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, starting in 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, which caught the whole worlds' attention, 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 that is currently happening in the world. The virus has caused over 3 million deaths, spread across 188 countries and it is currently in it's second wave. The governments all over the world have instructed people positive with 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 preexisting 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 preexisting 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 testing positive with Covid-19 when making decisions in whether hospitalization is important. 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 patient's are at high risk and require hospitalization. In addition, the app can assist people infected with Covid-19 in deciding whether to self-quarantine or go to a hospital. This can reduce overwhelming of hospital beds by people who are at low risk, which would also help reduce shortage of medical supplies and staff. Such app can assist in making better decisions during the second wave.

Objective The objective of this research is to use supervised learning to create a model that predicts rate of death depending on age and pre-existing medical conditions. This app is to be used by medical professionals to help them make decisions for hospitalization purposes and treatment of Covid-19 patients. Having an app that predicts possible death rate, can help hospitals decide which patients need hospitalization. In addition, the app can also be used by people who test positive with covid-19, to decide whether they need to visit the hospital. Having this app can reduce unnecessary surges to the hospital of people who are at low risk. The app is quick and easy to use and can be used by everyone. CAUTION with using app needed here.

Data Analysis The dataset that we are going to use was obtained from the Mexican government. It has over 560,000 anonymous records of patients who attended a hospital from March – July 2020 regarding

covid-19. The data was obtained from Kaggle.

Data Dictionary

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
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.

Label	Description
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)
```

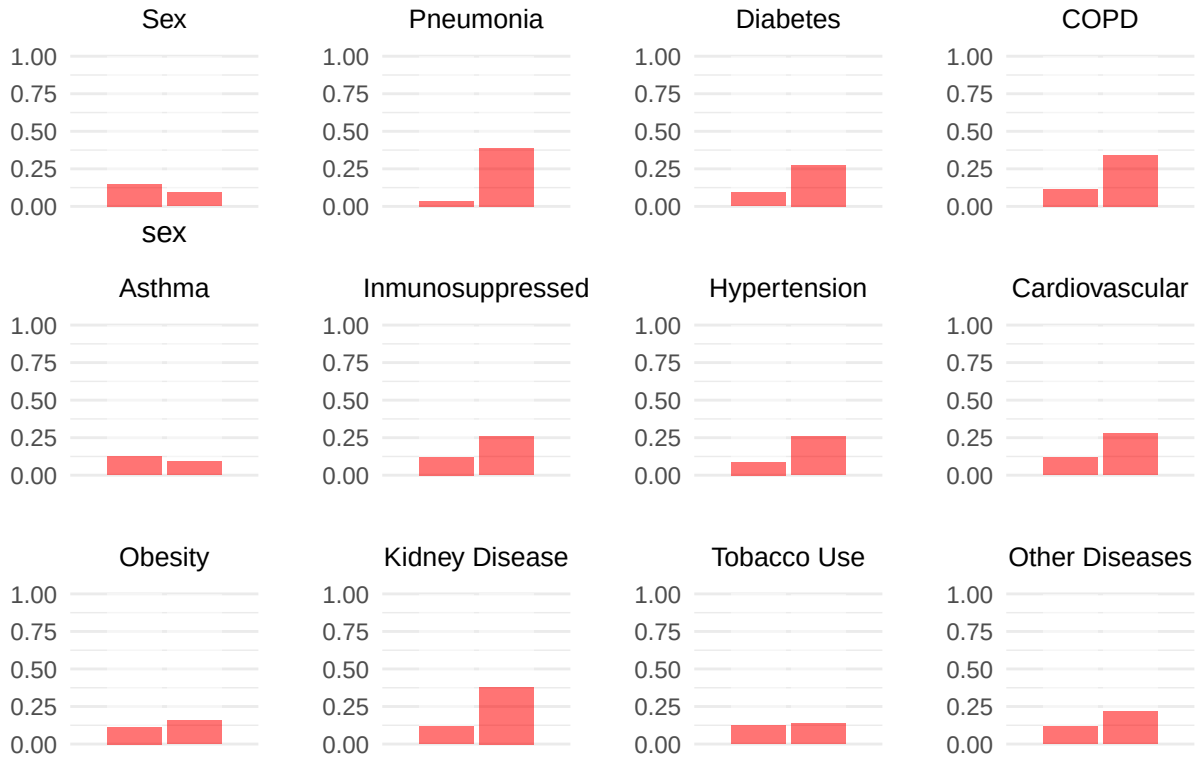
Data Exploration In order to explore the data further we will need to see how death is related to certain aspects of the data.

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

Missing Data To make the data appear more clean 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 precondition. Missing data in the columns was 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 left as 97 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 precondtion. This still left us with a dataset of 218275 cases.

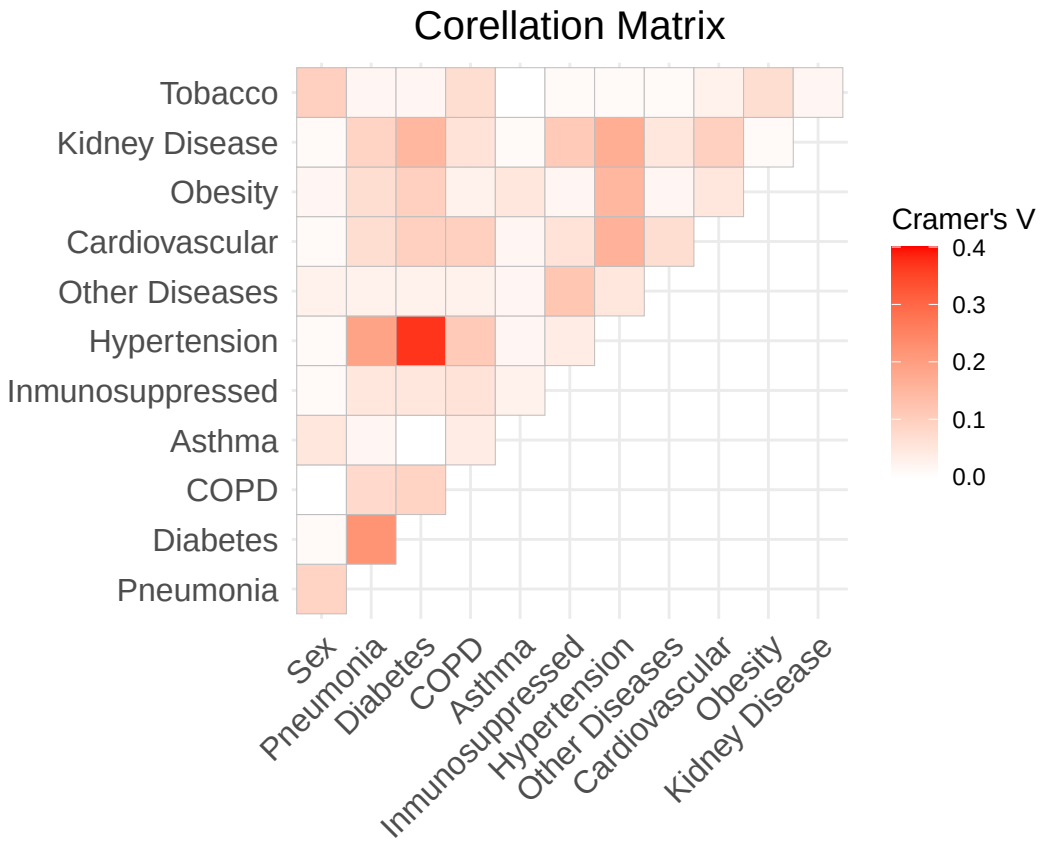
```
##                          cases death% mean_age median_age mean_age_death
## covid                   220657      0    45.69         45          61.00
## covid_complete_case     218275      0    45.66         45          61.03
## covid_incomplete_cases    2382      0    48.52         48          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 correlation of our independent variables to the column death. On the "Sex" variable graph we observe that the sex male is more strongly correlated to death from Covid-19 then the female sex. From the preconditions variable, the yes and no columns predict how strong the correlation between each condition and death is. The percentage of correlation can be seen in the y-axis. The strongest correlation to death is seen in patients' with pneumonia, kidney disease and COPD. Lowest correlation is seen in patients' who smoke, are obese and have asthma as a preexisting condition.

