Análisis Covid-19 México

Porfirio Ángel Díaz Sánchez

3 de julio de 2020

Introducción

La aparición del Covid-19 es uno de los mayores retos en materia de salud al que el mundo se ha enfrentado en los últimos años, tanto así que, en cuestión de meses, ha ocasionado muchos cambios en la vida cotidiana de las personas y se ha esparcido por muchos países alrededor del mundo. Si bien su nivel de mortalidad no es tan alto, dicha enfermedad es muy contagiosa, por lo que los sistemas de salud corren el riesgo de sobresaturación, provocando que no sea posible atender a todos los pacientes de forma adecuada, y así mismo, aumentando los efectos negativos en la salud de las personas.

Marco teórico

El aprendizaje automático o machine learning es una ciencia enfocada en crear sistemas que aprenden automáticamente a partir de los datos. Se relaciona estrechamente con la minería de datos, que consiste en la búsqueda de información útil en grandes bases de datos. Cabe destacar que la minería de datos involucra el uso de machine learning, pero no todo el machine learning involucra minería de datos [1].

Reglas de clasificación

Este tipo de algoritmos representan el conocimiento por medio de sentencias if-else y son adecuadas para datos nominales.

Las reglas de clasificación utilizan la heurística conocida como separa y conquistarás, que consiste en encontrar una regla que cubra un subconjunto de ejemplos, y los separa de los datos restantes. Estas acciones se van repitiendo hasta que todo el dataset haya sido cubierto [1].

KNN

Es un algoritmo de clasificación de vecinos más cercanos, que clasifica registros no etiquetados asignándoles una clase de ejemplos similares, su funcionamiento consiste en buscar los k vecinos más cercanos de acuerdo a la similitud en las características de los datos.

Cuando se aplica el algoritmo para datos nominales y faltantes, es necesario llevar a cabo procesamiento adicional.

Árboles de decisión

Los árboles de decisión modelan las relaciones existentes entre las características por medio de estructuras de árbol y sus salidas, además de proporcionar una salida simple y legible para cualquier persona.

Utilizan la heurística divide y conquistarás, que consiste en dividir los datos en subconjuntos, que se siguen dividiendo repetidamente hasta que el algoritmo determina que los datos son suficientemente homogéneos.

Naive Bayes

Este algoritmo utiliza el Teorema de Bayes en problemas de clasificación, utiliza los datos de entrenamiento para el cálculo de la probabilidad de cada salida.

Se aplica a problemas donde la información de muchos atributos debe ser considerada de manera simultánea para estimar la probabilidad de una salida [1].

Regresión Lineal

Los métodos de regresión se utilizan para el tratamiento de datos numéricos, utilizan una variable dependiente (valor a predecir), y un conjunto de variables independientes (los predictores).

El uso de datos categóricos requiere procesamiento adicional.

Obtención de los datos

Por medio del sitio oficial de Datos Abiertos del Gobierno de la República, el gobierno mexicano pone al alcance de los ciudadanos diversos catálogos de datos digitales acerca de diversos temas de interés nacional para su libre consulta y análisis.

La base de datos utilizada en el presente trabajo registra aspectos de salud, resultados de pruebas, evolución de enfermedad, entre otras características, de los pacientes que presentaron síntomas de Covid-19 en México. Estos datos se utilizan con el objetivo de determinar los aspectos que influyen en que un paciente con síntomas de Covid-19, muera.

```
# Lee el dataset.
data <-
  read.csv('data/200630COVID19MEXICO.csv', stringsAsFactors = TRUE)
```

Descripción de la base de datos

ORIGEN: Identifica si la unidad de atención se encuentra dentro del sistema de Unidades de Salud Monitoras de Enfermedades Respiratorias.

SECTOR: Identifica el tipo de institución del Sistema Nacional de Salud que brindó la atención.

SEXO: Identifica al sexo del paciente.

TIPO_PACIENTE: Identifica si el paciente regresó a su casa o si fue hospitalizado.

INTUBADO: Identifica si el paciente requirió intubación.

NEUMONIA: Identifica si al paciente se le diagnosticó con neumonía.

EDAD: Identifica la edad del paciente.

EMBARAZO: Identifica si la paciente está embarazada.

DIABETES: Identifica si el paciente tiene un diagnóstico de diabetes.

EPOC: Identifica si el paciente tiene un diagnóstico de EPOC.

ASMA: Identifica si el paciente tiene un diagnóstico de asma.

INMUSUPR: Identifica si el paciente presenta inmunosupresión.

HIPERTENSION: Identifica si el paciente tiene un diagnóstico de hipertensión.

OTRA_COM: Identifica si el paciente tiene diagnóstico de otras enfermedades.

CARDIOVASCULAR: Identifica si el paciente tiene diagnóstico de enfermedades cardiovasculares.

OBESIDAD: Identifica si el paciente tiene diagnóstico de obesidad.

RENAL_CRONICA: Identifica si el paciente tiene diagnóstico de insuficiencia renal crónica.

TABAQUISMO: Identifica si el paciente tiene hábito de tabaquismo.

OTRO_CASO: Identifica si el paciente tuvo contacto con algún otro caso diagnósticado con SARS CoV-2

RESULTADO: Identifica el resultado de la prueba de Covid-19.

UCI: Identifica si el paciente ingresó a una Unidad de Cuidados Intensivos.

MURIO: Identifica si el paciente murió.

DIAS_INGRESO: Días transcurridos desde que el paciente presentó síntomas, hasta que fue atendido por la unidad médica.

DIAS_ENFERMEDAD: Días transcurridos desde que el paciente presentó síntomas, hasta que murió.

DIAS_HOSPITALIZACION: Días transcurridos desde que el paciente fue atendido por la unidad médica, hasta que murió

Metodología

- Búsqueda del dataset.
- Previsualización de los datos.
- Selección y extracción de características.
- Aplicación de algoritmos de Machine Learning.
- Interpretación de resultados.

Exploración y preparación de los datos

Una vez importado el dataset, se procede a previsualizar su contenido, con la finalidad de ver cómo está compuesto el mismo y qué tipo de información ofrece:

```
head(data)
```

```
##   FECHA_ACTUALIZACION ID_REGISTRO ORIGEN SECTOR ENTIDAD_UM SEXO ENTIDAD_NAC
## 1          2020-06-30      04f3dd      2      3         25    2          25
## 2          2020-06-30      1b7c4b      2      3         27    1          27
## 3          2020-06-30      03f6dd      2      4          9    1           9
## 4          2020-06-30      187fc7      2      4         15    2          15
## 5          2020-06-30      1795a0      2      3          2    2           2
## 6          2020-06-30      172400      2      4         21    1          21
##   ENTIDAD_RES MUNICIPIO_RES TIPO_PACIENTE FECHA_INGRESO FECHA_SINTOMAS
## 1          25             6             1    2020-05-11     2020-05-09
## 2          27             5             2    2020-05-22     2020-05-20
## 3          15            58             1    2020-04-17     2020-04-14
## 4          15           122             2    2020-04-21     2020-04-21
## 5           2             2             1    2020-06-01     2020-06-01
## 6          21           114             1    2020-03-31     2020-03-28
##    FECHA_DEF INTUBADO NEUMONIA EDAD NACIONALIDAD EMBARAZO HABLA_LENGUA_INDIG
## 1 9999-99-99       97        2   27            1       97                 2
## 2 9999-99-99        2        2   52            1        2                 2
```

```
## 3 9999-99-99       97        2   55            1        2                  2
## 4 9999-99-99        2        2   59            1       97                  2
## 5 9999-99-99       97        2   33            1       97                  2
## 6 9999-99-99       97        2   44            1        2                  2
##   DIABETES EPOC ASMA INMUSUPR HIPERTENSION OTRA_COM CARDIOVASCULAR OBESIDAD
## 1        2    2    2        2            2        2              2        2
## 2        1    2    2        2            1        2              1        2
## 3        1    2    2        2            1        2              2        1
## 4        2    2    2        2            2        2              2        2
## 5        2    2    2        2            2        2              2        2
## 6        1    2    2        2            2        2              2        2
##   RENAL_CRONICA TABAQUISMO OTRO_CASO RESULTADO MIGRANTE PAIS_NACIONALIDAD
## 1             2          2         1         1       99            México
## 2             2          2         2         1       99            México
## 3             2          2        99         1       99            México
## 4             2          2        99         1       99            México
## 5             2          2         1         1       99            México
## 6             2          2        99         1       99            México
##   PAIS_ORIGEN UCI
## 1          99  97
## 2          99   2
## 3          99  97
## 4          99   2
## 5          99  97
## 6          99  97
```

```
summary(data)
```

```
##  FECHA_ACTUALIZACION  ID_REGISTRO         ORIGEN          SECTOR
##  2020-06-30:581580   000002 :     1   Min.   :1.000   Min.   : 1.00
##                      000008 :     1   1st Qu.:1.000   1st Qu.: 4.00
##                      00000e :     1   Median :2.000   Median :12.00
##                      000013 :     1   Mean   :1.644   Mean   : 9.73
##                      000015 :     1   3rd Qu.:2.000   3rd Qu.:12.00
##                      000019 :     1   Max.   :2.000   Max.   :99.00
##                      (Other):581574
##    ENTIDAD_UM         SEXO        ENTIDAD_NAC     ENTIDAD_RES
##  Min.   : 1.00   Min.   :1.000   Min.   : 1.00   Min.   : 1.00
##  1st Qu.: 9.00   1st Qu.:1.000   1st Qu.: 9.00   1st Qu.: 9.00
##  Median :14.00   Median :2.000   Median :15.00   Median :15.00
##  Mean   :15.35   Mean   :1.506   Mean   :16.31   Mean   :15.62
##  3rd Qu.:21.00   3rd Qu.:2.000   3rd Qu.:22.00   3rd Qu.:21.00
##  Max.   :32.00   Max.   :2.000   Max.   :99.00   Max.   :32.00
##
##  MUNICIPIO_RES    TIPO_PACIENTE      FECHA_INGRESO       FECHA_SINTOMAS
##  Min.   :  1.00   Min.   :1.000   2020-06-15: 13330   2020-06-01: 13883
##  1st Qu.:  7.00   1st Qu.:1.000   2020-06-16: 12839   2020-06-15: 13855
##  Median : 21.00   Median :1.000   2020-06-23: 12662   2020-06-10: 12920
##  Mean   : 38.71   Mean   :1.214   2020-06-26: 12544   2020-06-20: 12770
##  3rd Qu.: 50.00   3rd Qu.:1.000   2020-06-22: 12493   2020-06-08: 11193
##  Max.   :999.00   Max.   :2.000   2020-06-29: 12392   2020-05-25: 10807
##                                   (Other)   :505320   (Other)   :506152
##       FECHA_DEF         INTUBADO        NEUMONIA           EDAD
##  9999-99-99:544456   Min.   : 1.00   Min.   : 1.000   Min.   :  0.00
##  2020-06-16:   709   1st Qu.:97.00   1st Qu.: 2.000   1st Qu.: 31.00
```

```
##  2020-06-08:   706   Median :97.00   Median : 2.000   Median : 41.00
##  2020-06-10:   697   Mean   :76.68   Mean   : 1.847   Mean   : 42.62
##  2020-06-12:   679   3rd Qu.:97.00   3rd Qu.: 2.000   3rd Qu.: 53.00
##  2020-06-09:   662   Max.   :99.00   Max.   :99.000   Max.   :120.00
##  (Other)   : 33671
##   NACIONALIDAD      EMBARAZO     HABLA_LENGUA_INDIG    DIABETES
##  Min.   :1.000   Min.   : 1.00   Min.   : 1.000     Min.   : 1.000
##  1st Qu.:1.000   1st Qu.: 2.00   1st Qu.: 2.000     1st Qu.: 2.000
##  Median :1.000   Median :97.00   Median : 2.000     Median : 2.000
##  Mean   :1.006   Mean   :50.37   Mean   : 5.151     Mean   : 2.209
##  3rd Qu.:1.000   3rd Qu.:97.00   3rd Qu.: 2.000     3rd Qu.: 2.000
##  Max.   :2.000   Max.   :98.00   Max.   :99.000     Max.   :98.000
##
##       EPOC            ASMA           INMUSUPR       HIPERTENSION
##  Min.   : 1.00   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.00   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 2.00   Median : 2.000   Median : 2.000   Median : 2.000
##  Mean   : 2.28   Mean   : 2.264   Mean   : 2.319   Mean   : 2.145
##  3rd Qu.: 2.00   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :98.00   Max.   :98.000   Max.   :98.000   Max.   :98.000
##
##     OTRA_COM      CARDIOVASCULAR      OBESIDAD      RENAL_CRONICA
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median : 2.000
##  Mean   : 2.408   Mean   : 2.285   Mean   : 2.137   Mean   : 2.282
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :98.000
##
##    TABAQUISMO       OTRO_CASO       RESULTADO        MIGRANTE
##  Min.   : 1.000   Min.   : 1.00   Min.   :1.000   Min.   : 1.0
##  1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.:1.000   1st Qu.:99.0
##  Median : 2.000   Median : 2.00   Median :2.000   Median :99.0
##  Mean   : 2.237   Mean   :31.54   Mean   :1.735   Mean   :98.6
##  3rd Qu.: 2.000   3rd Qu.:99.00   3rd Qu.:2.000   3rd Qu.:99.0
##  Max.   :98.000   Max.   :99.00   Max.   :3.000   Max.   :99.0
##
##                  PAIS_NACIONALIDAD                         PAIS_ORIGEN
##  México                    :578290   99                          :580849
##  Estados Unidos de América:   807   Rep\xfablica de Honduras    :   123
##  Colombia                  :   302   Estados Unidos de Am\xe9rica:   110
##  Cuba                      :   249   Colombia                    :    69
##  Venezuela                 :   239   Venezuela                   :    66
##  República de Honduras     :   171   Cuba                        :    61
##  (Other)                   :  1522   (Other)                     :   302
##       UCI
##  Min.   : 1.00
##  1st Qu.:97.00
##  Median :97.00
##  Mean   :76.68
##  3rd Qu.:97.00
##  Max.   :99.00
##
```

