient_type	Describes the type of care the patient received, where 1 means discharged the same day, 2 means hospitalized, 99 means non-specified
entry_date	Hospital entry date
date_symptoms	The first date the patient started experiencing symptoms
date_died	The date the patient died if date is shown or 9999.99 if patient did not die or it is not-specified whether they died or not
intubed	Whether a patient was intubed in the hospital, where 1 means yes, 2 means no, 97 means not applicable. Our analysis showed that 97 was used for patients who were discharged the same day.
pneumonia	Whether the patient had pneumonia as a preexisting condition, where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
age	Patient's age
pregnancy	Whether a female patient was pregnant or not. Our analysis showed that 97 that means not applicable was used for male patients
diabetes	If the patient had diabetes as a preexisting condition, where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified)
copd	Whether the patient had copd - chronic obstructive pulmonary disease as a preexisting condition , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
asthma	Whether the patient had asthma as a preexisting condition , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
inmsupr	Whether the patient was immunocompromised, where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
hypertension	Whether the patient had hypertension as a preexisting condition , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
other_disease	Whether the patient had another disease not listed as a preexisting condition, where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
cardiovascular	Whether the patient had a cardiovascular disease as a preexisting condition, where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
obesity	Whether the patient was obese , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
renal_chronic	Whether the patient had chronic renal disease as a preexisting condition , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified
tobacco	Whether the patient smoked tabacco, where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified

Label	Description
contact_other	Whether the patient had contact with apositive case prior to contacting the virus , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified.
covid_res	Whether the patient tested positive or not , where 1 means yes, 2 means no, 97 means not applicable, 98 means ignored/not specified, 99 means not specified.
icu	Whether a patient was intubed in the hospital, where 1 means yes, 2 means no, 97 means not applicable. 97 was used for patients who were discharged the same day.

```
covid <- read_csv("covid.csv")
covid <- covid %>% filter(covid_res == 1)
```

Data Exploration In order to explore the data further we will need to see how death is related to certain aspects of the data.

```
covid<- covid %>%
  mutate(deathyn = ifelse(date_died == '9999-99-99', '0','1'))
```

Missing Data

We first visualized the missing data.

```
## Loading required package: colorspace
```

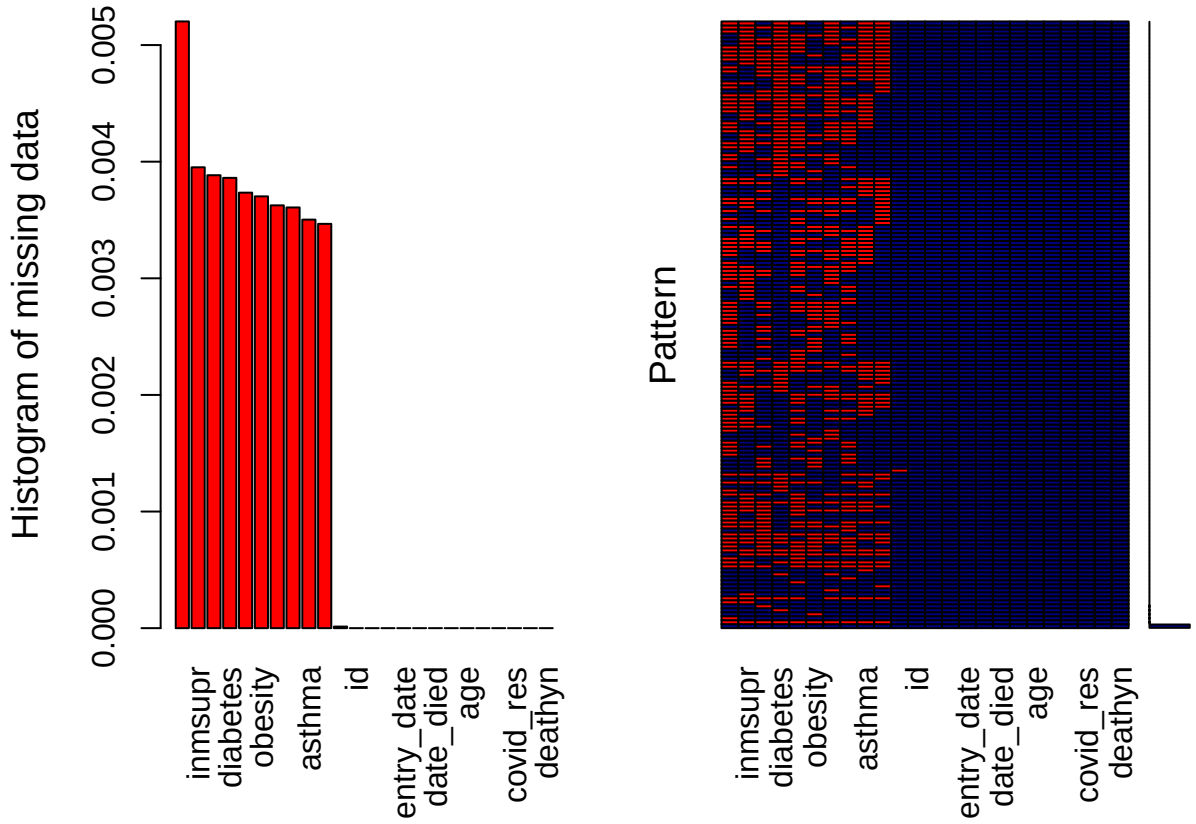
```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
## Warning in plot.aggr(res, ...): not enough vertical space to display frequencies
## (too many combinations)
```



