Chose your own project: COVID-19 patient death rate analysis

Iván de Luna

26/7/2021

The database used in this exercise can be downloaded from the Dirección General de Epidemiología Datos Abiertos website https://www.gob.mx/salud/documentos/datos-abiertos-152127 We will use the july 10th 2021 database, it is recommended that if the database does not load succesfully, try another time given the server side load.
###

Introduction

There has been a lot of debate about the causes of death regarding COVID-19, given the prevalence of the desease and the array of diferent sympstons that can complicate the treatment. In the Mexico case, the database that has been collected oficially has changed because of the further research that has given more light on what can be a probable rease a patient did not survived treatment. At the begining it was implied that any preexisting respiratory conditions may have a certain increase in the probability of not surviving treatment but after some time, factor such as hypertension, chronic renal disease and obeisty have become the common suspects.

This exercise will analyse the causes of death regarding registered conditions given the database selected and create a data model with the most relevant ones in trying to improve accuracy.

Metodology

The database that can be downloaded from the DGE website contains a csv zipped file with all the information from the whole National Health Service System, including private and public hospitals and medical facilities. There is an additional file names "Dictionario de datos" which contains the description and possible values in the database. There are some considerations regarding this analysis, in which we will define the deceased conditions as having a defunction date as stated in the database. Also we will not emphasise in pregnancy condition, or any other condition that cannot be directly related to a health pre-existing or developed condition.

Exploratory and data analysis

The database, which contains 7,732,694 records with 40 variables, not all of them will be useful for this analysys. For example, there is a unique identifier for every person that has been registered in a COVID-19 related case. This does not mean that the patient is positive.

```
## NULL
```

There are also variables that specify where does the patient comes from, where is being treated or, given the ambulatory patients, whom did not stayed at a hospital, this may be usefull in term of analyzing infraestructure or health provider availability.

```
##   FECHA_ACTUALIZACION ID_REGISTRO ORIGEN SECTOR ENTIDAD_UM SEXO
ENTIDAD_NAC
## 3          2021-07-10      z23d9d      1     12         22    2
24
## 4          2021-07-10      z24953      1     12          9    1
9
## 6          2021-07-10      z1b0d1      1     12          1    1
1
## 7          2021-07-10      z2d0c4      1     12          9    1
9
## 8          2021-07-10      z26b82      2     12          9    1
9
##   ENTIDAD_RES MUNICIPIO_RES TIPO_PACIENTE FECHA_INGRESO FECHA_SINTOMAS
## 3          22             9             1    2021-01-05     2021-01-05
## 4           9            10             1    2020-10-15     2020-10-15
## 6           1             3             1    2020-04-23     2020-04-21
## 7           9             6             1    2020-10-15     2020-10-14
## 8           9             7             1    2021-01-14     2021-01-10
##    FECHA_DEF INTUBADO NEUMONIA EDAD NACIONALIDAD EMBARAZO
HABLA_LENGUA_INDIG
## 3 9999-99-99       97        2   29            1       97
2
## 4 9999-99-99       97        2   40            1       98
99
## 6 9999-99-99       97        2   48            1        2
2
## 7 9999-99-99       97        2   60            1        2
2
## 8 9999-99-99       97        2   20            1        2
2
##   INDIGENA DIABETES EPOC ASMA INMUSUPR HIPERTENSION OTRA_COM
CARDIOVASCULAR
## 3        2        2    2    2        2            2        2
2
## 4       99        2    2    2        2            2        2
2
## 6        2        1    2    2        2            2        2
```

```
2
## 7        2        2    2    2        2            2        2
2
## 8        2        2    2    2        2            2        2
2
##   OBESIDAD RENAL_CRONICA TABAQUISMO OTRO_CASO TOMA_MUESTRA_LAB
RESULTADO_LAB
## 3       98             2          2         2                2
97
## 4        2             2          2         1                1
2
## 6        1             2          2        99                1
4
## 7        1             2          2         2                1
2
## 8        2             2          2         2                2
97
##   TOMA_MUESTRA_ANTIGENO RESULTADO_ANTIGENO CLASIFICACION_FINAL
MIGRANTE
## 3                     2                 97                   6
99
## 4                     2                 97                   7
99
## 6                     2                 97                   5
99
## 7                     2                 97                   7
99
## 8                     1                  2                   7
99
##   PAIS_NACIONALIDAD PAIS_ORIGEN UCI
## 3           MÃ©xico          97  97
## 4           MÃ©xico          97  97
## 6           MÃ©xico          97  97
## 7           MÃ©xico          97  97
## 8           MÃ©xico          97  97
```

For the date variables, there is an update record (fecha_actualizacion), a registration record (fecha_ingreso), an initial symptons record (fecha_sintomas) and a defunction record (fecha_def). Only the last variable will be useful in order to diferentiate patients that survived or died.

```
## Warning: 7429199 failed to parse.
```

Given that the database has a lot of records, for speed and memory management purposes we will subset it by state, we will chose the number 5, which refers to the Coahuila state and we will subset the variables to the conditions that we are interested in:

1. id_registro - Unique identifier
2. sector - the type of medical unit the patient was registered

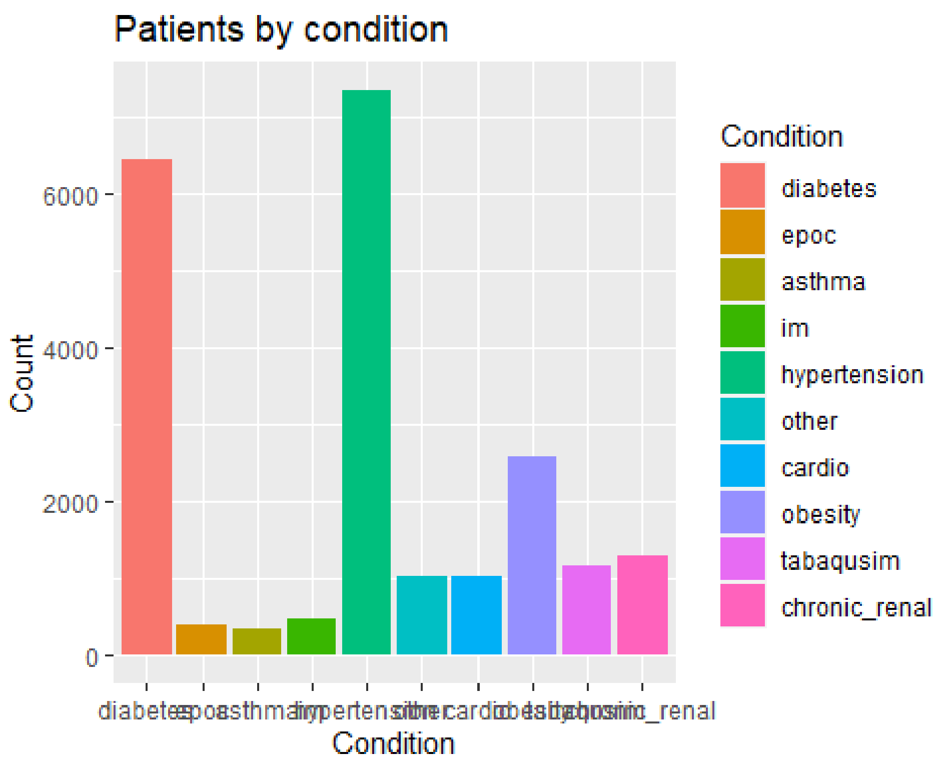
3. sexo - gender of the patient
4. fecha_def - date when the patient died
5. intubado - the patient was assigned a ventilator
6. neumonia - Neumony diagnose
7. edad - patient age
8. diabetes - Diabetes condition
9. epoc - EPOC condition
10. asma - Asthma condition
11. inmusupr - Immunosuppresion condition 12.hipertension - Hypertension condition
12. otra_com - Other conditions
13. cardiovascular - Cardiovascular condition
14. obesidad - Obesity condition
15. renal_cronica - Chronic renal condition
16. tabaquismo - Smoker condition
17. clasificacion_final - Final results for COVID19 test