Como puede observarse en los resultados de los comandos anteriores, el dataset se compone básicamente de datos categóricos que determinan padecimientos de salud, resultados de la prueba, datos de residencia, etc. Las características categóricas que se incluyen ya vienen codificadas numéricamente, pero aún así, se convierten a factor porque más allá de ver un resumen de sus medidas de tendencia central, es de mayor interés verlo en términos de las apariciones de cada una de sus clases:

```
data$ORIGEN <- as.factor(data$ORIGEN)
data$SECTOR <- as.factor(data$SECTOR)
data$ENTIDAD_UM <- as.factor(data$ENTIDAD_UM)
data$SEXO <- as.factor(data$SEXO)
data$ENTIDAD_NAC <- as.factor(data$ENTIDAD_NAC)
data$ENTIDAD_RES <- as.factor(data$ENTIDAD_RES)
data$MUNICIPIO_RES <- as.factor(data$MUNICIPIO_RES)
data$TIPO_PACIENTE <- as.factor(data$TIPO_PACIENTE)
data$INTUBADO <- as.factor(data$INTUBADO)
data$NEUMONIA <- as.factor(data$NEUMONIA)
data$NACIONALIDAD <- as.factor(data$NACIONALIDAD)
data$EMBARAZO <- as.factor(data$EMBARAZO)
data$HABLA_LENGUA_INDIG <- as.factor(data$HABLA_LENGUA_INDIG)
data$DIABETES <- as.factor(data$DIABETES)
data$EPOC <- as.factor(data$EPOC)
data$ASMA <- as.factor(data$ASMA)
data$INMUSUPR <- as.factor(data$INMUSUPR)
data$HIPERTENSION <- as.factor(data$HIPERTENSION)
data$OTRA_COM <- as.factor(data$OTRA_COM)
data$CARDIOVASCULAR <- as.factor(data$CARDIOVASCULAR)
data$OBESIDAD <- as.factor(data$OBESIDAD)
data$RENAL_CRONICA <- as.factor(data$RENAL_CRONICA)
data$TABAQUISMO <- as.factor(data$TABAQUISMO)
data$OTRO_CASO <- as.factor(data$OTRO_CASO)
data$RESULTADO <- as.factor(data$RESULTADO)
data$MIGRANTE <- as.factor(data$MIGRANTE)
data$UCI <- as.factor(data$UCI)
```

```
str(data)
```

```
## 'data.frame':    581580 obs. of  35 variables:
##  $ FECHA_ACTUALIZACION: Factor w/ 1 level "2020-06-30": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ID_REGISTRO        : Factor w/ 581580 levels "000002","000008",..: 94493 523675 75725 466842 44913
##  $ ORIGEN             : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...
##  $ SECTOR             : Factor w/ 13 levels "1","2","3","4",..: 3 3 4 4 3 4 3 3 4 4 ...
##  $ ENTIDAD_UM         : Factor w/ 32 levels "1","2","3","4",..: 25 27 9 15 2 21 27 8 2 2 ...
##  $ SEXO               : Factor w/ 2 levels "1","2": 2 1 1 2 2 1 1 2 2 2 ...
##  $ ENTIDAD_NAC        : Factor w/ 33 levels "1","2","3","4",..: 25 27 9 15 2 21 27 8 25 7 ...
##  $ ENTIDAD_RES        : Factor w/ 32 levels "1","2","3","4",..: 25 27 15 15 2 21 27 8 2 2 ...
##  $ MUNICIPIO_RES      : Factor w/ 409 levels "1","2","3","4",..: 6 5 58 122 2 114 4 37 2 4 ...
##  $ TIPO_PACIENTE      : Factor w/ 2 levels "1","2": 1 2 1 2 1 1 2 1 2 2 ...
##  $ FECHA_INGRESO      : Factor w/ 182 levels "2020-01-01","2020-01-02",..: 132 143 108 112 153 91 145
##  $ FECHA_SINTOMAS     : Factor w/ 182 levels "2020-01-01","2020-01-02",..: 130 141 105 112 153 88 139
##  $ FECHA_DEF          : Factor w/ 125 levels "2020-01-13","2020-01-14",..: 125 125 125 125 125 125 88
##  $ INTUBADO           : Factor w/ 4 levels "1","2","97","99": 3 2 3 2 3 3 2 3 2 2 ...
##  $ NEUMONIA           : Factor w/ 3 levels "1","2","99": 2 2 2 2 2 2 1 1 2 1 ...
##  $ EDAD               : int  27 52 55 59 33 44 68 48 52 58 ...
##  $ NACIONALIDAD       : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EMBARAZO           : Factor w/ 4 levels "1","2","97","98": 3 2 2 3 3 2 2 3 3 3 ...
```

```
##  $ HABLA_LENGUA_INDIG : Factor w/ 3 levels "1","2","99": 2 2 2 2 2 2 2 2 2 2 ...
##  $ DIABETES           : Factor w/ 3 levels "1","2","98": 2 1 1 2 2 1 1 2 1 1 ...
##  $ EPOC               : Factor w/ 3 levels "1","2","98": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ASMA               : Factor w/ 3 levels "1","2","98": 2 2 2 2 2 2 2 2 2 2 ...
##  $ INMUSUPR           : Factor w/ 3 levels "1","2","98": 2 2 2 2 2 2 2 2 2 2 ...
##  $ HIPERTENSION       : Factor w/ 3 levels "1","2","98": 2 1 1 2 2 2 1 2 1 1 ...
##  $ OTRA_COM           : Factor w/ 3 levels "1","2","98": 2 2 2 2 2 2 2 2 2 2 ...
##  $ CARDIOVASCULAR     : Factor w/ 3 levels "1","2","98": 2 1 2 2 2 2 2 2 2 2 ...
##  $ OBESIDAD           : Factor w/ 3 levels "1","2","98": 2 2 1 2 2 2 2 2 1 2 ...
##  $ RENAL_CRONICA      : Factor w/ 3 levels "1","2","98": 2 2 2 2 2 2 2 2 2 2 ...
##  $ TABAQUISMO         : Factor w/ 3 levels "1","2","98": 2 2 2 2 2 2 2 2 2 2 ...
##  $ OTRO_CASO          : Factor w/ 3 levels "1","2","99": 1 2 3 3 1 3 3 2 3 3 ...
##  $ RESULTADO          : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MIGRANTE           : Factor w/ 3 levels "1","2","99": 3 3 3 3 3 3 3 3 3 3 ...
##  $ PAIS_NACIONALIDAD  : Factor w/ 99 levels "Alemania","Archipiélago de Svalbard",..: 61 61 61 61 61
##  $ PAIS_ORIGEN        : Factor w/ 56 levels "99","Alemania",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ UCI                : Factor w/ 4 levels "1","2","97","99": 3 2 3 2 3 3 2 3 2 2 ...
```

```
summary(data)
```

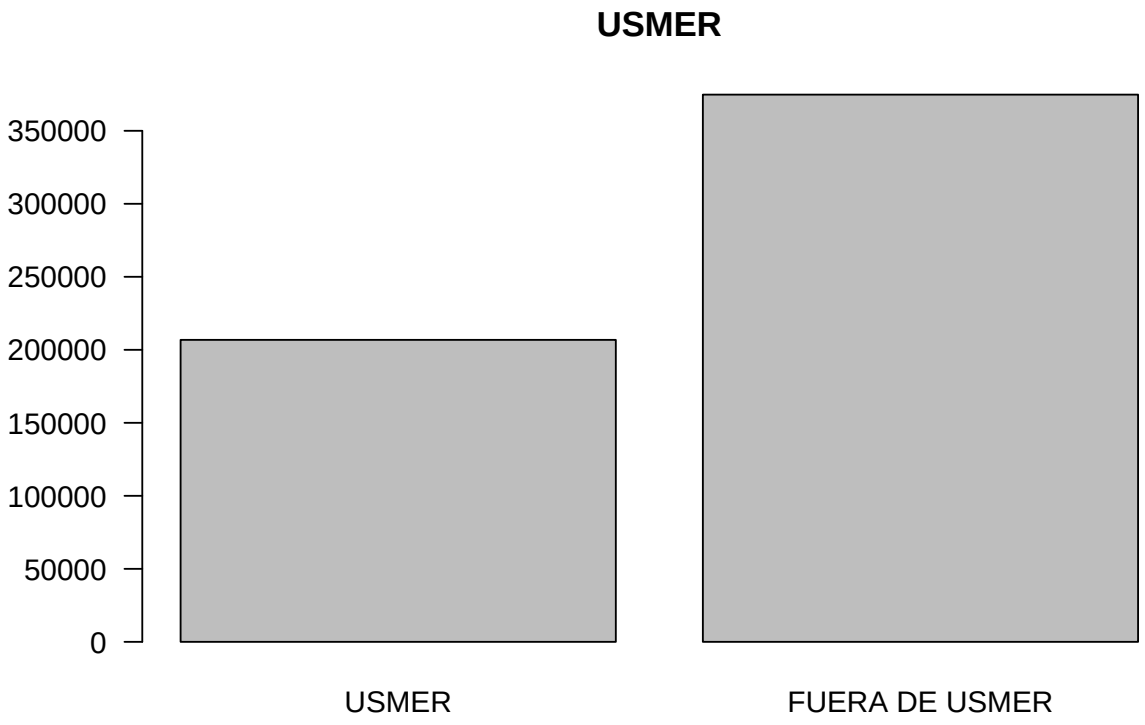
```
##  FECHA_ACTUALIZACION  ID_REGISTRO     ORIGEN         SECTOR
##  2020-06-30:581580   000002 :     1   1:206811   12     :344924
##                      000008 :     1   2:374769   4      :161488
##                      00000e :     1              6      : 22575
##                      000013 :     1              9      : 22260
##                      000015 :     1              3      : 12226
##                      000019 :     1              8      :  5629
##                      (Other):581574              (Other): 12478
##    ENTIDAD_UM     SEXO        ENTIDAD_NAC      ENTIDAD_RES     MUNICIPIO_RES
##  9      :142935   1:287056   9      :128407   9      :118208   7      : 28281
##  15     : 57938   2:294524   15     : 68591   15     : 81422   5      : 24980
##  19     : 30106              11     : 26078   19     : 29807   2      : 22461
##  11     : 27355              30     : 25884   11     : 27301   4      : 22054
##  14     : 25433              21     : 25559   14     : 25252   39     : 20973
##  21     : 25313              14     : 24561   21     : 25172   6      : 18474
##  (Other):272500              (Other):282500   (Other):274418   (Other):444357
##  TIPO_PACIENTE    FECHA_INGRESO       FECHA_SINTOMAS        FECHA_DEF
##  1:457186      2020-06-15: 13330   2020-06-01: 13883   9999-99-99:544456
##  2:124394      2020-06-16: 12839   2020-06-15: 13855   2020-06-16:   709
##                2020-06-23: 12662   2020-06-10: 12920   2020-06-08:   706
##                2020-06-26: 12544   2020-06-20: 12770   2020-06-10:   697
##                2020-06-22: 12493   2020-06-08: 11193   2020-06-12:   679
##                2020-06-29: 12392   2020-05-25: 10807   2020-06-09:   662
##                (Other)   :505320   (Other)   :506152   (Other)   : 33671
##  INTUBADO    NEUMONIA         EDAD        NACIONALIDAD EMBARAZO
##  1 : 10241   1 : 89914   Min.   :  0.00   1:578289     1 :  4170
##  2 :114031   2 :491655   1st Qu.: 31.00   2:  3291     2 :281242
##  97:457186   99:    11   Median : 41.00                97:294524
##  99:   122               Mean   : 42.62                98:  1644
##                          3rd Qu.: 53.00
##                          Max.   :120.00
##
##  HABLA_LENGUA_INDIG DIABETES    EPOC        ASMA        INMUSUPR
##  1 :  5563          1 : 72626   1 :  9304   1 : 18429   1 :  9134
##  2 :557070          2 :506929   2 :570481   2 :561358   2 :570416
```

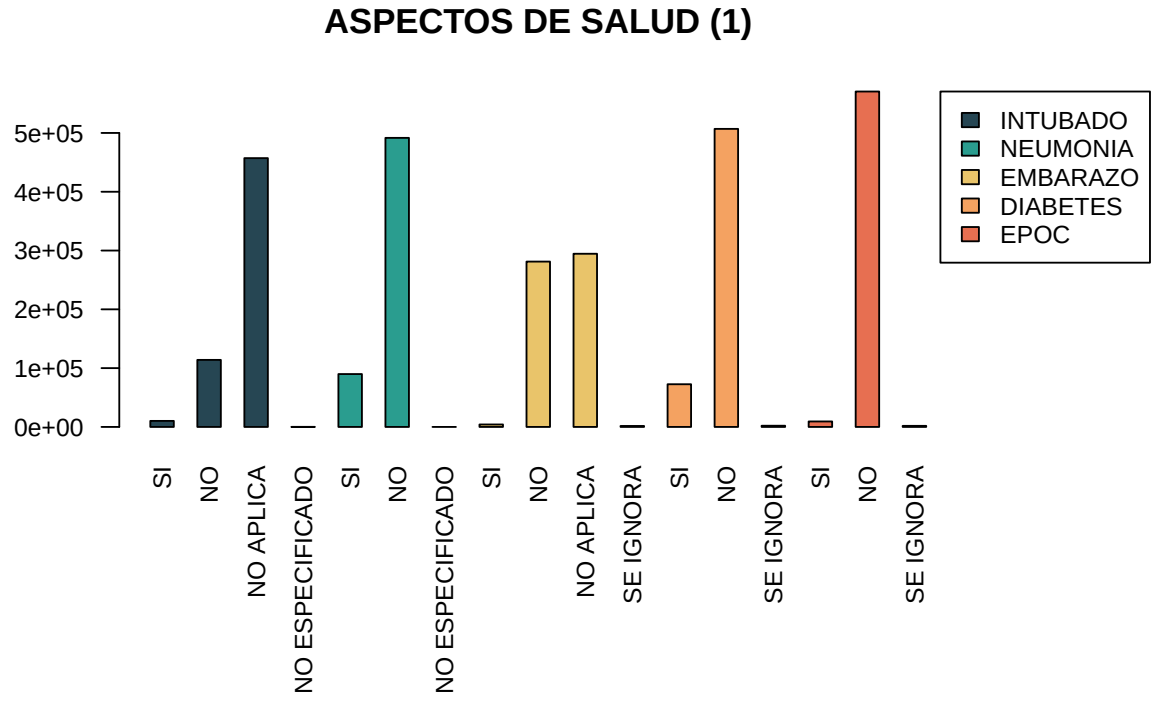
```
##  99: 18947          98:  2025   98:  1795   98:  1793   98:  2030
##
##
##
##
##  HIPERTENSION OTRA_COM    CARDIOVASCULAR OBESIDAD    RENAL_CRONICA TABAQUISMO
##  1 : 94914    1 : 17421   1 : 13048      1 : 94615   1 : 11551     1 : 49110
##  2 :484799    2 :561504   2 :566668      2 :485147   2 :568198     2 :530523
##  98:  1867    98:  2655   98:  1864      98:  1818   98:  1831     98:  1947
##
##
##
##
##  OTRO_CASO   RESULTADO  MIGRANTE                    PAIS_NACIONALIDAD
##  1 :227917   1:226089   1 :   731   México                   :578290
##  2 :174179   2:283450   2 :  1635   Estados Unidos de América:   807
##  99:179484   3: 72041   99:579214   Colombia                 :   302
##                                     Cuba                     :   249
##                                     Venezuela                :   239
##                                     República de Honduras    :   171
##                                     (Other)                  :  1522
##                        PAIS_ORIGEN     UCI
##  99                          :580849   1 : 10313
##  Rep\xfablica de Honduras    :   123   2 :113958
##  Estados Unidos de Am\xe9rica:   110   97:457186
##  Colombia                    :    69   99:   123
##  Venezuela                   :    66
##  Cuba                        :    61
##  (Other)                     :   302
```

```
nrow(data)
```

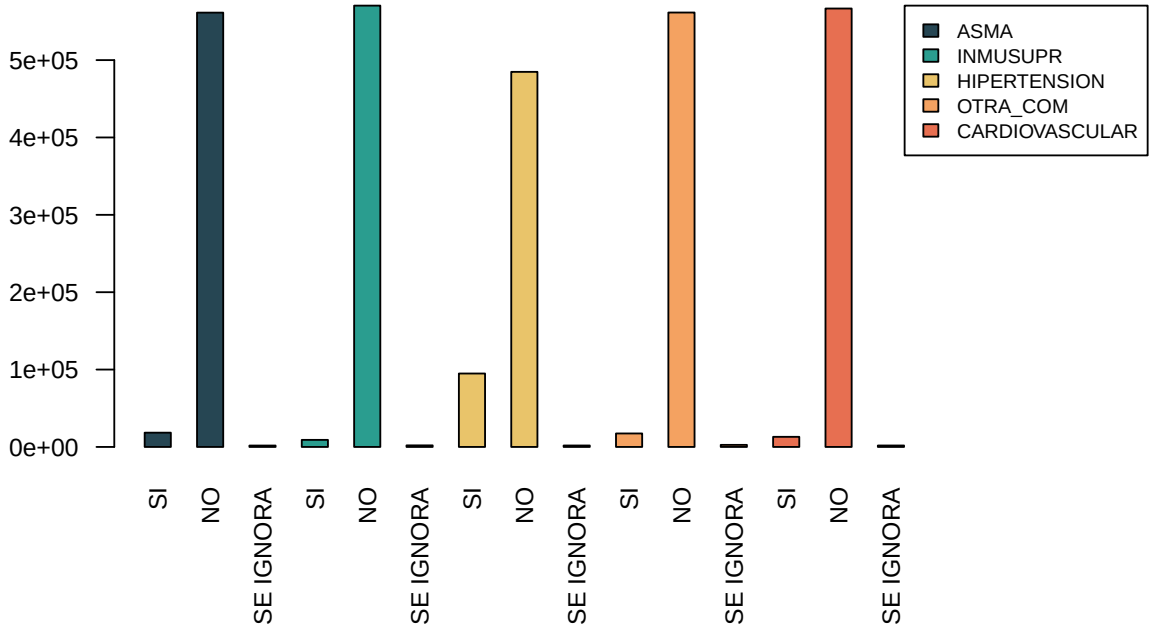
```
## [1] 581580
```

El conjunto de datos contiene 581580 registros, se tiene el conocimiento de la distribución de las clases categóricas y de las fechas donde fueron ocurriendo casos, ingresos a unidad médica y defunciones. En los siguientes apartados se visualiza de manera gráfica algunas de las características.