Correlation of variables To perform the correlation between our independent variables used the chi-squared test. However, we were unable to use the p-value provided by the test because our sample size is too large. We used Cramér's V test instead, which takes into account the sample size. Higher value is more correlated.



Description of correlation matrix and its findings, nothing above 0.5 suggests we are safe correlation wise.

Takeaways from Data Exploration During data exploration we learned that

Data Imputation Data exploration showed that there were patients' of age 0 up to 120. As 120 years old could be an error in imputation or an outliar, to reduce error we used 3 standard deviations to replace the upper age range. Any age above 95 years old was replaced with 95. Age of 0 was left as it is, as there could be patients newborn or under the age of 1 years old.

```
mean_age <- mean(covid_complete_case$age)
sd_age <- sd(covid_complete_case$age)
imputed_age <- round(mean_age + 3*sd_age)
covid_complete_case$age[covid_complete_case$age > imputed_age] = imputed_age
```

Modeling and Evaluation

Feature Selection

Train and Test Data

Logistic Regression Model

```
##     (Intercept)            sex1      pneumonia1             age      pregnancy1
##     -5.80930786     -0.45302764      2.25528335      0.05246929      0.07019434
##     pregnancy97       diabetes1           copd1         asthma1        inmsupr1
##              NA      0.32907552      0.15899211     -0.12368372      0.30847526
##   hypertension1  other_disease1 cardiovascular1        obesity1  renal_chronic1
##      0.18531198      0.48283785     -0.09276075      0.24359063      0.74618758
##        tobacco1
##     -0.07552174
```

```
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    0    1
##          0 84.7  8.2
##          1  3.1  4.0
##
##  Accuracy (average) : 0.8866
```

Decision Tree Model

```
##     pneumonia1            age  hypertension1 renal_chronic1      diabetes1
##    9828.048878    2253.679953     231.048502      74.896075      58.411769
##     pregnancy1       obesity1    pregnancy97           sex1        asthma1
##      17.397825      10.096798       8.878814       8.878814       5.530805
```

```
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    0    1
##          0 85.0  8.5
##          1  2.8  3.7
##
##  Accuracy (average) : 0.8868
```

Random Forest Model

```
##                MeanDecreaseGini
## sex                     2050.195
## pneumonia              35298.290
## age                    13258.784
## pregnancy              10605.882
## diabetes                5081.880
## copd                    8700.317
## asthma                  1837.096
## inmsupr                 4693.982
## hypertension            3125.686
## other_disease           3389.804
## cardiovascular          5289.382
```

```
## obesity                2015.949
## renal_chronic         11908.067
## tobacco                1881.653
```

```
##           0          1
## 0 0.4517650 0.04834727
## 1 0.0377368 0.46215096
```

Gradient Boosting

```
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    0    1
##          0 84.8  8.3
##          1  3.0  3.9
##
##  Accuracy (average) : 0.8867
```

Model Selection

Deployment