There are some considerations regarding the database, in which the conditions are arranged in a factor format classified as:

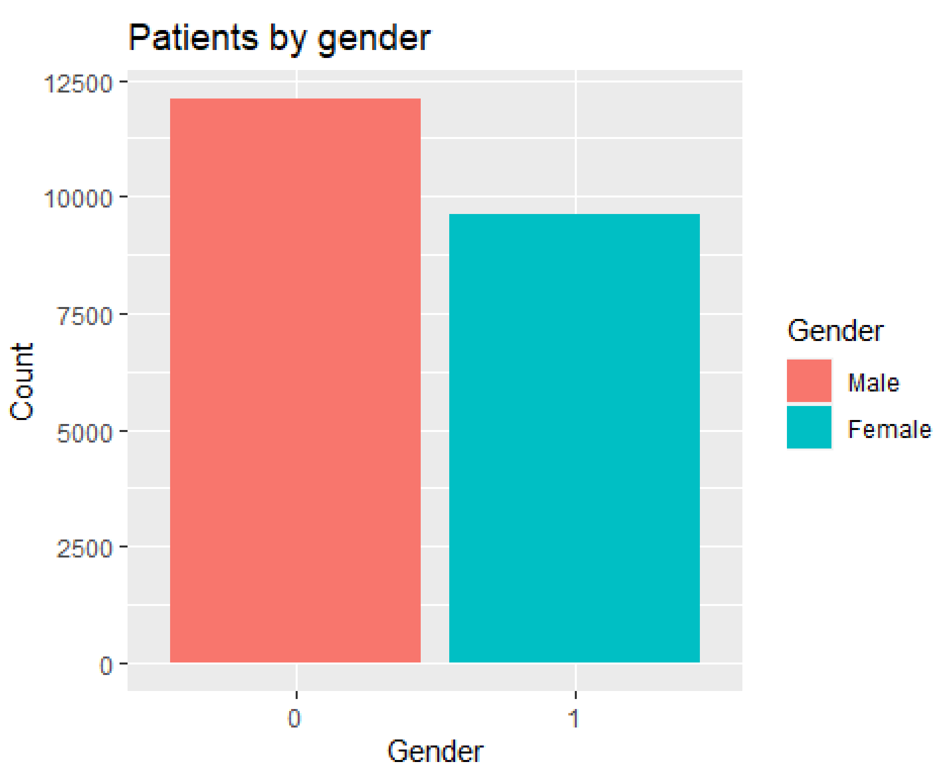
- 1 - the patient has the condition
- 2 - the patient does not have the condition
- 97 - the condition does not apply (such as pregnancy in male patients)
- 98 - there is not enough information to determine if the condition exists
- 99 - not specified

Which will be further arranged as a binomial option of having or not having certain condition, of which the most prevalent are hypertension, diabetes and obesity.

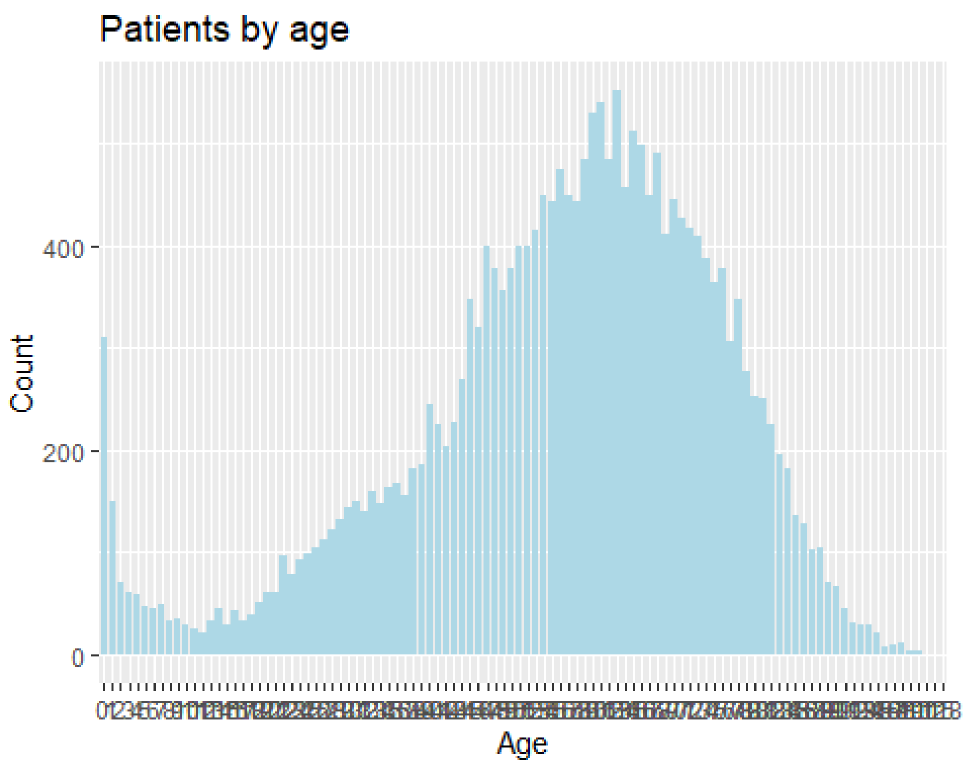
```
## No id variables; using all as measure variables
```