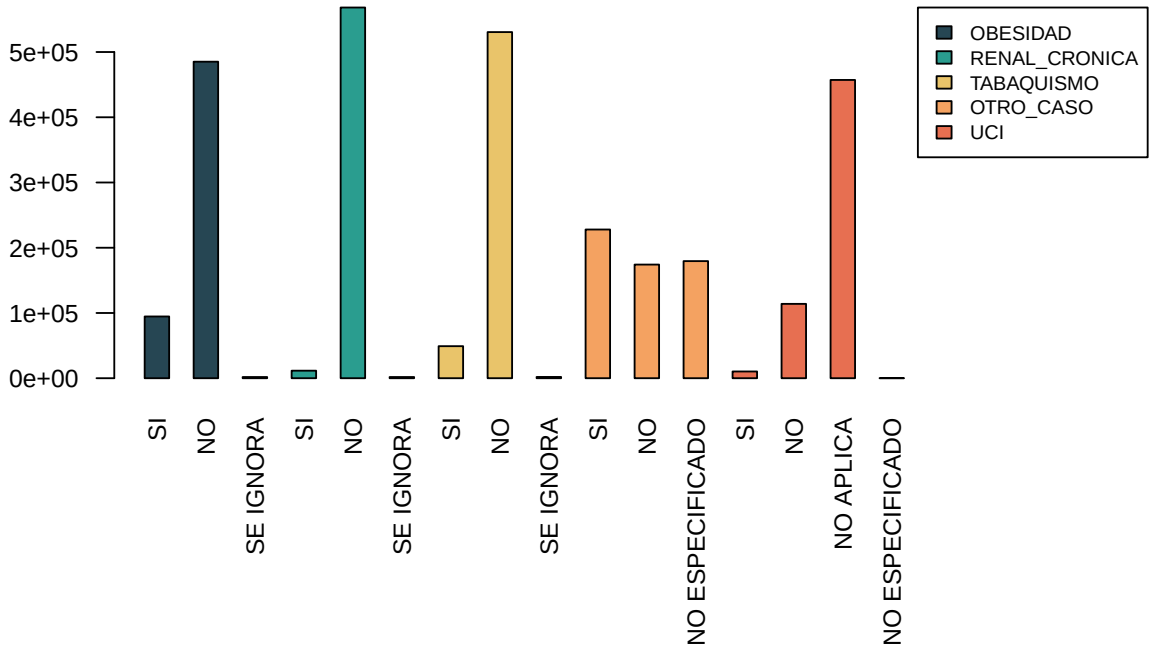

USMER
350000
300000
250000
200000
150000
100000
50000
0
USMER
FUERA DE USMER


ASPECTOS DE SALUD (1)
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
INTUBADO
NEUMONIA
EMBARAZO
DIABETES
EPOC
SI
NO
NO APLICA
NO ESPECIFICADO
SI
NO
NO ESPECIFICADO
SI
NO
NO APLICA
SE IGNORA
SI
NO
SE IGNORA
SI
NO
SE IGNORA

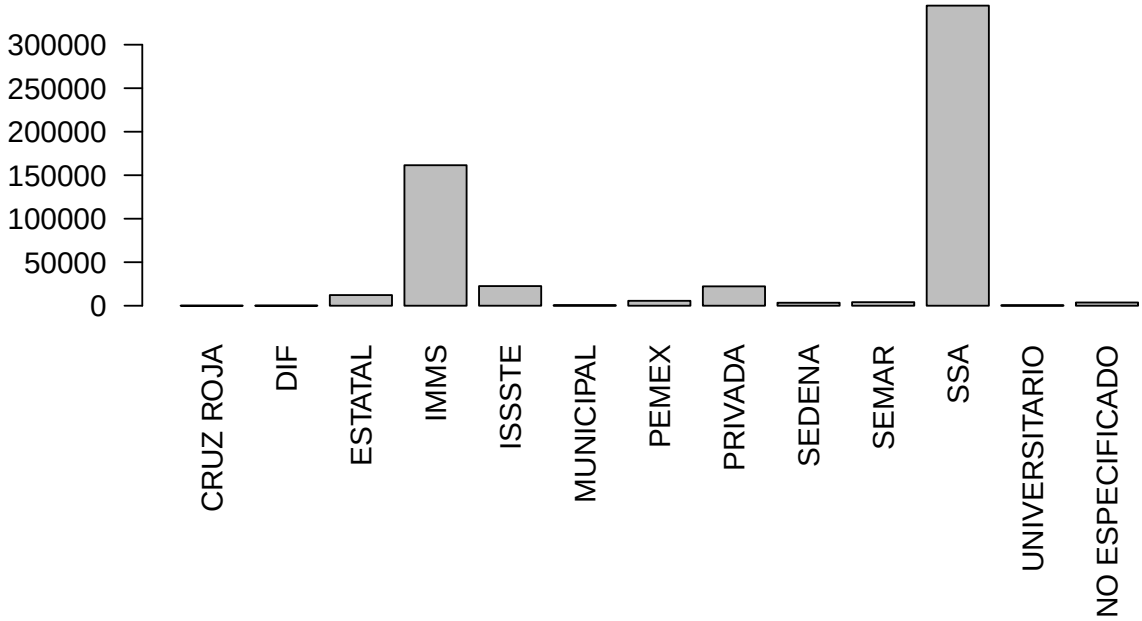
ASPECTOS DE SALUD (2)



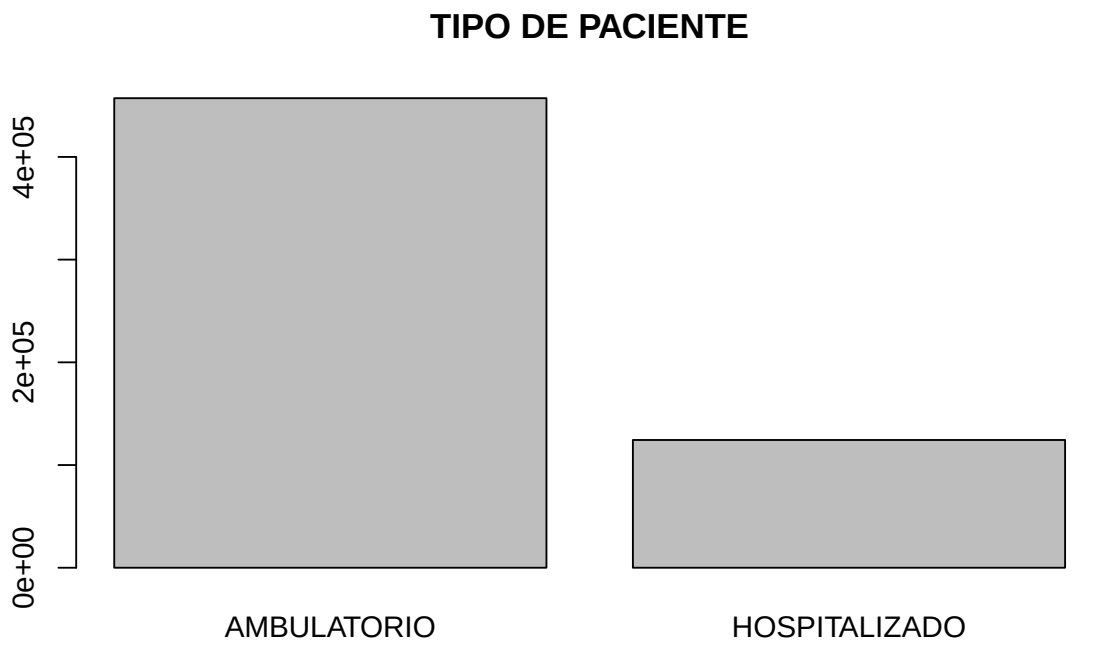
ASPECTOS DE SALUD (3)