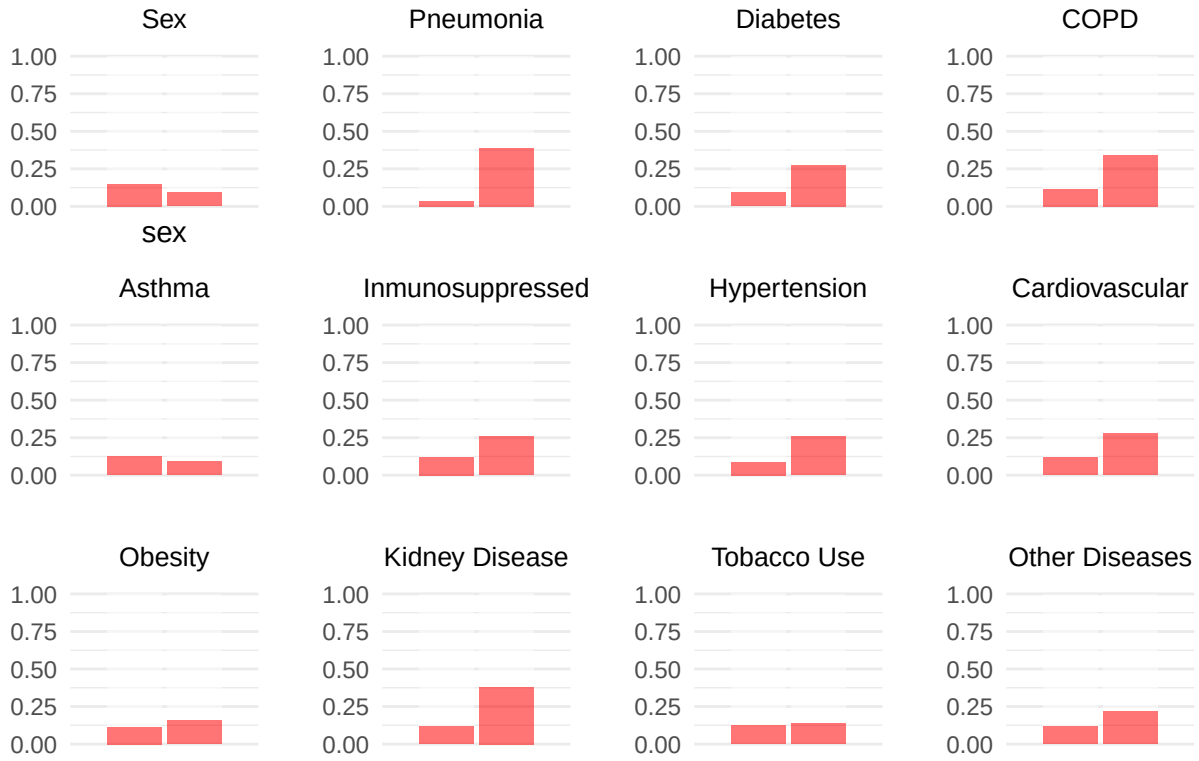

```
##
##  Variables sorted by number of missings:
##              Variable        Count
##        other_disease 5.202645e-03
##              inmsupr 3.951835e-03
##              tobacco 3.883856e-03
##             diabetes 3.861196e-03
##       cardiovascular 3.734303e-03
##              obesity 3.702579e-03
##         hypertension 3.625536e-03
##        renal_chronic 3.607409e-03
##               asthma 3.503175e-03
##                 copd 3.466919e-03
##            pneumonia 1.359576e-05
##                   id 0.000000e+00
##                  sex 0.000000e+00
##         patient_type 0.000000e+00
##           entry_date 0.000000e+00
##        date_symptoms 0.000000e+00
##            date_died 0.000000e+00
##              intubed 0.000000e+00
##                  age 0.000000e+00
##            pregnancy 0.000000e+00
##  contact_other_covid 0.000000e+00
##            covid_res 0.000000e+00
##                  icu 0.000000e+00
```

```
##              deathyn 0.000000e+00
```