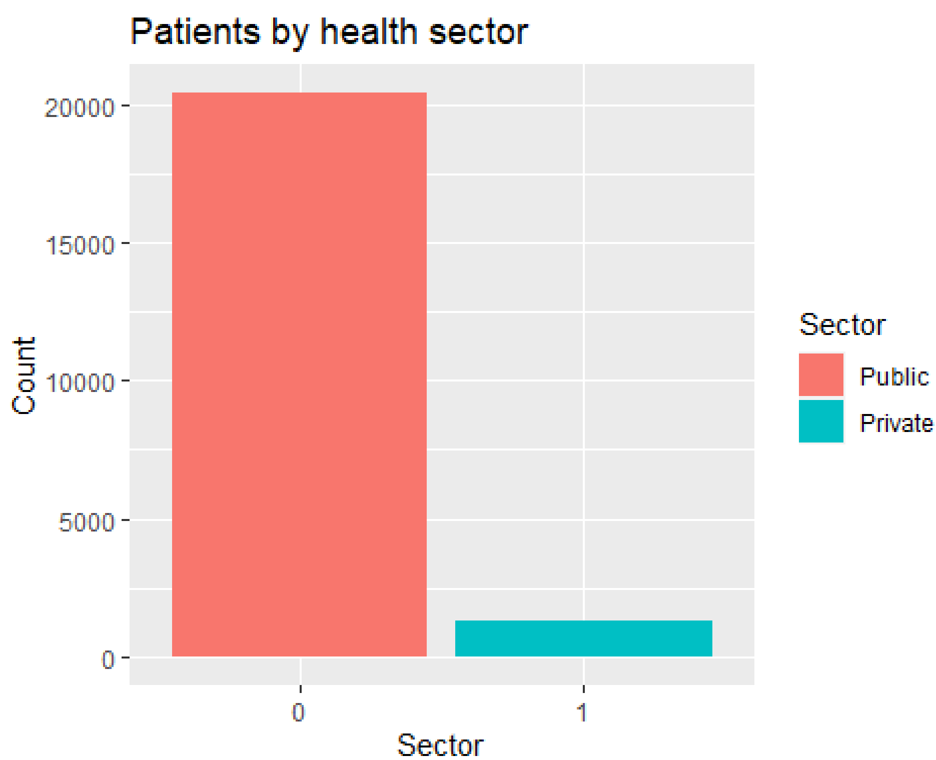
Given gender, there are slightly more men than women in the database



and most of the patients are adults.

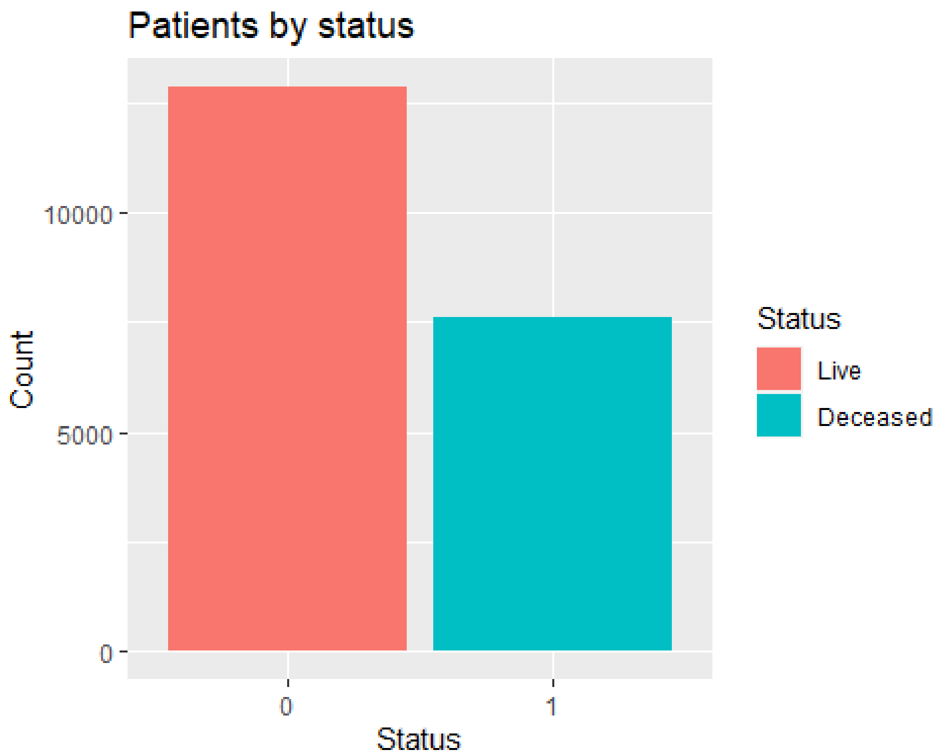


Regarding the health service sector, Mexico public health system has a wide coverage wich can be reflected by the substantially large proportion of attendance to this sector.



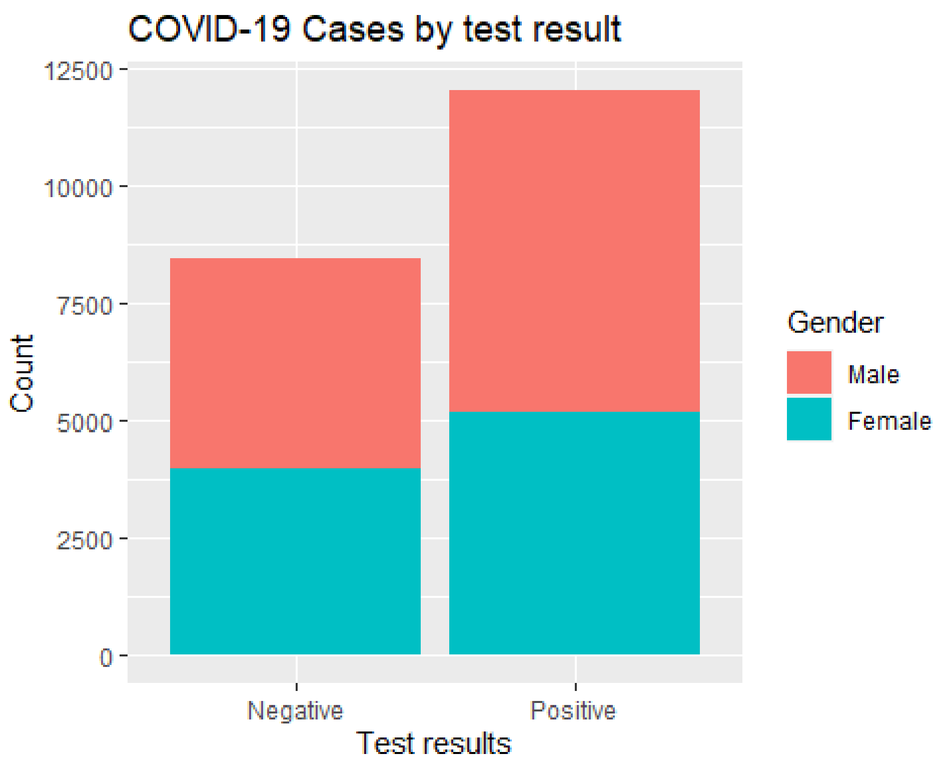
As of the positivity of COVID-19, it can be implied that most of the people that attended or requested a health service with covid-19 related symptoms did in fact have the virus.