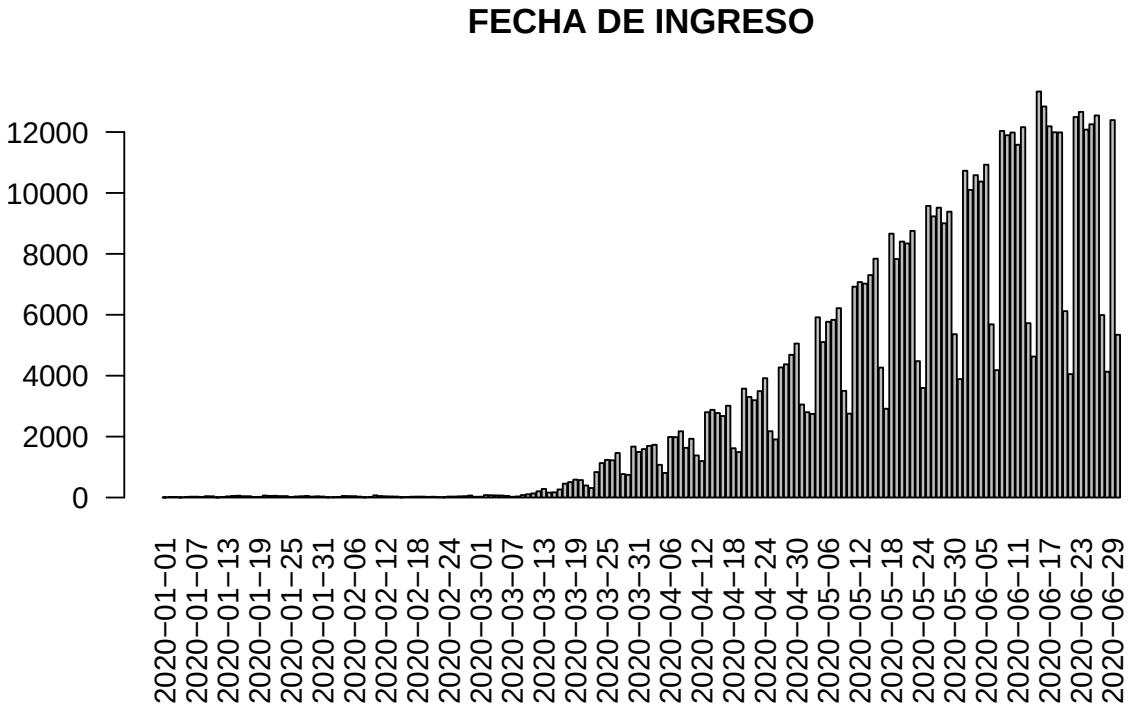


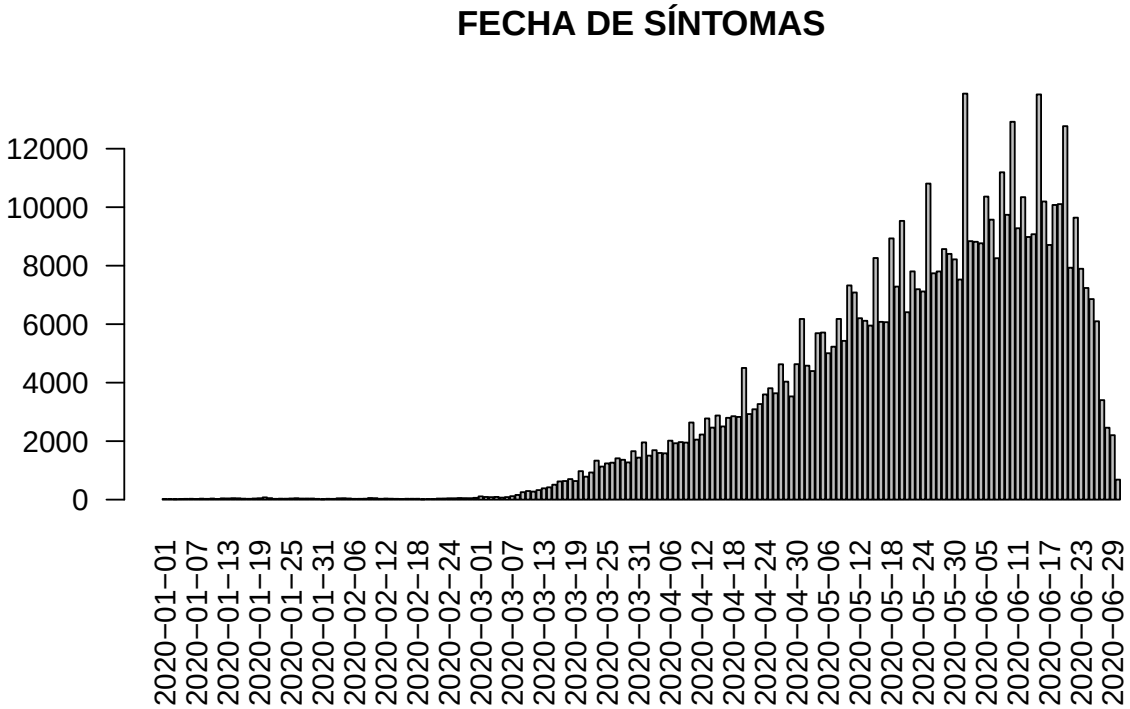
INSTITUCIÓN QUE BRINDÓ LA ATENCIÓN

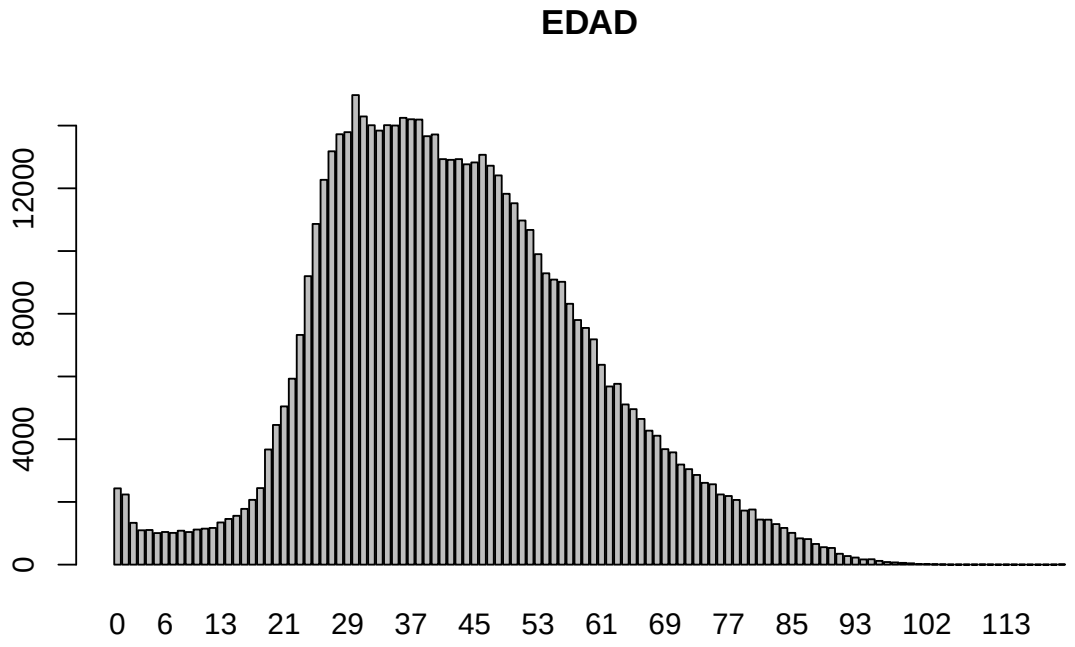


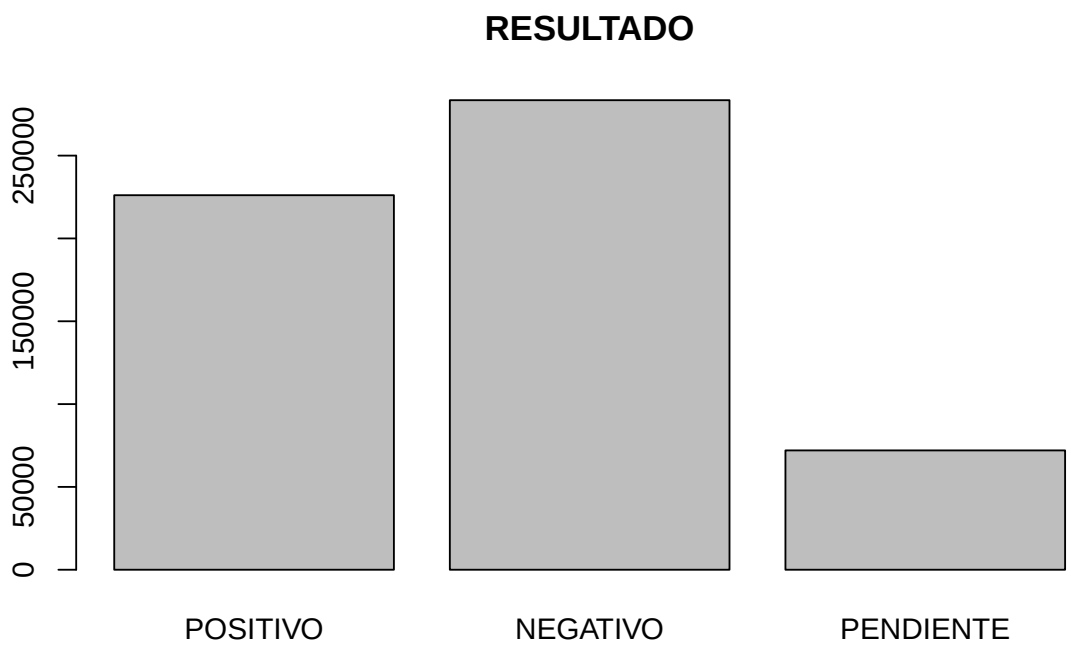
SEXO DEL PACIENTE


TIPO DE PACIENTE
4e+05
2e+05
0e+00
AMBULATORIO
HOSPITALIZADO


FECHA DE INGRESO
12000
10000
8000
6000
4000
2000
0
2020−01−01
2020−01−07
2020−01−13
2020−01−19
2020−01−25
2020−01−31
2020−02−06
2020−02−12
2020−02−18
2020−02−24
2020−03−01
2020−03−07
2020−03−13
2020−03−19
2020−03−25
2020−03−31
2020−04−06
2020−04−12
2020−04−18
2020−04−24
2020−04−30
2020−05−06
2020−05−12
2020−05−18
2020−05−24
2020−05−30
2020−06−05
2020−06−11
2020−06−17
2020−06−23
2020−06−29


FECHA DE SÍNTOMAS
12000
10000
8000
6000
4000
2000
0
2020−01−01
2020−01−07
2020−01−13
2020−01−19
2020−01−25
2020−01−31
2020−02−06
2020−02−12
2020−02−18
2020−02−24
2020−03−01
2020−03−07
2020−03−13
2020−03−19
2020−03−25
2020−03−31
2020−04−06
2020−04−12
2020−04−18
2020−04−24
2020−04−30
2020−05−06
2020−05−12
2020−05−18
2020−05−24
2020−05−30
2020−06−05
2020−06−11
2020−06−17
2020−06−23
2020−06−29


EDAD
0
4000
8000
12000
0
6
13
21
29
37
45
53
61
69
77
85
93
102
113



De acuerdo a las gráficas anteriores es notable que la distribución de las clases no es homogénea alrededor del conjunto de datos, pues los resultados negativos sobre todo en los padecimientos de salud, aforturnadamente la mayoría son negativos. Sin embargo, esto puede afectar al análisis de los datos, más adelante se abordará este problema.

Por otro lado, con base a los intereses del estudio y tomando en cuenta que los datos son de México, se eliminan todas aquellas características que describen el lugar de origen y/o residencia de los pacientes.

```
data$FECHA_ACTUALIZACION <- NULL
data$ID_REGISTRO <- NULL
data$ENTIDAD_UM <- NULL
data$ENTIDAD_NAC <- NULL
data$ENTIDAD_RES <- NULL
data$MUNICIPIO_RES <- NULL
data$NACIONALIDAD <- NULL
data$HABLA_LENGUA_INDIG <- NULL
data$MIGRANTE <- NULL
data$PAIS_NACIONALIDAD <- NULL
data$PAIS_ORIGEN <- NULL
```

```
summary(data)
```

```
##  ORIGEN          SECTOR       SEXO       TIPO_PACIENTE    FECHA_INGRESO
##  1:206811   12      :344924   1:287056   1:457186      2020-06-15: 13330
##  2:374769   4       :161488   2:294524   2:124394      2020-06-16: 12839
##             6       : 22575                            2020-06-23: 12662
##             9       : 22260                            2020-06-26: 12544
##             3       : 12226                            2020-06-22: 12493
##             8       :  5629                            2020-06-29: 12392
##             (Other): 12478                             (Other)   :505320
```

```
##     FECHA_SINTOMAS        FECHA_DEF      INTUBADO    NEUMONIA
##  2020-06-01: 13883   9999-99-99:544456   1 : 10241   1 : 89914
##  2020-06-15: 13855   2020-06-16:   709   2 :114031   2 :491655
##  2020-06-10: 12920   2020-06-08:   706   97:457186   99:    11
##  2020-06-20: 12770   2020-06-10:   697   99:   122
##  2020-06-08: 11193   2020-06-12:   679
##  2020-05-25: 10807   2020-06-09:   662
##  (Other)   :506152   (Other)   : 33671
##       EDAD         EMBARAZO    DIABETES    EPOC         ASMA         INMUSUPR
##  Min.   :  0.00   1 :  4170   1 : 72626   1 :  9304   1 : 18429   1 :  9134
##  1st Qu.: 31.00   2 :281242   2 :506929   2 :570481   2 :561358   2 :570416
##  Median : 41.00   97:294524   98:  2025   98:  1795   98:  1793   98:  2030
##  Mean   : 42.62   98:  1644
##  3rd Qu.: 53.00
##  Max.   :120.00
##
##  HIPERTENSION OTRA_COM    CARDIOVASCULAR OBESIDAD    RENAL_CRONICA TABAQUISMO
##  1 : 94914    1 : 17421   1 : 13048      1 : 94615   1 : 11551     1 : 49110
##  2 :484799    2 :561504   2 :566668      2 :485147   2 :568198     2 :530523
##  98:  1867    98:  2655   98:  1864      98:  1818   98:  1831     98:  1947
##
##
##
##
##  OTRO_CASO   RESULTADO  UCI
##  1 :227917   1:226089   1 : 10313
##  2 :174179   2:283450   2 :113958
##  99:179484   3: 72041   97:457186
##                         99:   123
##
##
##
```

```
names(data)
```

```
##  [1] "ORIGEN"         "SECTOR"         "SEXO"           "TIPO_PACIENTE"
##  [5] "FECHA_INGRESO"  "FECHA_SINTOMAS" "FECHA_DEF"      "INTUBADO"
##  [9] "NEUMONIA"       "EDAD"           "EMBARAZO"       "DIABETES"
## [13] "EPOC"           "ASMA"           "INMUSUPR"       "HIPERTENSION"
## [17] "OTRA_COM"       "CARDIOVASCULAR" "OBESIDAD"       "RENAL_CRONICA"
## [21] "TABAQUISMO"     "OTRO_CASO"      "RESULTADO"      "UCI"
```

Después del procesamiento que se le ha dado al dataset, se continúa con el manejo de las fechas, que si bien no se usarán directamente en los algoritmos, servirán de ayuda para obtener otra información importante que sí ayudará al procesamiento.

```
print('FECHA DE SINTOMAS')
```

```
## [1] "FECHA DE SINTOMAS"
```

```
min(as.character(data$FECHA_SINTOMAS))
```

```
## [1] "2020-01-01"
```

```
max(as.character(data$FECHA_SINTOMAS))
```

```
## [1] "2020-06-30"
```

```r
print('FECHA DE INGRESO')
```

```
## [1] "FECHA DE INGRESO"
```

```r
min(as.character(data$FECHA_INGRESO))
```

```
## [1] "2020-01-01"
```

```r
max(as.character(data$FECHA_INGRESO))
```

```
## [1] "2020-06-30"
```

```r
print('FECHA DE DEFUNCION')
```

```
## [1] "FECHA DE DEFUNCION"
```

```r
min(as.character(data$FECHA_DEF))
```

```
## [1] "2020-01-13"
```

```r
max(as.character(data$FECHA_DEF))
```

```
## [1] "9999-99-99"
```

De acuerdo a la fecha de defunción se agrega un nuevo campo al dataset que indica de forma binaria si el paciente murió o no.

```r
# Establece si el paciente murió de acuerdo a su fecha de defunción.
data$MURIO <- data$FECHA_DEF != '9999-99-99'
data$MURIO <- ifelse(data$MURIO, 1, 2)
data$MURIO <- as.factor(data$MURIO)

# Muestra resumen de datos agrupando de acuerdo a si el paciente murió o no.
data %>% split(data$MURIO) %>% map(summary)
```

```
## $`1`
##  ORIGEN        SECTOR      SEXO      TIPO_PACIENTE    FECHA_INGRESO
##  1:21627   4      :20416   1:13072   1: 3720       2020-06-01:  702
##  2:15497   12     :11524   2:24052   2:33404       2020-06-08:  699
##            6      : 2532                           2020-05-25:  672
##            3      :  772                           2020-05-18:  665
##            9      :  573                           2020-06-02:  645
##            8      :  516                           2020-06-04:  643
##            (Other):  791                           (Other)   :33098
##     FECHA_SINTOMAS       FECHA_DEF     INTUBADO   NEUMONIA        EDAD
##  2020-06-01:  978   2020-06-16:  709   1 : 6008   1 :26930   Min.   :  0.0
##  2020-05-10:  786   2020-06-08:  706   2 :27342   2 :10194   1st Qu.: 51.0
##  2020-05-25:  766   2020-06-10:  697   97: 3720   99:    0   Median : 62.0
##  2020-05-20:  742   2020-06-12:  679   99:   54              Mean   : 60.8
##  2020-05-18:  679   2020-06-09:  662                         3rd Qu.: 72.0
##  2020-05-15:  653   2020-06-17:  651                         Max.   :103.0
##  (Other)   :32520   (Other)   :33020
##  EMBARAZO   DIABETES   EPOC       ASMA       INMUSUPR   HIPERTENSION OTRA_COM
##  1 :   49   1 :13607   1 : 2081   1 :  759   1 : 1393   1 :15591     1 : 2325
##  2 :12987   2 :23257   2 :34783   2 :36112   2 :35456   2 :21290     2 :34425
##  97:24052   98:  260   98:  260   98:  253   98:  275   98:  243     98:  374
##  98:   36
##
##
```

```
##
##  CARDIOVASCULAR OBESIDAD   RENAL_CRONICA TABAQUISMO OTRO_CASO  RESULTADO
##  1 : 2287       1 : 8633   1 : 2876      1 : 3425   1 : 3477   1:27769
##  2 :34561       2 :28223   2 :33990      2 :33433   2 :10844   2: 7158
##  98:  276       98:  268   98:  258      98:  266   99:22803   3: 2197
##
##
##
##
##  UCI        MURIO
##  1 : 4061   1:37124
##  2 :29289   2:    0
##  97: 3720
##  99:   54
##
##
##
##
## $`2`
##  ORIGEN         SECTOR       SEXO       TIPO_PACIENTE    FECHA_INGRESO
##  1:185184   12     :333400   1:273984   1:453466      2020-06-15: 12755
##  2:359272   4      :141072   2:270472   2: 90990      2020-06-23: 12423
##             9      : 21687                            2020-06-26: 12422
##             6      : 20043                            2020-06-16: 12378
##             3      : 11454                            2020-06-29: 12376
##             8      :  5113                            2020-06-22: 12228
##             (Other): 11687                            (Other)   :469874
##     FECHA_SINTOMAS        FECHA_DEF      INTUBADO    NEUMONIA
##  2020-06-15: 13396   9999-99-99:544456   1 :  4233   1 : 62984
##  2020-06-01: 12905   2020-01-13:     0   2 : 86689   2 :481461
##  2020-06-20: 12549   2020-01-14:     0   97:453466   99:    11
##  2020-06-10: 12373   2020-01-15:     0   99:    68
##  2020-06-08: 10585   2020-01-29:     0
##  2020-05-25: 10041   2020-01-30:     0
##  (Other)   :472607   (Other)   :     0
##       EDAD         EMBARAZO    DIABETES    EPOC         ASMA         INMUSUPR
##  Min.   :  0.00   1 :  4121   1 : 59019   1 :  7223   1 : 17670   1 :  7741
##  1st Qu.: 30.00   2 :268255   2 :483672   2 :535698   2 :525246   2 :534960
##  Median : 40.00   97:270472   98:  1765   98:  1535   98:  1540   98:  1755
##  Mean   : 41.38   98:  1608
##  3rd Qu.: 51.00
##  Max.   :120.00
##
##  HIPERTENSION OTRA_COM    CARDIOVASCULAR OBESIDAD    RENAL_CRONICA TABAQUISMO
##  1 : 79323    1 : 15096   1 : 10761      1 : 85982   1 :  8675     1 : 45685
##  2 :463509    2 :527079   2 :532107      2 :456924   2 :534208     2 :497090
##  98:  1624    98:  2281   98:  1588      98:  1550   98:  1573     98:  1681
##
##
##
##
##  OTRO_CASO   RESULTADO  UCI         MURIO
##  1 :224440   1:198320   1 :  6252   1:     0
##  2 :163335   2:276292   2 : 84669   2:544456
```

```
##  99:156681   3: 69844   97:453466
##                         99:    69
##
##
##
```

El anterior resumen agrupa los resultados de acuerdo al resultado de la muerte del paciente, y cabe destacar que el registro de muertes no corresponden totalmente a pacientes con Coronavirus, sino que de las 37124 muertes, 7158 corresponden a personas cuya prueba resultó negativa, así como 2197 fallecidos que aún están en espera de resultados.