To clean the data, we converted all 2's that represent "no"s into zero's, since we are using a binary system for all preconditions. This excludes age as it is not a binary categorical variable. Missing data in the columns were characterized by 97 = not-applicable, 98 = ignored or not specified or 99 = not specified. We had to understand what these variables meant. We analyzed the data and realized that 97 was used in 3 columns: pregnancy, intubed or icu. Considering that pregnancy only applies to females, we checked whether 97 indicated that the person was male. Our analysis showed that 97 meant male patient. It was changed to zero in our table. We also explored 97 value in intubed and icu columns. We realized that it would be inapplicable to say a patient was intubed or on icu if they were discharged the same day. We did ran an analysis and noticed that 97 was only used for patients who were not hospitalized in both cases. Since patient's were not hospitalized, that means that they did not need to be intubed or placed in icu, therefore 97 means "no" in these two columns. We substituted the 97 values with 0's (no) in our tables. The rows that contained the values 98 and 99, which mean ignored or not specified were removed from our data. These values were ignored in the icu and intubed columns which were not included in our analysis, as they are not a precondition. This still left us with a dataset of 218275 cases.

```
##                         cases death% mean_age median_age mean_age_death
## covid                  220657      0    45.69         45          61.00
## covid_complete_case    218275      0    45.66         45          61.03
## covid_incomplete_cases   2382      0    48.52         48          58.98
## covid_ignored_cases      2379      0    48.54         48          58.98
## covid_blank_cases         536      0    49.17         47          60.26
##                        mean_days_died std_death_days male%
## covid                           11.51           7.39 54.75
## covid_complete_case             11.51           7.39 54.83
## covid_incomplete_cases          11.14           7.59 47.02
## covid_ignored_cases             11.14           7.59 46.99
## covid_blank_cases               10.62           7.54 64.93
```

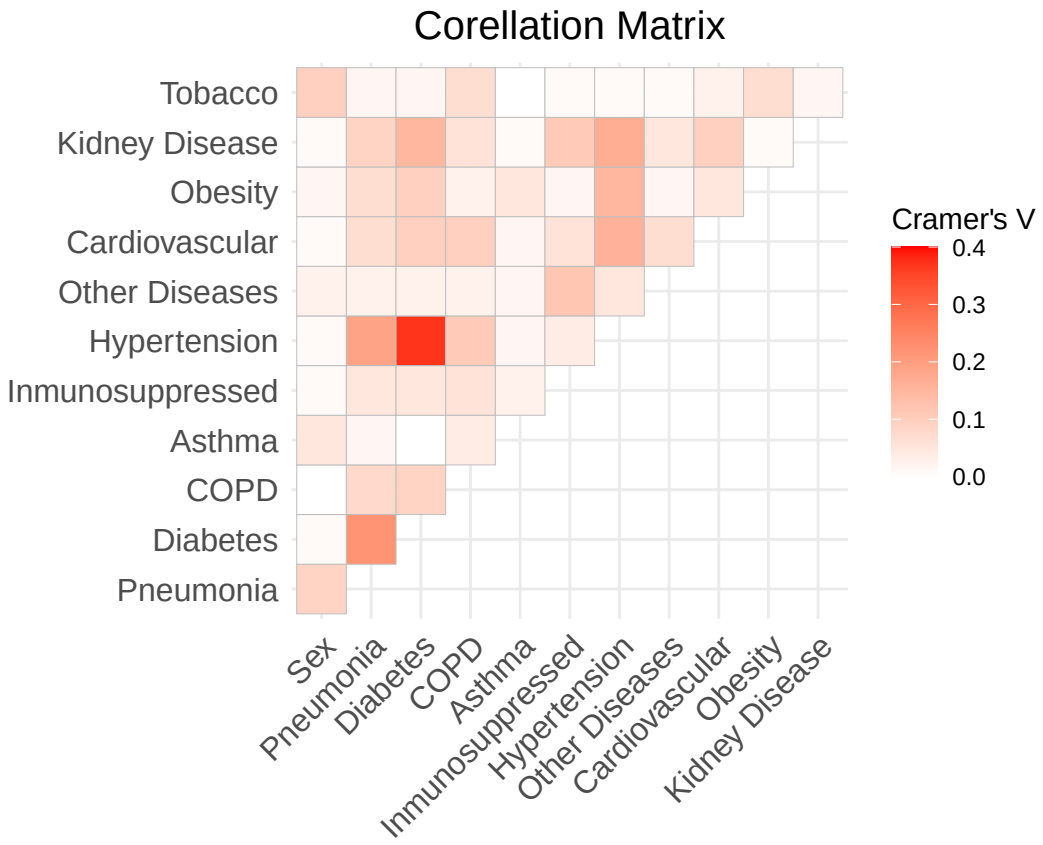
We plotted individual graphs to predict the relationship between our independent variables and our target variable death.



The graphs above show the comparison between our predictor variables and death. The proportion of deaths is represented by the y-axis. For example, in the "Sex" variable graph, we observe that the male sex has a greater proportion of individuals who died from Covid-19 compared to the female sex. The highest proportion of death is seen in patients with pneumonia, kidney disease, and COPD.

Correlation of variables

To perform the correlation between our independent variables, we used the Chi-squared test, as a standard correlogram using Pearson's R does not accept categorical variables. However, we were unable to use the original p-value provided by the test as samples higher than a few hundred typically always return $p < 0.05$. To adjust for our large sample size, we used Cramér's V test instead. A higher Cramér's V value denotes stronger correlation.



The above correlogram shows correlations between the categorical predictor variables. A value greater than 0.5 typically represents collinearity; as all values fall below that threshold, we can assume independence.