But the outcome of the treatment was mostly favorable and the proportion of survivors is substantially large in comparison with deceased patients.



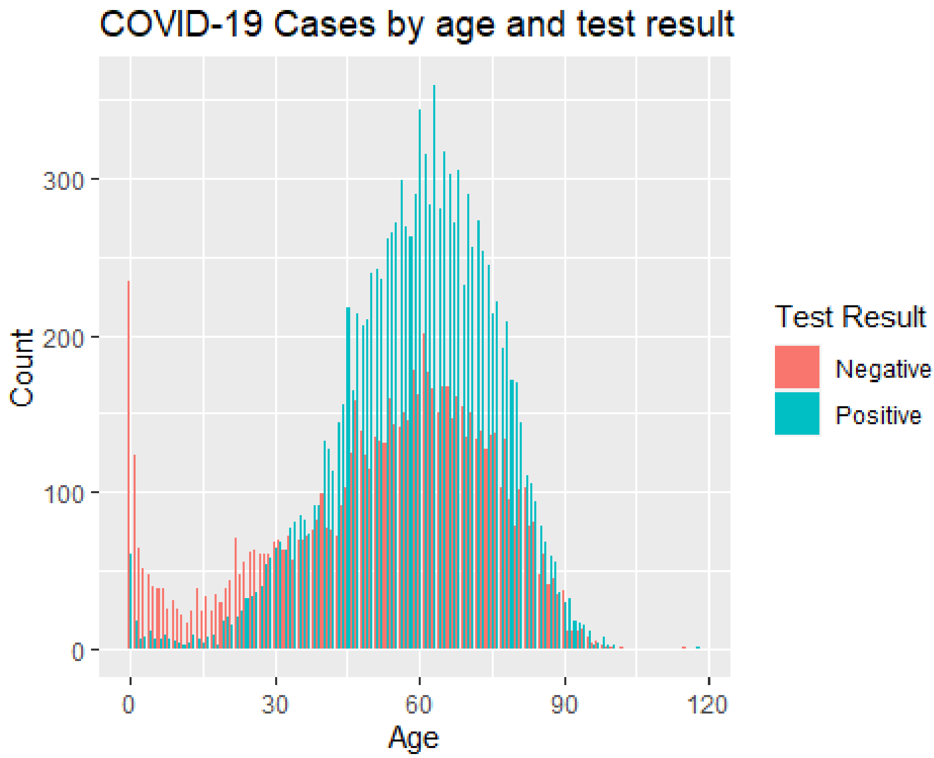
But differentiating by COVID-19 results will give us more insight regarding the disease, such as that male patients are more probable to be infected that female patients.

```
## `summarise()` has grouped output by 'sexo'. You can override using the
`.groups` argument.
```



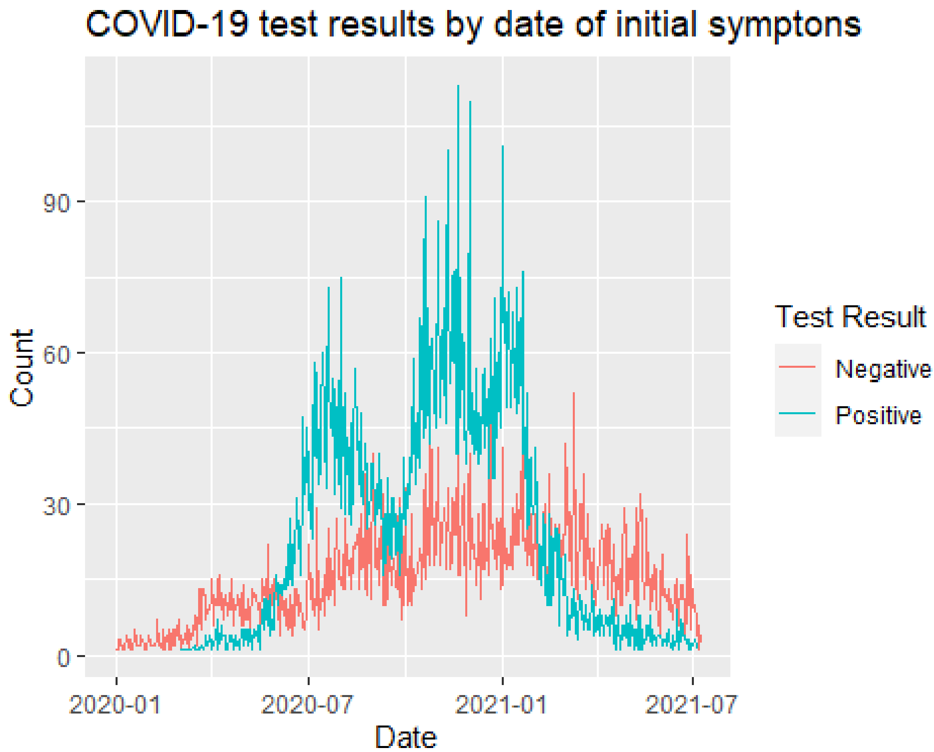
In terms of age, most of the cases, independently from the COVID-19 test result, are from adults and centered around 60 years old,

```
## `summarise()` has grouped output by 'edad'. You can override using the
`.groups` argument.
```