A partir de las fechas de síntomas, ingreso a la unidad médica, y de defunción, se generan tres nuevos valores numéricos: días de ingreso, que indica cuánto tardó el paciente en ingresar a la unidad médica a partir de cuando comenzó con síntomas, así como cuanto tiempo duró desde que presentó síntomas hasta el momento del fallecimiento, etc. Una vez obtenidos estos datos, las categorías de fechas ya no son de utilidad, por lo que se eliminan de la base de datos.

```r
# Parsea las fechas como texto.
data$FECHA_SINTOMAS <- as.character(data$FECHA_SINTOMAS)
data$FECHA_INGRESO <- as.character(data$FECHA_INGRESO)
data$FECHA_DEF <- as.character(data$FECHA_DEF)
data$FECHA_DEF[data$FECHA_DEF == '9999-99-99'] <- '2019-01-01'

# Parsea las fechas como date.
data$FECHA_SINTOMAS <- as.Date(data$FECHA_SINTOMAS)
data$FECHA_INGRESO <- as.Date(data$FECHA_INGRESO)
data$FECHA_DEF <- as.Date(data$FECHA_DEF)

# Calcula tiempos de la enfermedad.
data$DIAS_INGRESO <- as.numeric(difftime(data$FECHA_INGRESO, data$FECHA_SINTOMAS)) / 60 / 60 / 24
data$DIAS_ENFERMEDAD <- as.numeric(difftime(data$FECHA_DEF, data$FECHA_SINTOMAS)) / 60 / 60 / 24
data$DIAS_HOSPITALIZACION <- as.numeric(difftime(data$FECHA_DEF, data$FECHA_INGRESO)) / 60 / 60 / 24

# Elimina las fechas del dataset.
data$FECHA_SINTOMAS <- NULL
data$FECHA_INGRESO <- NULL
data$FECHA_DEF <- NULL
```

```r
print('DÍAS DESDE APARICIÓN DE SÍNTOMAS HASTA INGRESO A UNIDAD MÉDICA')
```

```
## [1] "DÍAS DESDE APARICIÓN DE SÍNTOMAS HASTA INGRESO A UNIDAD MÉDICA"
```

```r
summary(data$DIAS_INGRESO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   3.672   5.000 115.000
```

```r
print('DÍAS DESDE APARICIÓN DE SÍNTOMAS HASTA DEFUNCIÓN')
```

```
## [1] "DÍAS DESDE APARICIÓN DE SÍNTOMAS HASTA DEFUNCIÓN"
```

```r
summary(data$DIAS_ENFERMEDAD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -546.0  -527.0  -511.0  -475.3  -488.0    93.0
```

```r
print('DÍAS DE HOSPITALIZACIÓN DEL PACIENTE')
```

```
## [1] "DÍAS DE HOSPITALIZACIÓN DEL PACIENTE"
```

```r
summary(data$DIAS_HOSPITALIZACION)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -546    -532    -515    -479    -492      91
```

```r
data %>% split(data$MURIO) %>% map(summary)
```

```
## $`1`
##  ORIGEN        SECTOR      SEXO      TIPO_PACIENTE INTUBADO   NEUMONIA
##  1:21627   4      :20416   1:13072   1: 3720       1 : 6008   1 :26930
##  2:15497   12     :11524   2:24052   2:33404       2 :27342   2 :10194
##            6      : 2532                           97: 3720   99:    0
##            3      :  772                           99:   54
##            9      :  573
##            8      :  516
##            (Other):  791
##       EDAD       EMBARAZO   DIABETES   EPOC       ASMA       INMUSUPR
##  Min.   :  0.0   1 :   49   1 :13607   1 : 2081   1 :  759   1 : 1393
##  1st Qu.: 51.0   2 :12987   2 :23257   2 :34783   2 :36112   2 :35456
##  Median : 62.0   97:24052   98:  260   98:  260   98:  253   98:  275
##  Mean   : 60.8   98:   36
##  3rd Qu.: 72.0
##  Max.   :103.0
##
##  HIPERTENSION OTRA_COM   CARDIOVASCULAR OBESIDAD   RENAL_CRONICA TABAQUISMO
##  1 :15591     1 : 2325   1 : 2287       1 : 8633   1 : 2876      1 : 3425
##  2 :21290     2 :34425   2 :34561       2 :28223   2 :33990      2 :33433
##  98:  243     98:  374   98:  276       98:  268   98:  258      98:  266
##
##
##
##
##  OTRO_CASO  RESULTADO UCI        MURIO      DIAS_INGRESO   DIAS_ENFERMEDAD
##  1 : 3477   1:27769   1 : 4061   1:37124   Min.   : 0.00   Min.   :-41.00
##  2 :10844   2: 7158   2 :29289   2:    0   1st Qu.: 1.00   1st Qu.:  6.00
##  99:22803   3: 2197   97: 3720             Median : 4.00   Median : 10.00
##                       99:   54             Mean   : 4.05   Mean   : 11.07
##                                            3rd Qu.: 6.00   3rd Qu.: 15.00
##                                            Max.   :41.00   Max.   : 93.00
##
##  DIAS_HOSPITALIZACION
##  Min.   :-41.000
##  1st Qu.:  2.000
##  Median :  5.000
##  Mean   :  7.019
##  3rd Qu.: 10.000
##  Max.   : 91.000
##
##
## $`2`
##  ORIGEN         SECTOR       SEXO       TIPO_PACIENTE INTUBADO    NEUMONIA
##  1:185184   12     :333400   1:273984   1:453466      1 :  4233   1 : 62984
##  2:359272   4      :141072   2:270472   2: 90990      2 : 86689   2 :481461
##             9      : 21687                           97:453466   99:    11
```

```
##             6      : 20043                           99:    68
##             3      : 11454
##             8      :  5113
##             (Other): 11687
##       EDAD         EMBARAZO    DIABETES    EPOC         ASMA         INMUSUPR
##  Min.   :  0.00   1 :  4121   1 : 59019   1 :  7223   1 : 17670   1 :  7741
##  1st Qu.: 30.00   2 :268255   2 :483672   2 :535698   2 :525246   2 :534960
##  Median : 40.00   97:270472   98:  1765   98:  1535   98:  1540   98:  1755
##  Mean   : 41.38   98:  1608
##  3rd Qu.: 51.00
##  Max.   :120.00
##
##  HIPERTENSION OTRA_COM    CARDIOVASCULAR OBESIDAD    RENAL_CRONICA TABAQUISMO
##  1 : 79323    1 : 15096   1 : 10761      1 : 85982   1 :  8675     1 : 45685
##  2 :463509    2 :527079   2 :532107      2 :456924   2 :534208     2 :497090
##  98:  1624    98:  2281   98:  1588      98:  1550   98:  1573     98:  1681
##
##
##
##
##  OTRO_CASO   RESULTADO  UCI         MURIO       DIAS_INGRESO
##  1 :224440   1:198320   1 :  6252   1:     0   Min.   :  0.000
##  2 :163335   2:276292   2 : 84669   2:544456   1st Qu.:  1.000
##  99:156681   3: 69844   97:453466              Median :  3.000
##                         99:    69              Mean   :  3.647
##                                                3rd Qu.:  5.000
##                                                Max.   :115.000
##
##  DIAS_ENFERMEDAD  DIAS_HOSPITALIZACION
##  Min.   :-546.0   Min.   :-546.0
##  1st Qu.:-528.0   1st Qu.:-532.0
##  Median :-514.0   Median :-518.0
##  Mean   :-508.5   Mean   :-512.1
##  3rd Qu.:-494.0   3rd Qu.:-497.0
##  Max.   :-365.0   Max.   :-365.0
##
```

Generación de modelos

Balanceo del dataset

Como se mecionó anteriormente, la clase que se predecirá, es decir, si el paciente murió o no, está desbalanceada, ya que la gran mayoría de las personas estudiadas no murieron. Es por ello que es necesario balancear los registros para tener un equilibrio entre las clases de la característica de interés.

El método que se utiliza para balancear el dataset es **undersampling**, por medio del cual se reducen las observaciones de la clase que tiene mayor número de coincidencias y de este modo, los datos quedan equilibrados. Es importante aclarar que no se usó **oversampling**, que prácticamente es lo opuesto del primero porque este último agrega más registros en las clases con menor cantidad de incidencias para equilibrarlas con la más repetida, pero como de por sí el dataset ya es muy grande, agregarle miles de registros más pudiera no ser tan conveniente para el óptimo rendimiento de los algoritmos.

```
balanced <- downSample(data, data$MURIO)
```

```
balanced %>% split(balanced$MURIO) %>% map(summary)
```

```
## $`1`
##  ORIGEN        SECTOR      SEXO      TIPO_PACIENTE INTUBADO   NEUMONIA
##  1:21627   4      :20416   1:13072   1: 3720       1 : 6008   1 :26930
##  2:15497   12     :11524   2:24052   2:33404       2 :27342   2 :10194
##            6      : 2532                           97: 3720   99:    0
##            3      :  772                           99:   54
##            9      :  573
##            8      :  516
##            (Other):  791
##       EDAD        EMBARAZO   DIABETES   EPOC        ASMA        INMUSUPR
##  Min.   :  0.0   1 :   49   1 :13607   1 : 2081   1 :  759   1 : 1393
##  1st Qu.: 51.0   2 :12987   2 :23257   2 :34783   2 :36112   2 :35456
##  Median : 62.0   97:24052   98:  260   98:  260   98:  253   98:  275
##  Mean   : 60.8   98:   36
##  3rd Qu.: 72.0
##  Max.   :103.0
##
##  HIPERTENSION OTRA_COM   CARDIOVASCULAR OBESIDAD   RENAL_CRONICA TABAQUISMO
##  1 :15591     1 : 2325   1 : 2287       1 : 8633   1 : 2876      1 : 3425
##  2 :21290     2 :34425   2 :34561       2 :28223   2 :33990      2 :33433
##  98:  243     98:  374   98:  276       98:  268   98:  258      98:  266
##
##
##
##
##  OTRO_CASO  RESULTADO UCI        MURIO      DIAS_INGRESO   DIAS_ENFERMEDAD
##  1 : 3477   1:27769   1 : 4061   1:37124   Min.   : 0.00   Min.   :-41.00
##  2 :10844   2: 7158   2 :29289   2:    0   1st Qu.: 1.00   1st Qu.:  6.00
##  99:22803   3: 2197   97: 3720             Median : 4.00   Median : 10.00
##                       99:   54             Mean   : 4.05   Mean   : 11.07
##                                            3rd Qu.: 6.00   3rd Qu.: 15.00
##                                            Max.   :41.00   Max.   : 93.00
##
##  DIAS_HOSPITALIZACION Class
##  Min.   :-41.000      1:37124
##  1st Qu.:  2.000      2:    0
##  Median :  5.000
##  Mean   :  7.019
##  3rd Qu.: 10.000
##  Max.   : 91.000
##
##
## $`2`
##  ORIGEN        SECTOR      SEXO      TIPO_PACIENTE INTUBADO   NEUMONIA
##  1:12588   12     :22815   1:18764   1:31077       1 :  310   1 : 4162
##  2:24536   4      : 9490   2:18360   2: 6047       2 : 5734   2 :32962
##            9      : 1462                           97:31077   99:    0
##            6      : 1381                           99:    3
##            3      :  777
##            8      :  333
##            (Other):  866
##       EDAD        EMBARAZO   DIABETES   EPOC        ASMA        INMUSUPR
```

```
##  Min.   :  0.00   1 :  276   1 : 3959   1 :  512   1 : 1167   1 :  552
##  1st Qu.: 30.00   2 :18360   2 :33049   2 :36511   2 :35851   2 :36452
##  Median : 40.00   97:18360   98:  116   98:  101   98:  106   98:  120
##  Mean   : 41.44   98:  128
##  3rd Qu.: 51.00
##  Max.   :114.00
##
##  HIPERTENSION OTRA_COM   CARDIOVASCULAR OBESIDAD   RENAL_CRONICA TABAQUISMO
##  1 : 5444     1 : 1035   1 :  743       1 : 5879   1 :  603      1 : 3207
##  2 :31568     2 :35930   2 :36274       2 :31133   2 :36414      2 :33802
##  98:  112     98:  159   98:  107       98:  112   98:  107      98:  115
##
##
##
##
##  OTRO_CASO  RESULTADO UCI        MURIO      DIAS_INGRESO    DIAS_ENFERMEDAD
##  1 :15271   1:13597   1 :  414   1:    0   Min.   : 0.000   Min.   :-546.0
##  2 :11295   2:18811   2 : 5630   2:37124   1st Qu.: 1.000   1st Qu.:-529.0
##  99:10558   3: 4716   97:31077             Median : 3.000   Median :-514.0
##                       99:    3             Mean   : 3.648   Mean   :-508.5
##                                            3rd Qu.: 5.000   3rd Qu.:-494.0
##                                            Max.   :58.000   Max.   :-365.0
##
##  DIAS_HOSPITALIZACION Class
##  Min.   :-546.0       1:    0
##  1st Qu.:-532.0       2:37124
##  Median :-518.0
##  Mean   :-512.2
##  3rd Qu.:-498.0
##  Max.   :-365.0
##
```

De acuerdo al resultado del resumen generado, el dataset fue reducido para tener 37124 registros de pacientes que murieron y esta misma cantidad para pacientes que siguen vivos.

```
backup <- balanced
```

En la tranformación de fechas, se calcularon los días que duró vivo un paciente desde que presentó síntomas, así como lo que duró hospitalizado. Como la mayoría de pacientes no murieron, estas características tienen a tener un valor negativo, lo cual no sería de utilidad para los resultados que se pretenden obtener. Por ahora se decide descartarlas también, sin cerrarse a que posteriormente pudieran tratarse y hacer que realmente sirvan.