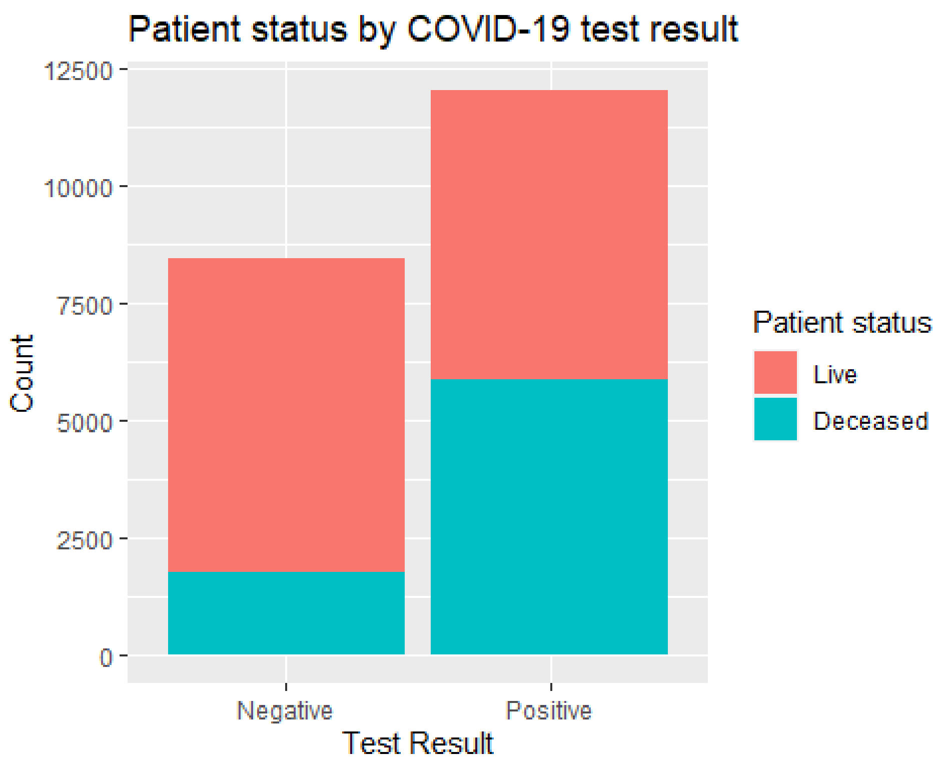
In terms of date of symptons, we can identify three peaks around july 2020 and the last quarter of the same year. It seems that the infection rate has decreased considerably in the last year, while the non-positive results remained constant over the whole time frame.

```
## `summarise()` has grouped output by 'fecha_sintomas'. You can override
using the `.groups` argument.
```



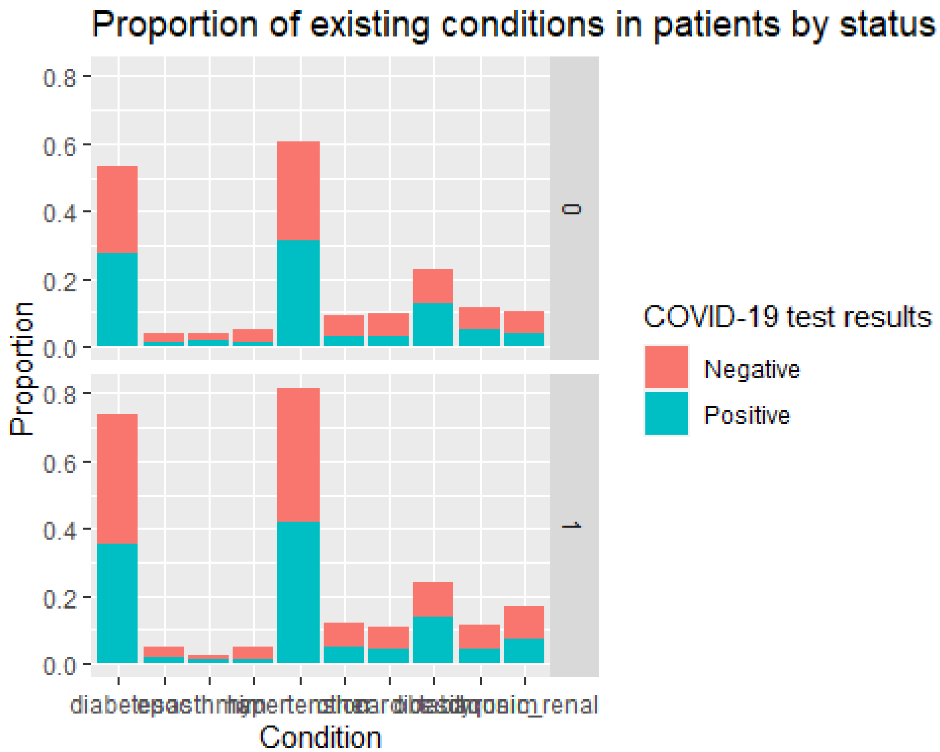
Regarding the final status, there seems to by an equal proportion of lived and deceased patients given the test results.

```
## `summarise()` has grouped output by 'clasificacion_final'. You can
override using the `.groups` argument.
```



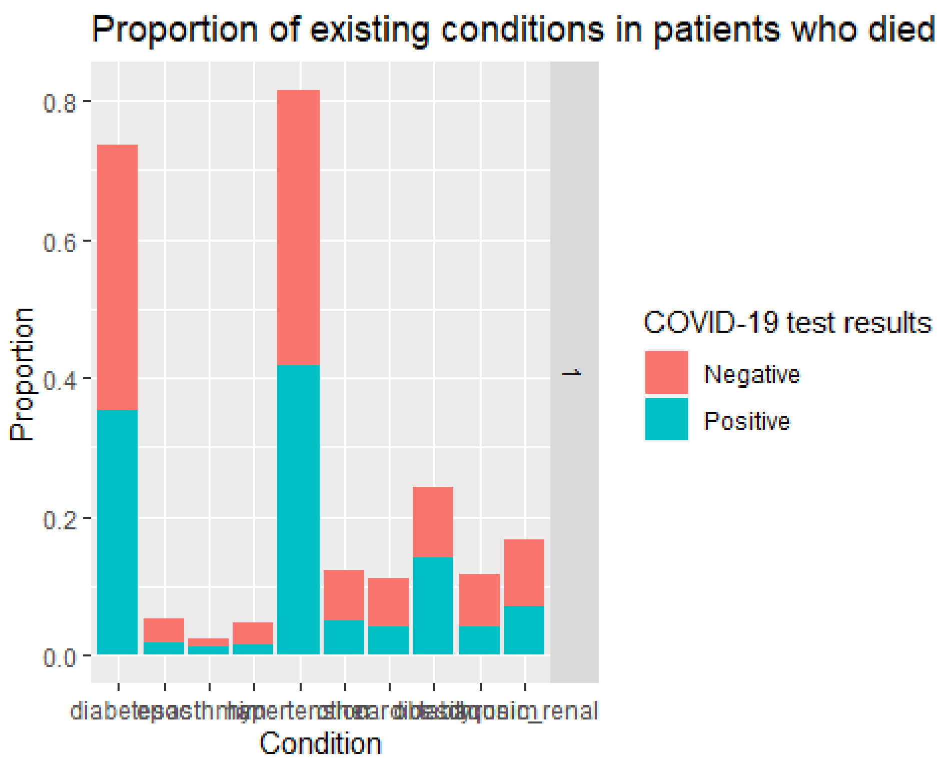
We can also check if there is an effect on previous conditions or the ones that may have developed by COVID-19, as mentioned before and confirming the relevant pre-existing conditions, we see that diabetes, hypertension and obesity are the most common.

```
## `summarise()` has grouped output by 'status'. You can override using
the `.groups` argument.
```



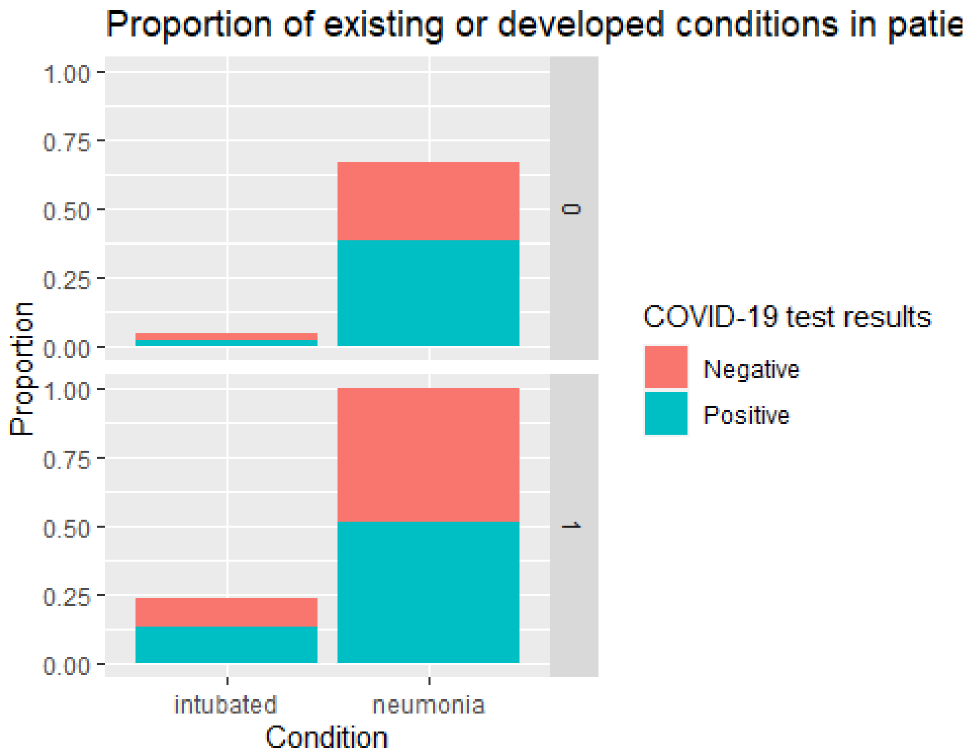
But filtering by the status of the patient, those who did not survived the disease seem to have a higher rate of the above mentioned conditions.

```
## `summarise()` has grouped output by 'status'. You can override using
the `.groups` argument.
```



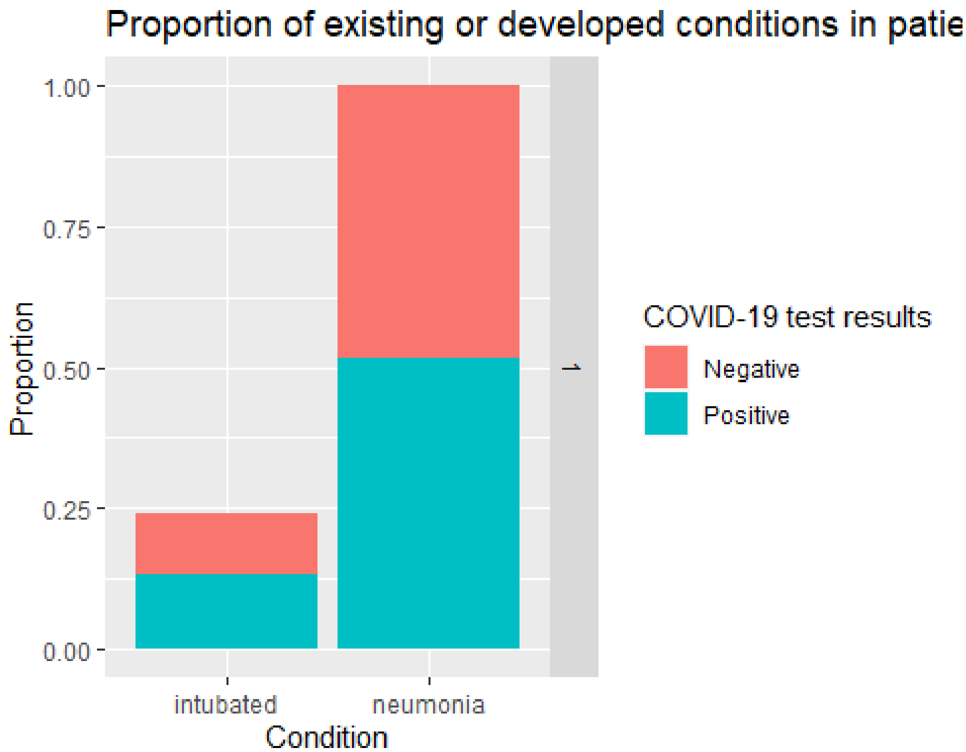
The same effect can be seen if we take into account the probably developed conditions such as neumonia or intubation.

```
## `summarise()` has grouped output by 'status'. You can override using
the `.groups` argument.
```



Where it does not seem to be much difference in the neumonia condition between both status, but the intubation increases in the patients that did not survived the disease.

```
## `summarise()` has grouped output by 'status'. You can override using
the `.groups` argument.
```



Data Modeling

To create a predictive model of not surviving the disease, we will concentrate in the patients that have a positive COVID-19 test result. In such case, we will use as reference the mean deceased rate of:

```
## [1] 0.4867874
```

and a Root mean square error of:

```
## [1] 0.4998474
```

Linear regression (Binomial)