Para efectos de pruebas, con el fin de reducir aún más la cantidad de registros, se lleva a cabo un segundo balanceo con respecto al resultado de la prueba de Coronavirus, que es la razón de ser del estudio.

```
balanced <- backup
balanced$DIAS_ENFERMEDAD <- NULL
balanced$DIAS_HOSPITALIZACION <- NULL
balanced$Class <- NULL
balanced <- downSample(balanced, balanced$RESULTADO)
```

```
nrow(balanced)
```

```
## [1] 20739
```

```
balanced %>% split(balanced$MURIO) %>% map(summary)
```

```
## $`1`
##  ORIGEN       SECTOR     SEXO     TIPO_PACIENTE INTUBADO  NEUMONIA
##  1:4896   4      :4942   1:3146   1: 884        1 :1394   1 :6007
##  2:3827   12     :2386   2:5577   2:7839        2 :6433   2 :2716
##           6      : 539                          97: 884   99:   0
##           9      : 234                          99:  12
##           10     : 196
##           3      : 156
##           (Other): 270
##       EDAD        EMBARAZO  DIABETES  EPOC      ASMA      INMUSUPR
##  Min.   :  0.00   1 :   9   1 :3160   1 : 488   1 : 162   1 : 331
##  1st Qu.: 52.00   2 :3128   2 :5503   2 :8178   2 :8506   2 :8333
##  Median : 62.00   97:5577   98:  60   98:  57   98:  55   98:  59
##  Mean   : 61.24   98:   9
##  3rd Qu.: 72.00
##  Max.   :103.00
##
##  HIPERTENSION OTRA_COM  CARDIOVASCULAR OBESIDAD  RENAL_CRONICA TABAQUISMO
##  1 :3667      1 : 531   1 : 557        1 :1963   1 : 691       1 : 784
##  2 :5003      2 :8112   2 :8104        2 :6705   2 :7976       2 :7883
##  98:  53      98:  80   98:  62        98:  55   98:  56       98:  56
##
##
##
##
##  OTRO_CASO RESULTADO UCI       MURIO     DIAS_INGRESO    Class
##  1 : 727   1:4669    1 : 965   1:8723   Min.   : 0.000   1:4669
##  2 :2540   2:1857    2 :6862   2:   0   1st Qu.: 1.000   2:1857
##  99:5456   3:2197    97: 884            Median : 3.000   3:2197
##                      99:  12            Mean   : 3.943
##                                         3rd Qu.: 6.000
##                                         Max.   :35.000
##
##
## $`2`
##  ORIGEN       SECTOR     SEXO     TIPO_PACIENTE INTUBADO   NEUMONIA
##  1:4027   12     :7076   1:6077   1:10059       1 :  115   1 : 1300
##  2:7989   4      :3075   2:5939   2: 1957       2 : 1842   2 :10716
##           9      : 761                          97:10059   99:    0
##           6      : 441                          99:    0
##           3      : 250
##           11     : 137
##           (Other): 276
##       EDAD        EMBARAZO  DIABETES   EPOC       ASMA       INMUSUPR
##  Min.   :  0.00   1 : 104   1 : 1301   1 :  153   1 :  361   1 :  180
##  1st Qu.: 30.00   2 :5925   2 :10675   2 :11829   2 :11616   2 :11793
##  Median : 40.00   97:5939   98:   40   98:   34   98:   39   98:   43
##  Mean   : 41.42   98:  48
##  3rd Qu.: 51.00
##  Max.   :110.00
##
##  HIPERTENSION OTRA_COM   CARDIOVASCULAR OBESIDAD   RENAL_CRONICA TABAQUISMO
```

```
##  1 : 1773     1 :  338   1 :  217       1 : 1898   1 :  185      1 : 1016
##  2 :10202     2 :11624   2 :11762       2 :10080   2 :11793      2 :10962
##  98:   41     98:   54   98:   37       98:   38   98:   38      98:   38
##
##
##
##
##  OTRO_CASO RESULTADO UCI        MURIO      DIAS_INGRESO    Class
##  1 :4884   1:2244    1 :  147   1:    0   Min.   : 0.000   1:2244
##  2 :3714   2:5056    2 : 1810   2:12016   1st Qu.: 1.000   2:5056
##  99:3418   3:4716    97:10059             Median : 3.000   3:4716
##                      99:    0             Mean   : 3.635
##                                           3rd Qu.: 5.000
##                                           Max.   :38.000
##
```

Reglas de clasificación: JRip

```
covid_jrip <- JRip(MURIO ~ ., data = balanced)
```

```
covid_jrip
```

```
## JRIP rules:
## ===========
##
## (TIPO_PACIENTE = 2) and (RESULTADO = 1) and (EDAD >= 52) => MURIO=1 (3454.0/222.0)
## (TIPO_PACIENTE = 2) and (SECTOR = 4) => MURIO=1 (3457.0/712.0)
## (TIPO_PACIENTE = 2) and (RESULTADO = 1) and (NEUMONIA = 1) => MURIO=1 (545.0/102.0)
## (NEUMONIA = 1) and (INTUBADO = 1) => MURIO=1 (570.0/63.0)
## (TIPO_PACIENTE = 2) and (EDAD >= 49) and (OTRO_CASO = 99) => MURIO=1 (159.0/58.0)
## (NEUMONIA = 1) and (TIPO_PACIENTE = 1) and (EDAD >= 65) => MURIO=1 (187.0/20.0)
## (NEUMONIA = 1) and (EDAD >= 51) and (RESULTADO = 2) => MURIO=1 (334.0/86.0)
## (TIPO_PACIENTE = 2) and (INTUBADO = 1) => MURIO=1 (47.0/13.0)
## (NEUMONIA = 1) and (EDAD >= 44) and (RESULTADO = 1) => MURIO=1 (149.0/24.0)
## (TIPO_PACIENTE = 2) and (SECTOR = 10) => MURIO=1 (134.0/31.0)
## (TIPO_PACIENTE = 2) and (NEUMONIA = 1) and (EDAD >= 72) => MURIO=1 (113.0/39.0)
## (TIPO_PACIENTE = 2) and (NEUMONIA = 1) and (ORIGEN = 1) and (RESULTADO = 2) and (EDAD >= 33) and (SE
## (SECTOR = 4) and (EDAD >= 60) => MURIO=1 (322.0/142.0)
## (NEUMONIA = 1) and (TIPO_PACIENTE = 1) and (EDAD >= 58) => MURIO=1 (28.0/10.0)
## (NEUMONIA = 1) and (EDAD >= 53) and (HIPERTENSION = 1) and (DIABETES = 1) and (EDAD <= 59) => MURIO=
## (TIPO_PACIENTE = 2) and (ORIGEN = 1) and (OBESIDAD = 1) and (EDAD >= 57) => MURIO=1 (25.0/6.0)
## (TIPO_PACIENTE = 2) and (RESULTADO = 1) and (OBESIDAD = 1) => MURIO=1 (21.0/6.0)
## (TIPO_PACIENTE = 2) and (OTRO_CASO = 2) and (EDAD >= 40) and (EPOC = 1) => MURIO=1 (17.0/5.0)
## (TIPO_PACIENTE = 2) and (NEUMONIA = 1) and (DIAS_INGRESO <= 3) and (DIABETES = 1) => MURIO=1 (35.0/13
## (TIPO_PACIENTE = 2) and (EDAD >= 38) and (HIPERTENSION = 2) and (SECTOR = 3) => MURIO=1 (25.0/10.0)
## (NEUMONIA = 1) and (OTRO_CASO = 2) and (RESULTADO = 2) and (DIAS_INGRESO >= 4) and (EDAD >= 31) and
## (NEUMONIA = 1) and (OTRO_CASO = 2) and (TIPO_PACIENTE = 1) and (DIAS_INGRESO >= 5) => MURIO=1 (37.0/1
##  => MURIO=2 (11000.0/576.0)
##
## Number of Rules : 23
```

```
summary(covid_jrip)
```

```
##
```

```
## === Summary ===
##
## Correctly Classified Instances       18571               89.5463 %
## Incorrectly Classified Instances      2168               10.4537 %
## Kappa statistic                          0.7889
## Mean absolute error                      0.1725
## Root mean squared error                  0.2937
## Relative absolute error                 35.3894 %
## Root relative squared error             59.4891 %
## Total Number of Instances            20739
##
## === Confusion Matrix ===
##
##      a     b   <-- classified as
##   8147   576 |     a = 1
##   1592 10424 |     b = 2
```

El primer algoritmo implementado fue JRip para reglas de clasificación, por medio de este algoritmo se obtienen una serie de condiciones para llegar a un resultado, ideal para ser usado con la naturaliza de los datos analizados.

El resultado puede decirse que fue bueno, ya que clasificó correctamente el 89% de los registros. Lo único preocupante a partir de estos resultados es el uso que se les pueda dar, ya que por ejemplo, el algoritmo determinó que 681 personas morirían, pero en realidad no fue así. Este dato puede ser desde una simple "buena noticia", hasta una tragedia, ya que como las esperanzas de vida eran pocas según la predicción, es posible que la atención médica se haya otorgado a alguien más, y de esta manera ocasionar más muertes por Covid-19, que de haber sido antendido el paciente, no hubieran ocurrido.

De esta misma manera pueden interpretarse los otros errores, se tienen 1584 pacientes que se determinó que no morirían pero sí ocurrió el fallecimiento. Todo depende de qué resultado erróneo sea más grave, y en base a esto pueden implementarse mejoras al algoritmo.

Vecinos Cercanos: Knn

```
# Divide datos en entrenamiento y prueba.
dt <- sort(sample(nrow(balanced), nrow(balanced) * .7))
dt_z <- as.data.frame(scale(dt[-1]))
train <- balanced[dt, ]
train_labels <- train$MURIO
test <- balanced[-dt, ]
test_labels <- test$MURIO

# Elige k
k <- round(sqrt(nrow(balanced)))
k <- if (k %% 2) k else k + 1
k
```

```
## [1] 145
```

```
# Aplica el algoritmo kNN.
test_pred <- knn(train = train, test = test,
                      cl = train_labels, k = k)

# Compara resultados de la predicción en el dataset de prueba.
CrossTable(x = test_labels, y = test_pred, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  6222
##
##
##              | test_pred
##  test_labels |         1 |         2 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |      2391 |       227 |      2618 |
##              |     0.913 |     0.087 |     0.421 |
##              |     0.810 |     0.069 |           |
##              |     0.384 |     0.036 |           |
## -------------|-----------|-----------|-----------|
##            2 |       561 |      3043 |      3604 |
##              |     0.156 |     0.844 |     0.579 |
##              |     0.190 |     0.931 |           |
##              |     0.090 |     0.489 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      2952 |      3270 |      6222 |
##              |     0.474 |     0.526 |           |
## -------------|-----------|-----------|-----------|
##
##
```

El siguiente algoritmo aplicado fue Knn, que para el subconjunto de datos de prueba, el 9% de las predicciones donde se dijo que el paciente iba a morir, no fue así, y por otro lado, el 15% de los pacientes cuyo resultado en la predicción se establecía que no morirían, en realidad sí murieron.

Árboles de decisión: C5.0

```
covid_c5 <- C5.0(MURIO ~ ., data = balanced, rules = TRUE)
```

```
summary(covid_c5)
```

```
##
## Call:
## C5.0.formula(formula = MURIO ~ ., data = balanced, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Jul  7 23:22:15 2020
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 20739 cases (24 attributes) from undefined.data
```

```
##
## Rules:
##
## Rule 1: (23, lift 2.3)
##  INTUBADO = 97
##  NEUMONIA = 1
##  EDAD <= 50
##  DIABETES = 1
##  RESULTADO = 1
##  ->  class 1  [0.960]
##
## Rule 2: (2963/168, lift 2.2)
##  NEUMONIA = 1
##  EDAD > 50
##  RESULTADO = 1
##  ->  class 1  [0.943]
##
## Rule 3: (958/56, lift 2.2)
##  SECTOR in {3, 6, 7, 8, 9, 12, 13, 99}
##  INTUBADO in {1, 99}
##  EDAD > 49
##  ->  class 1  [0.941]
##
## Rule 4: (14, lift 2.2)
##  SECTOR = 10
##  INTUBADO = 97
##  NEUMONIA = 2
##  EDAD > 43
##  EDAD <= 55
##  RESULTADO = 3
##  ->  class 1  [0.938]
##
## Rule 5: (38/2, lift 2.2)
##  NEUMONIA = 1
##  DIABETES = 98
##  DIAS_INGRESO <= 7
##  ->  class 1  [0.925]
##
## Rule 6: (1906/143, lift 2.2)
##  SECTOR in {4, 8}
##  EDAD > 55
##  RESULTADO = 1
##  ->  class 1  [0.925]
##
## Rule 7: (1521/115, lift 2.2)
##  INTUBADO in {1, 99}
##  ->  class 1  [0.924]
##
## Rule 8: (3573/273, lift 2.2)
##  INTUBADO in {1, 2, 99}
##  NEUMONIA = 1
##  EDAD > 18
##  RESULTADO = 1
##  ->  class 1  [0.923]
```

```
##
## Rule 9: (10, lift 2.2)
##  SECTOR = 6
##  INTUBADO in {1, 2}
##  NEUMONIA = 1
##  TABAQUISMO = 2
##  RESULTADO = 3
##  UCI = 1
##  ->  class 1  [0.917]
##
## Rule 10: (1348/134, lift 2.1)
##  SECTOR in {4, 8}
##  EDAD > 73
##  ->  class 1  [0.900]
##
## Rule 11: (2964/318, lift 2.1)
##  NEUMONIA = 1
##  EDAD > 64
##  ->  class 1  [0.892]
##
## Rule 12: (353/38, lift 2.1)
##  SECTOR in {4, 8}
##  EDAD > 55
##  RENAL_CRONICA = 1
##  ->  class 1  [0.890]
##
## Rule 13: (49/5, lift 2.1)
##  NEUMONIA = 1
##  DIABETES = 98
##  ->  class 1  [0.882]
##
## Rule 14: (4794/569, lift 2.1)
##  NEUMONIA = 1
##  EDAD > 50
##  OTRO_CASO in {2, 99}
##  ->  class 1  [0.881]
##
## Rule 15: (93/12, lift 2.1)
##  SECTOR = 10
##  NEUMONIA = 2
##  EDAD > 55
##  ->  class 1  [0.863]
##
## Rule 16: (728/101, lift 2.0)
##  INTUBADO in {1, 2, 99}
##  RENAL_CRONICA = 1
##  ->  class 1  [0.860]
##
## Rule 17: (153/24, lift 2.0)
##  SECTOR in {1, 7, 10}
##  NEUMONIA = 2
##  EDAD > 43
##  ->  class 1  [0.839]
##
```

```
## Rule 18: (9796/1957, lift 1.9)
##  INTUBADO in {1, 2, 99}
##  ->  class 1  [0.800]
##
## Rule 19: (3699/18, lift 1.7)
##  INTUBADO = 97
##  EDAD <= 64
##  OTRO_CASO = 1
##  RESULTADO in {2, 3}
##  ->  class 2  [0.995]
##
## Rule 20: (4385/58, lift 1.7)
##  SECTOR = 12
##  NEUMONIA = 2
##  EDAD <= 51
##  EMBARAZO in {2, 97, 98}
##  OTRA_COM = 2
##  OBESIDAD = 2
##  ->  class 2  [0.987]
##
## Rule 21: (6687/98, lift 1.7)
##  INTUBADO = 97
##  EDAD <= 50
##  RESULTADO in {2, 3}
##  ->  class 2  [0.985]
##
## Rule 22: (8555/139, lift 1.7)
##  SECTOR in {2, 3, 4, 6, 8, 9, 11, 12, 13, 99}
##  INTUBADO = 97
##  NEUMONIA = 2
##  EDAD <= 55
##  ->  class 2  [0.984]
##
## Rule 23: (7641/127, lift 1.7)
##  SECTOR in {1, 2, 3, 6, 7, 9, 11, 12, 13, 99}
##  INTUBADO = 97
##  NEUMONIA = 2
##  ->  class 2  [0.983]
##
## Rule 24: (7731/152, lift 1.7)
##  INTUBADO = 97
##  EDAD <= 50
##  DIABETES = 2
##  ->  class 2  [0.980]
##
## Rule 25: (8034/168, lift 1.7)
##  INTUBADO = 97
##  NEUMONIA = 2
##  EDAD <= 73
##  RENAL_CRONICA = 2
##  RESULTADO in {2, 3}
##  ->  class 2  [0.979]
##
## Rule 26: (3558/77, lift 1.7)
```

```
##  SEXO = 1
##  INTUBADO = 97
##  NEUMONIA = 2
##  RENAL_CRONICA = 2
##  DIAS_INGRESO <= 4
##  ->  class 2  [0.978]
##
## Rule 27: (5503/123, lift 1.7)
##  SECTOR = 12
##  NEUMONIA = 2
##  EMBARAZO in {2, 97, 98}
##  RESULTADO in {2, 3}
##  ->  class 2  [0.977]
##
## Rule 28: (2916/84, lift 1.7)
##  SECTOR in {3, 6, 7, 8, 9, 12, 13, 99}
##  NEUMONIA = 2
##  OTRA_COM = 2
##  RESULTADO = 3
##  ->  class 2  [0.971]
##
## Rule 29: (2066/79, lift 1.7)
##  SECTOR in {2, 3, 6, 9, 11, 12, 13}
##  EDAD <= 30
##  INMUSUPR in {2, 98}
##  RENAL_CRONICA = 2
##  RESULTADO in {2, 3}
##  ->  class 2  [0.961]
##
## Rule 30: (178/7, lift 1.6)
##  SECTOR in {8, 9, 13}
##  NEUMONIA = 2
##  RESULTADO = 2
##  ->  class 2  [0.956]
##
## Rule 31: (2404/124, lift 1.6)
##  SECTOR in {3, 6, 9, 11, 12, 13}
##  EDAD <= 49
##  RENAL_CRONICA = 2
##  RESULTADO = 3
##  ->  class 2  [0.948]
##
## Rule 32: (104/6, lift 1.6)
##  EMBARAZO = 1
##  OTRA_COM = 2
##  ->  class 2  [0.934]
##
## Rule 33: (1010/70, lift 1.6)
##  SECTOR in {6, 8, 9, 10, 11, 99}
##  NEUMONIA = 2
##  EDAD <= 51
##  ->  class 2  [0.930]
##
## Rule 34: (12, lift 1.6)
```

```
##  INTUBADO = 2
##  EDAD <= 18
##  OTRO_CASO in {1, 2}
##  RESULTADO = 1
##  ->  class 2  [0.929]
##
## Rule 35: (162/12, lift 1.6)
##  SECTOR in {6, 11, 13}
##  EDAD <= 49
##  RESULTADO = 2
##  ->  class 2  [0.921]
##
## Rule 36: (9, lift 1.6)
##  SECTOR = 6
##  INTUBADO = 2
##  TABAQUISMO = 1
##  RESULTADO = 3
##  ->  class 2  [0.909]
##
## Rule 37: (105/9, lift 1.6)
##  SECTOR in {3, 6, 9, 12, 99}
##  ASMA = 1
##  RESULTADO = 3
##  ->  class 2  [0.907]
##
## Rule 38: (47/5, lift 1.5)
##  EDAD <= 18
##  RESULTADO = 1
##  DIAS_INGRESO <= 2
##  ->  class 2  [0.878]
##
## Rule 39: (126/18, lift 1.5)
##  SECTOR = 9
##  INTUBADO = 2
##  EDAD <= 71
##  OTRA_COM = 2
##  RESULTADO = 3
##  ->  class 2  [0.852]
##
## Default class: 2
##
##
## Evaluation on training data (20739 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##      39 2115(10.2%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##    8146   577    (a): class 1
```

```
##    1538 10478    (b): class 2
##
##
##  Attribute usage:
##
##   97.30% INTUBADO
##   86.96% EDAD
##   81.12% NEUMONIA
##   70.65% SECTOR
##   66.56% RESULTADO
##   48.05% RENAL_CRONICA
##   41.01% OTRO_CASO
##   37.62% DIABETES
##   30.38% EMBARAZO
##   28.90% OTRA_COM
##   21.14% OBESIDAD
##   17.47% DIAS_INGRESO
##   17.16% SEXO
##    9.96% INMUSUPR
##    0.51% ASMA
##    0.09% TABAQUISMO
##    0.05% UCI
##
##
## Time: 0.2 secs
```

En esta ocasión de probó con árboles de decisión por medio del algoritmo C5.0, donde los resultados del entrenamiento arrojan que el 6% de los registros de pacientes que murieron se predijo que sobrevivirían, y el 14% de los que sobrevivieron se había establecido que morirían. Hasta ahora es el algoritmo que ha tenido un mejor rendimiento.

Naive Bayes

```
install_missing_packages(c('tm', 'SnowballC', 'wordcloud', 'e1071'))

library(tm)
```

```
## Loading required package: NLP
##
## Attaching package: 'NLP'
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(SnowballC)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(e1071)
library(gmodels)
```

```
# Divide datos en entrenamiento y prueba.
dt = sort(sample(nrow(balanced), nrow(balanced) * .7))
```

```r
train_data <- balanced[dt, ]
train_labels <- train$MURIO
test_data <- balanced[-dt, ]
test_labels <- test$MURIO

# Crea clasificador
classifier <- naiveBayes(train_data, train_labels)

# Aplica predicción.
test_pred <- predict(classifier, test_data)

# Compara las predicciones con los valores verdaderos.
CrossTable(
  test_pred,
  test_labels,
  prop.chisq = FALSE,
  prop.t = FALSE,
  dnn = c('predicted', 'actual')
)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  6222
##
##
##              | actual
##    predicted |         1 |         2 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |      1529 |      1060 |      2589 |
##              |     0.591 |     0.409 |     0.416 |
##              |     0.584 |     0.294 |           |
## -------------|-----------|-----------|-----------|
##            2 |      1089 |      2544 |      3633 |
##              |     0.300 |     0.700 |     0.584 |
##              |     0.416 |     0.706 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      2618 |      3604 |      6222 |
##              |     0.421 |     0.579 |           |
## -------------|-----------|-----------|-----------|
##
##
```

El algoritmo de Naive Bayes no tuvo el desempeño deseado, prácticamente el margen de error es de la mitad, por lo que este método no es el ideal con los datos presentados, y mucho menos tratándose de un tema de salud.

Regresión lineal

```
install_missing_packages(c('corrplot', 'psych', 'mnormt'))
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
summary(balanced)
```