The linear regression model gives us the possiblity of modeling the predicted value with more that just the mean. We can define it as y_hat = a + b_i + b_j + .. + b_n. In this case, for the first model all the conditions will be used, giving us a RMSE of

```
## [1] 0.4593629
```

and confirming that only a few variables are significant and the RMSE is slightly better.

```
## % latex table generated in R 4.1.0 by xtable 1.8-4 package
## % Tue Jul 27 01:18:59 2021
## \begin{table}[ht]
```

```
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & -0.0948 & 0.0229 & -4.14 & 0.0000 \\
##   sector & -0.3785 & 0.0311 & -12.17 & 0.0000 \\
##   sexo & -0.0493 & 0.0120 & -4.12 & 0.0000 \\
##   intubado & 0.4036 & 0.0221 & 18.29 & 0.0000 \\
##   neumonia & 0.1021 & 0.0120 & 8.54 & 0.0000 \\
##   edad & 0.0088 & 0.0004 & 23.61 & 0.0000 \\
##   diabetes & 0.0212 & 0.0140 & 1.52 & 0.1291 \\
##   epoc & 0.0146 & 0.0506 & 0.29 & 0.7726 \\
##   asma & -0.0413 & 0.0457 & -0.90 & 0.3666 \\
##   inmusupr & -0.0480 & 0.0472 & -1.02 & 0.3089 \\
##   hipertension & 0.0083 & 0.0139 & 0.59 & 0.5525 \\
##   otra\_com & 0.0972 & 0.0299 & 3.25 & 0.0012 \\
##   cardiovascular & -0.0453 & 0.0314 & -1.44 & 0.1499 \\
##   obesidad & 0.0238 & 0.0178 & 1.34 & 0.1806 \\
##   renal\_cronica & 0.1550 & 0.0270 & 5.74 & 0.0000 \\
##   tabaquismo & -0.0170 & 0.0301 & -0.56 & 0.5726 \\
##    \hline
## \end{tabular}
## \end{table}
```

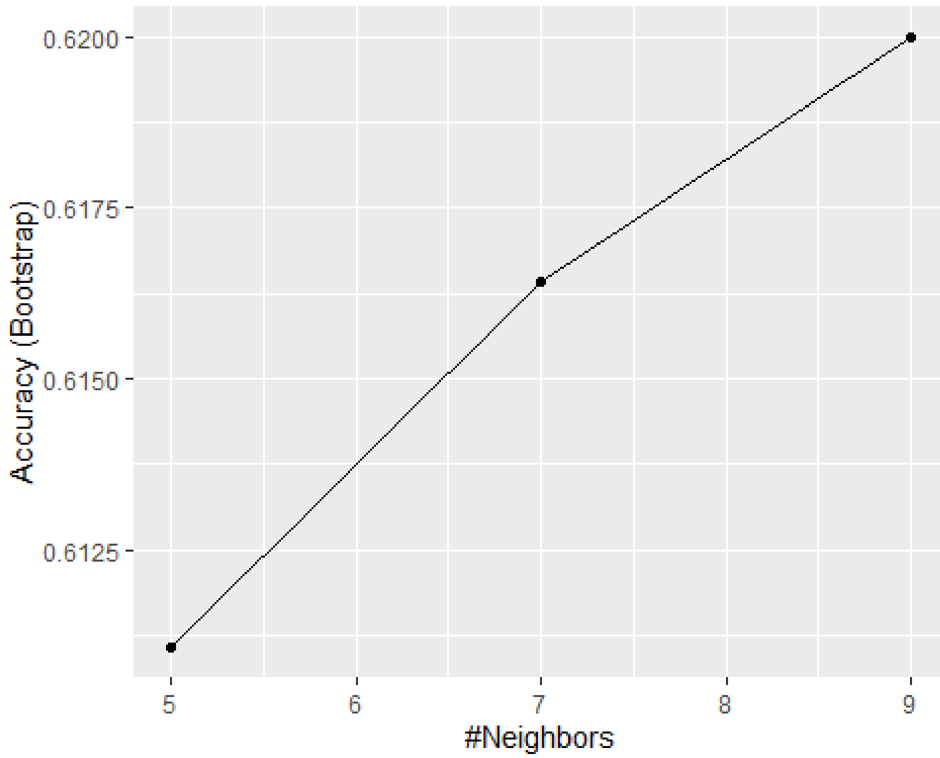
Linear regression with selected predictors

The second model uses only the relevant predictor as by their significance, Which gives us a slightly better RMSE.

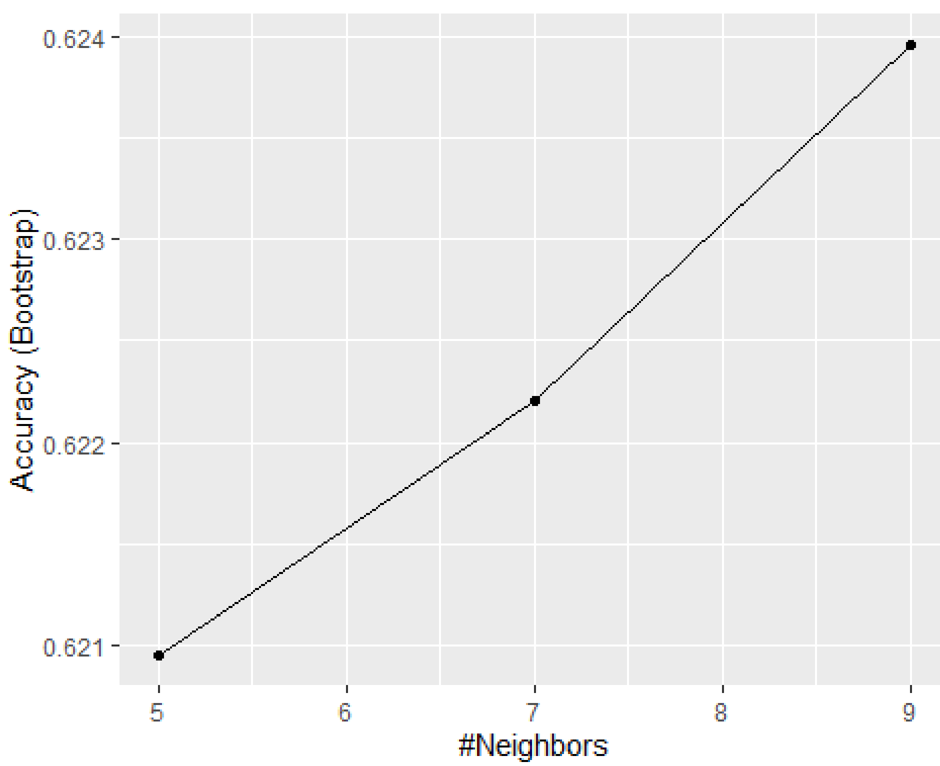
```
## [1] 0.4599529
```

Classification model

The classification model allows us to think in terms of groups, given that in theory people with certain conditions have more probability of not surviving the virus, this can be a better tool for modeling. With this model we can see that the best accuracy can be achieved at a certain level (k = 9), with a value of 0.62.



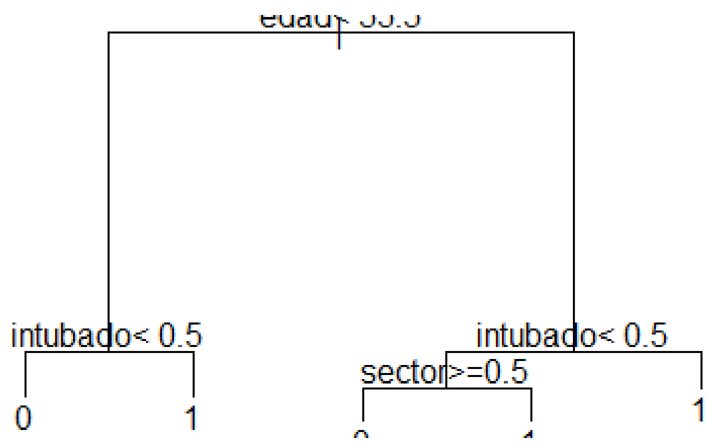
But it is not better that the linear model by means of the RMSE if we further improve the model using only the relevant variables.




```
## [1] 0.6975824
```

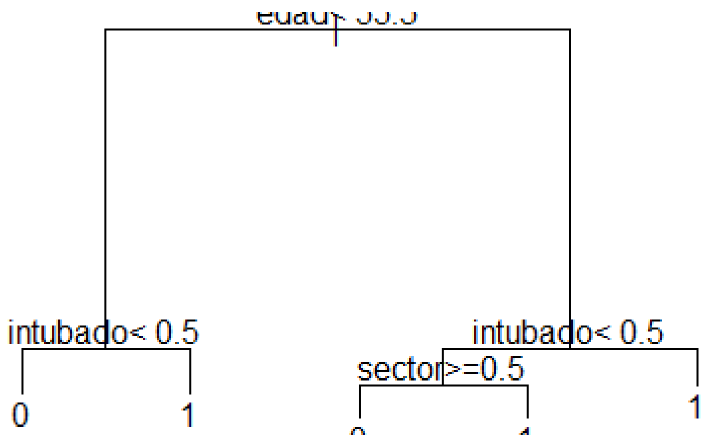
Regression Trees

This model allows us to classify the cases by cuts and have a better understanding of what is happening given the patient conditions. In this case, we have a cut at age 55.5, and then being intubated can become a probable cause of death for younger people. In the case of older people, the intubation condition is further down classfied by the sector condition, which implies that public and private service can have a significant difference in outcome.




```
## [1] 1.122529
```

#Regression trees with selected conditions If we further train our model with just the relevant variables as of the linear model, we can see that the cuts remain the same but the RMSE is not better than any of the previous models.

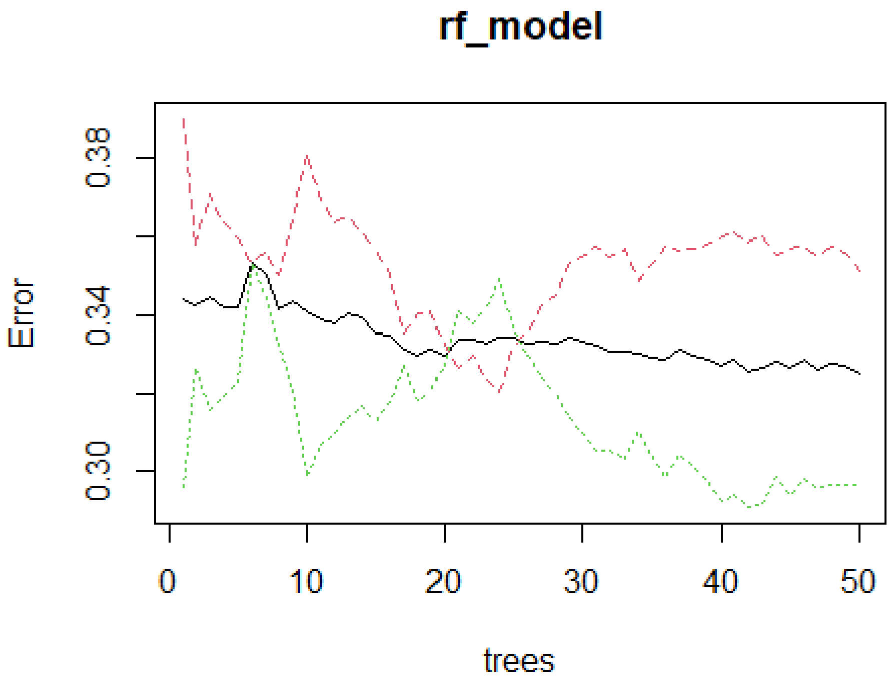



```
## [1] 1.122529
```

Random Forest

Using the random forest model, we can have a better accuracy rate, at around 32.47% error margin using the relevant variables as of the linear model.

```
##
## Call:
##  randomForest(formula = as.formula("status ~ sector + sexo + intubado
+ neumonia + edad + renal_cronica"),      data = train_set2, ntree = 50,
importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 32.47%
## Confusion matrix:
##      0    1 class.error
## 0 2003 1085   0.3513601
## 1  869 2060   0.2966883
```



and it can also tells us the relevant variables, confirming the intubated, sector and chronic renal failure conditions as the most relevant ones.

```
##                MeanDecreaseAccuracy
## sector                    11.925368
## sexo                       1.387952
## intubado                  18.907628
## neumonia                   6.246183
## edad                      16.326682
## renal_cronica             10.381874
```

Quadratic Discriminant Analysis

The Quadratic Discriminant Analysis or QDA can also give us insight on what are the relevant conditions, but by using the ones as specified by the linear model, we see that there is no better accuracy with the test set as other models. In other cases, this can be usefull to classify which conditions or group means can be usefull to predict the survival rate.

```
## [1] 0.5925974
```

Conclusion

As of the relevant conditions that can be confirmed by the data analysis and predictive models there is a substantial consideration regarding what makes a patient survive or not the COVID-19 disease. This will be further down understood as more and more research is done, but given the results of this analysis, pre-existing conditions such as chronic renal failure are relevant in survival rates. Given that neumonia can or not be a pre-existing condition, COVID-19 will only make it worse, and in addition to intubation and age, in increases considerable the probability of not surviving. This exercise gives insight on how difficult it is to predict the outcome of a recently discovered disease or virus, and how much does it plays the preexisting conditions and the still lack of understanding that exists regarding the nature and function of the complexity of the human body.