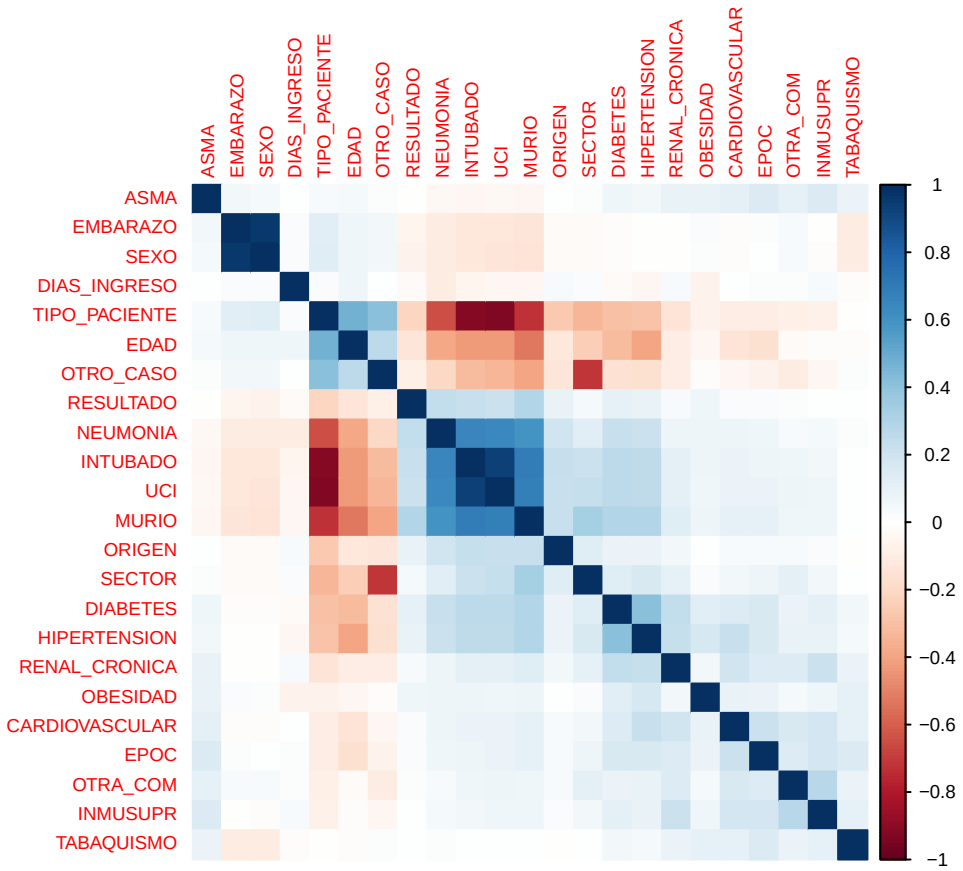
```
##  ORIGEN        SECTOR     SEXO      TIPO_PACIENTE INTUBADO   NEUMONIA
##  1: 8923   12     :9462   1: 9223   1:10943       1 : 1509   1 : 7307
##  2:11816   4      :8017   2:11516   2: 9796       2 : 8275   2 :13432
##            9      : 995                           97:10943   99:    0
##            6      : 980                           99:   12
##            3      : 406
##            10     : 282
##            (Other): 597
##       EDAD        EMBARAZO   DIABETES   EPOC       ASMA       INMUSUPR
##  Min.   :  0.00   1 :  113   1 : 4461   1 :  641   1 :  523   1 :  511
##  1st Qu.: 36.00   2 : 9053   2 :16178   2 :20007   2 :20122   2 :20126
##  Median : 50.00   97:11516   98:  100   98:   91   98:   94   98:  102
##  Mean   : 49.76   98:   57
##  3rd Qu.: 64.00
##  Max.   :110.00
##
##  HIPERTENSION OTRA_COM   CARDIOVASCULAR OBESIDAD   RENAL_CRONICA TABAQUISMO
##  1 : 5440     1 :  869   1 :  774       1 : 3861   1 :  876      1 : 1800
##  2 :15205     2 :19736   2 :19866       2 :16785   2 :19769      2 :18845
##  98:   94     98:  134   98:   99       98:   93   98:   94      98:   94
##
##
##
##
##  OTRO_CASO RESULTADO UCI        MURIO      DIAS_INGRESO    Class
##  1 :5611   1:6913    1 : 1112   1: 8723   Min.   : 0.000   1:6913
##  2 :6254   2:6913    2 : 8672   2:12016   1st Qu.: 1.000   2:6913
##  99:8874   3:6913    97:10943             Median : 3.000   3:6913
##                      99:   12             Mean   : 3.765
##                                           3rd Qu.: 6.000
##                                           Max.   :38.000
##
```

```
balanced$ORIGEN = as.numeric(balanced$ORIGEN)
balanced$SECTOR = as.numeric(balanced$SECTOR)
balanced$SEXO = as.numeric(balanced$SEXO)
balanced$TIPO_PACIENTE = as.numeric(balanced$TIPO_PACIENTE)
balanced$NEUMONIA = as.numeric(balanced$NEUMONIA)
```

```r
balanced$INTUBADO = as.numeric(balanced$INTUBADO)
balanced$EDAD = as.numeric(balanced$EDAD)
balanced$EMBARAZO = as.numeric(balanced$EMBARAZO)
balanced$DIABETES = as.numeric(balanced$DIABETES)
balanced$EPOC = as.numeric(balanced$EPOC)
balanced$ASMA = as.numeric(balanced$ASMA)
balanced$INMUSUPR = as.numeric(balanced$INMUSUPR)
balanced$HIPERTENSION = as.numeric(balanced$HIPERTENSION)
balanced$OTRA_COM = as.numeric(balanced$OTRA_COM)
balanced$CARDIOVASCULAR = as.numeric(balanced$CARDIOVASCULAR)
balanced$OBESIDAD = as.numeric(balanced$OBESIDAD)
balanced$RENAL_CRONICA = as.numeric(balanced$RENAL_CRONICA)
balanced$TABAQUISMO = as.numeric(balanced$TABAQUISMO)
balanced$OTRO_CASO = as.numeric(balanced$OTRO_CASO)
balanced$RESULTADO = as.numeric(balanced$RESULTADO)
balanced$UCI = as.numeric(balanced$UCI)
balanced$Class <- NULL

# Divide datos en entrenamiento y prueba.
dt = sort(sample(nrow(balanced), nrow(balanced) * .7))
train_data <- balanced[dt, ]
train_labels <- train$MURIO
test_data <- balanced[-dt, ]
test_labels <- test$MURIO
```

```r
train_data$MURIO = as.numeric(train_data$MURIO)
test_data$MURIO = as.numeric(test_data$MURIO)
correlacion = cor(train_data[names(train_data)])
corrplot(correlacion, method = "color", order = "AOE", tl.cex = 0.6, cl.cex = 0.6)
```




```
correlacion
```

```
##                      ORIGEN       SECTOR         SEXO TIPO_PACIENTE
## ORIGEN          1.0000000000  0.134386665 -0.022296012  -0.262308838
## SECTOR          0.1343866654  1.000000000 -0.028956107  -0.332238987
## SEXO           -0.0222960117 -0.028956107  1.000000000   0.133726708
## TIPO_PACIENTE  -0.2623088376 -0.332238987  0.133726708   1.000000000
## INTUBADO        0.2387265902  0.210605064 -0.129165344  -0.915333415
## NEUMONIA        0.1925161917  0.128113764 -0.109087098  -0.645968968
## EDAD           -0.1235062683 -0.246111588  0.062086520   0.472457596
## EMBARAZO       -0.0238870546 -0.022863553  0.968081180   0.128551343
## DIABETES        0.0859256369  0.138578099 -0.017034549  -0.292570393
## EPOC            0.0348972203  0.077102462  0.004653729  -0.094368225
## ASMA            0.0003661791  0.010674337  0.045005011   0.038444015
## INMUSUPR        0.0281713593  0.059552470 -0.010199170  -0.078420946
## HIPERTENSION    0.0880195735  0.160577305 -0.003260910  -0.288055982
## OTRA_COM        0.0340102951  0.117541924  0.030525544  -0.085881872
## CARDIOVASCULAR  0.0353776822  0.052272935 -0.015890781  -0.093722538
## OBESIDAD        0.0069324131  0.027036510  0.018061472  -0.060982870
## RENAL_CRONICA   0.0521197252  0.102386894 -0.008069951  -0.141727705
## TABAQUISMO     -0.0081445819  0.008358637 -0.107894056  -0.004136981
## OTRO_CASO      -0.1349649385 -0.715470052  0.057208309   0.411745364
## RESULTADO       0.0951012810  0.044231157 -0.061525182  -0.215942023
## UCI             0.2297361789  0.235281959 -0.131398028  -0.928250058
## MURIO           0.2245583438  0.330131111 -0.142605704  -0.727611213
## DIAS_INGRESO    0.0333596896  0.027700649  0.026422706   0.021410254
##                     INTUBADO    NEUMONIA        EDAD     EMBARAZO    DIABETES
```

```
## ORIGEN          0.2387265902  0.19251619 -0.12350627 -0.023887055  0.08592564
## SECTOR          0.2106050644  0.12811376 -0.24611159 -0.022863553  0.13857810
## SEXO           -0.1291653436 -0.10908710  0.06208652  0.968081180 -0.01703455
## TIPO_PACIENTE  -0.9153334146 -0.64596897  0.47245760  0.128551343 -0.29257039
## INTUBADO        1.0000000000  0.65533860 -0.42361603 -0.126384354  0.26859291
## NEUMONIA        0.6553385958  1.00000000 -0.38362102 -0.108216289  0.22848878
## EDAD           -0.4236160277 -0.38362102  1.00000000  0.067783744 -0.31005054
## EMBARAZO       -0.1263843537 -0.10821629  0.06778374  1.000000000 -0.01964270
## DIABETES        0.2685929088  0.22848878 -0.31005054 -0.019642701  1.00000000
## EPOC            0.0765078356  0.06777895 -0.16647164  0.010479356  0.16213907
## ASMA           -0.0404188938 -0.03188797  0.04408335  0.052451588  0.06338274
## INMUSUPR        0.0571500755  0.04182799 -0.01737981 -0.005124674  0.11379451
## HIPERTENSION    0.2574466773  0.21985353 -0.39494177 -0.005095981  0.41671537
## OTRA_COM        0.0618539117  0.03122162 -0.02835904  0.037105746  0.08630761
## CARDIOVASCULAR  0.0855920706  0.07407373 -0.14356923 -0.013935686  0.14659512
## OBESIDAD        0.0707607235  0.07560511 -0.04987411  0.020579647  0.12516004
## RENAL_CRONICA   0.1153667528  0.07632171 -0.09649717 -0.006421159  0.24331400
## TABAQUISMO      0.0009443972  0.01360464 -0.01065123 -0.102001701  0.05310441
## OTRO_CASO      -0.3109517525 -0.20352251  0.26255124  0.054130349 -0.15738117
## RESULTADO       0.2206778977  0.24961961 -0.13561845 -0.057152493  0.10098159
## UCI             0.9334223620  0.64638241 -0.42790489 -0.127292435  0.26945858
## MURIO           0.6944620167  0.59847858 -0.52728371 -0.139675185  0.29621459
## DIAS_INGRESO   -0.0549203867 -0.10656599  0.06173020  0.025821004 -0.02738424
##                        EPOC          ASMA     INMUSUPR HIPERTENSION    OTRA_COM
## ORIGEN          0.034897220  0.0003661791  0.028171359  0.088019574  0.03401030
## SECTOR          0.077102462  0.0106743366  0.059552470  0.160577305  0.11754192
## SEXO            0.004653729  0.0450050107 -0.010199170 -0.003260910  0.03052554
## TIPO_PACIENTE  -0.094368225  0.0384440154 -0.078420946 -0.288055982 -0.08588187
## INTUBADO        0.076507836 -0.0404188938  0.057150075  0.257446677  0.06185391
## NEUMONIA        0.067778951 -0.0318879735  0.041827991  0.219853531  0.03122162
## EDAD           -0.166471639  0.0440833485 -0.017379814 -0.394941768 -0.02835904
## EMBARAZO        0.010479356  0.0524515884 -0.005124674 -0.005095981  0.03710575
## DIABETES        0.162139066  0.0633827384  0.113794508  0.416715373  0.08630761
## EPOC            1.000000000  0.1521678640  0.189512801  0.165761618  0.14537370
## ASMA            0.152167864  1.0000000000  0.150752138  0.057540464  0.10719922
## INMUSUPR        0.189512801  0.1507521375  1.000000000  0.095355951  0.27607178
## HIPERTENSION    0.165761618  0.0575404639  0.095355951  1.000000000  0.09215101
## OTRA_COM        0.145373701  0.1071992173  0.276071781  0.092151008  1.00000000
## CARDIOVASCULAR  0.210313607  0.1142625428  0.182311084  0.225680750  0.16908573
## OBESIDAD        0.083670573  0.0975413950  0.069714160  0.175082632  0.04643648
## RENAL_CRONICA   0.140804321  0.0960995982  0.214081190  0.236087211  0.14051756
## TABAQUISMO      0.158732074  0.0802117117  0.109116983  0.048702268  0.08809845
## OTRO_CASO      -0.066883296  0.0169599437 -0.046453700 -0.165299072 -0.10450318
## RESULTADO       0.021073305 -0.0088004334  0.008516877  0.098663801  0.01530625
## UCI             0.082216245 -0.0344207106  0.068952612  0.259330001  0.07188421
## MURIO           0.103020408 -0.0406358739  0.060174264  0.298155289  0.06367474
## DIAS_INGRESO    0.019789850  0.0085639517  0.032081169 -0.040669951  0.01621556
##                CARDIOVASCULAR     OBESIDAD RENAL_CRONICA    TABAQUISMO
## ORIGEN           0.035377682  0.006932413   0.052119725 -0.0081445819
## SECTOR           0.052272935  0.027036510   0.102386894  0.0083586365
## SEXO            -0.015890781  0.018061472  -0.008069951 -0.1078940556
## TIPO_PACIENTE   -0.093722538 -0.060982870  -0.141727705 -0.0041369814
## INTUBADO         0.085592071  0.070760724   0.115366753  0.0009443972
## NEUMONIA         0.074073726  0.075605111   0.076321714  0.0136046426
```

```
## EDAD             -0.143569233 -0.049874105  -0.096497172 -0.0106512300
## EMBARAZO         -0.013935686  0.020579647  -0.006421159 -0.1020017008
## DIABETES          0.146595119  0.125160039   0.243313999  0.0531044090
## EPOC              0.210313607  0.083670573   0.140804321  0.1587320740
## ASMA              0.114262543  0.097541395   0.096099598  0.0802117117
## INMUSUPR          0.182311084  0.069714160   0.214081190  0.1091169833
## HIPERTENSION      0.225680750  0.175082632   0.236087211  0.0487022677
## OTRA_COM          0.169085727  0.046436478   0.140517564  0.0880984536
## CARDIOVASCULAR    1.000000000  0.098942488   0.190622561  0.1055916092
## OBESIDAD          0.098942488  1.000000000   0.053104574  0.1032813748
## RENAL_CRONICA     0.190622561  0.053104574   1.000000000  0.0861697454
## TABAQUISMO        0.105591609  0.103281375   0.086169745  1.0000000000
## OTRO_CASO        -0.046472461 -0.018843020  -0.096434393  0.0128116506
## RESULTADO         0.025235068  0.069129323   0.034123634  0.0028105997
## UCI               0.085329192  0.061520668   0.118559244  0.0018064390
## MURIO             0.100932427  0.078968253   0.138583332  0.0042320333
## DIAS_INGRESO      0.006261908 -0.060138440   0.036824417 -0.0152298203
##                    OTRO_CASO    RESULTADO          UCI        MURIO
## ORIGEN         -0.1349649385  0.095101281  0.229736179  0.224558344
## SECTOR         -0.7154700522  0.044231157  0.235281959  0.330131111
## SEXO            0.0572083092 -0.061525182 -0.131398028 -0.142605704
## TIPO_PACIENTE   0.4117453641 -0.215942023 -0.928250058 -0.727611213
## INTUBADO       -0.3109517525  0.220677898  0.933422362  0.694462017
## NEUMONIA       -0.2035225135  0.249619611  0.646382412  0.598478577
## EDAD            0.2625512441 -0.135618447 -0.427904893 -0.527283706
## EMBARAZO        0.0541303485 -0.057152493 -0.127292435 -0.139675185
## DIABETES       -0.1573811716  0.100981587  0.269458582  0.296214588
## EPOC           -0.0668832962  0.021073305  0.082216245  0.103020408
## ASMA            0.0169599437 -0.008800433 -0.034420711 -0.040635874
## INMUSUPR       -0.0464537002  0.008516877  0.068952612  0.060174264
## HIPERTENSION   -0.1652990722  0.098663801  0.259330001  0.298155289
## OTRA_COM       -0.1045031761  0.015306255  0.071884210  0.063674741
## CARDIOVASCULAR -0.0464724611  0.025235068  0.085329192  0.100932427
## OBESIDAD       -0.0188430205  0.069129323  0.061520668  0.078968253
## RENAL_CRONICA  -0.0964343929  0.034123634  0.118559244  0.138583332
## TABAQUISMO      0.0128116506  0.002810600  0.001806439  0.004232033
## OTRO_CASO       1.0000000000 -0.081118638 -0.333213163 -0.395690514
## RESULTADO      -0.0811186378  1.000000000  0.213205264  0.293476725
## UCI            -0.3332131629  0.213205264  1.000000000  0.687003523
## MURIO          -0.3956905137  0.293476725  0.687003523  1.000000000
## DIAS_INGRESO    0.0007982205 -0.026851948 -0.048825783 -0.049168073
##                 DIAS_INGRESO
## ORIGEN          0.0333596896
## SECTOR          0.0277006489
## SEXO            0.0264227064
## TIPO_PACIENTE   0.0214102537
## INTUBADO       -0.0549203867
## NEUMONIA       -0.1065659863
## EDAD            0.0617301995
## EMBARAZO        0.0258210042
## DIABETES       -0.0273842379
## EPOC            0.0197898504
## ASMA            0.0085639517
## INMUSUPR        0.0320811689
```

```
## HIPERTENSION   -0.0406699506
## OTRA_COM        0.0162155578
## CARDIOVASCULAR  0.0062619076
## OBESIDAD       -0.0601384398
## RENAL_CRONICA   0.0368244174
## TABAQUISMO     -0.0152298203
## OTRO_CASO       0.0007982205
## RESULTADO      -0.0268519480
## UCI            -0.0488257833
## MURIO          -0.0491680735
## DIAS_INGRESO    1.0000000000
```

Para la regresión lineal primeramente se obtuvo un mapa de correlación entre los atributos, a simple vista se puede observar que los padecimientos de salud se encuentrar relacionados entre sí, por lo que se deduce que un padecimiento pudiera conducir a otro, y entre más problemas de salud se tengan, es más probable que un contagio por Covid-19 sea grave.

```
muerte_model <-
  lm(
    MURIO ~ TIPO_PACIENTE + EDAD + NEUMONIA + RESULTADO + SECTOR + INTUBADO + DIABETES + SEXO + HIPERTE
    data = train_data
  )

summary(muerte_model)
```

```
##
## Call:
## lm(formula = MURIO ~ TIPO_PACIENTE + EDAD + NEUMONIA + RESULTADO +
##     SECTOR + INTUBADO + DIABETES + SEXO + HIPERTENSION + OBESIDAD +
##     RENAL_CRONICA + INMUSUPR + TABAQUISMO, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03591 -0.14732 -0.01181  0.10808  1.26099
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.2734834  0.0625953  20.345  < 2e-16 ***
## TIPO_PACIENTE -0.2868593  0.0138294 -20.743  < 2e-16 ***
## EDAD          -0.0050555  0.0001628 -31.062  < 2e-16 ***
## NEUMONIA       0.1736545  0.0072322  24.011  < 2e-16 ***
## RESULTADO      0.0679643  0.0032024  21.223  < 2e-16 ***
## SECTOR         0.0168414  0.0008266  20.375  < 2e-16 ***
## INTUBADO       0.1326007  0.0104383  12.703  < 2e-16 ***
## DIABETES       0.0352760  0.0069107   5.105 3.36e-07 ***
## SEXO          -0.0419849  0.0051509  -8.151 3.90e-16 ***
## HIPERTENSION   0.0103986  0.0066609   1.561   0.1185
## OBESIDAD       0.0144630  0.0065157   2.220   0.0265 *
## RENAL_CRONICA  0.0492500  0.0125128   3.936 8.32e-05 ***
## INMUSUPR       0.0023697  0.0150456   0.158   0.8749
## TABAQUISMO    -0.0212581  0.0088568  -2.400   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3031 on 14503 degrees of freedom
```

```
## Multiple R-squared:  0.623,  Adjusted R-squared:  0.6226
## F-statistic:  1843 on 13 and 14503 DF,  p-value: < 2.2e-16
```

```
test_data$predicted <- predict(muerte_model, test_data)

summary(test_data$predicted - test_data$MURIO)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.184899 -0.109357  0.011833 -0.001907  0.149234  1.176765
```

El modelo de regresión se construyo asignando valores numéricos a las categorías, y aunque este pudiera no ser el mejor enfoque, se obtuvo un valor r cuadrada de 0.6217, que si bien no es muy alto, tampoco es inaceptable. Al final se comparan las diferencias entre el valor original y el que se predice y los errores rondan entre -1 y 1, lo cual no dice mucho, ya que las categorías tienen valor numérico de 1, 2, 3, etc. Simplemente esto se interpreta como el cambio de clase que puede resultar de un error.

Conclusiones

El análisis de datos es una herramienta muy útil para la vida cotidiana actual, donde a cada momento se está generando una cantidad enorme de información. Darle a los datos un procesamiento permite encontrar patrones, relaciones y comportamientos que pueden ayudar a la toma de decisiones y al entendimiento de algún fenómeno.

Como complemento a la tarea anteriormente descrita, los algoritmos de Machine Learning permiten automatizar los cálculos que hay que llevar a cabo para la formulación de conclusiones, cabe destacar que sin este tipo de herramientas sería muy complicado analizar cantidades grandes de datos, ya que la velocidad de procesamiento de los equipos de cómputo permiten llevar a cabo estos análisis de una forma rápida y exacta.

El campo de la medicina es una de las mayores áreas de interés para el análisis de datos, ya que se busca aprovechar las bondades de las ciencias computacionales para mejorar la atención médica que se otorga, incluso para obtener información que a simple vista, sería imposible coseguir.

La implementación de herramientas de análisis de datos aplicadas a situaciones como la actual pandemia por Covid-19, puede ayudar a entender mejor el comportamiento de la enfermedad en un sector de la población. En el presente trabajo se estudiaron los datos correspondientes al seguimiento de los pacientes en México, cuyas características particulares conllevan a que la gravedad de la enfermedad sea diferente en este país que en algún otro.

Por medio de los algoritmos de clasificación utilizados, en especial, del algoritmo C5.0 para árboles de decisión, fue posible determinar con un porcentaje de éxito importante (al rededor del 90%), si un paciente que presentó síntomas de Covid-19, así como con el estudio de otros factores relacionados a su salud, morirá o no.

Trabajo futuro

En relación al Covid-19 se abre un mar de posibilidades de estudio, en este trabajo se hizo hincapié en los aspectos de salud de un paciente para determinar si morirá o no, pero el descenlace de la enfermedad no se limita a ésto, sino que hay más variables que se pueden analizar para obtener mejores resultados.

Un objetivo de interés es el poder estimar el tiempo en el que un paciente se recuperará o morirá, hecho que ayudaría mucho a la implementación de la logística que se lleva a cabo para sel combate de la enfermedad en cuestión.

Por otro lado, sería conveniente probar con más tipos de algoritmos, sobre todo, determinar qué información arrojaría si se aplican redes neuronales a estos datos. Análisis que por cuestiones de tiempo y del equipo de cómputo con el que se cuenta, fue difícil incluirlo en este trabajo.