Contents

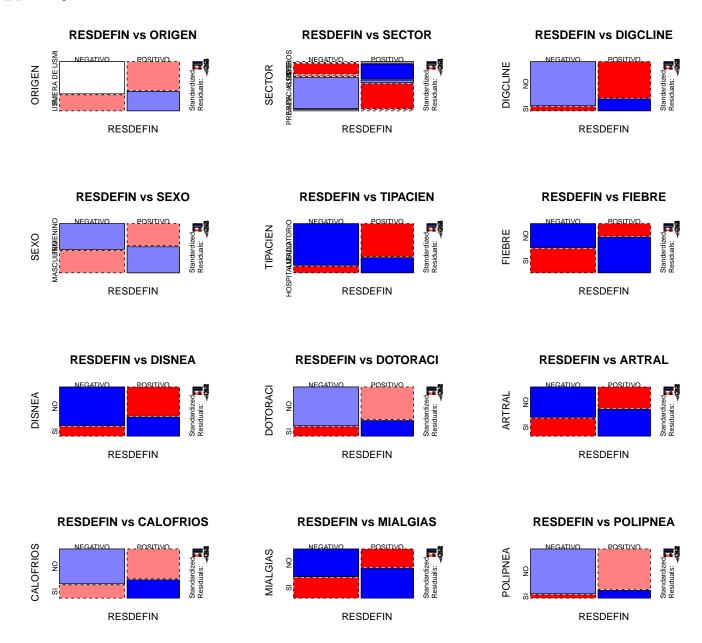
| Manipulación de los datos | 2 |
|---------------------------|---|
| RESDEFIN | 2 |
| Positivos a Covid | 3 |
| Evolución | 4 |
| Descriptivo | 5 |
| Resultados | 7 |
| Conclusión | 8 |

Manipulación de los datos

Para el análisis que se llevará a cabo en este proyecto, además del preprocesamiento que se hizo en clase sobre las variables de ocupación, sector y resultado definitivo se imputarán los "SE IGNORA" de la mayoría de variables dicotómicas de manera aleatoria con probabilidad justa, además creamos la variable que mide los días transcurridos entre el inicio de síntomas y el ingreso al sistema.

RESDEFIN

##Descriptivo



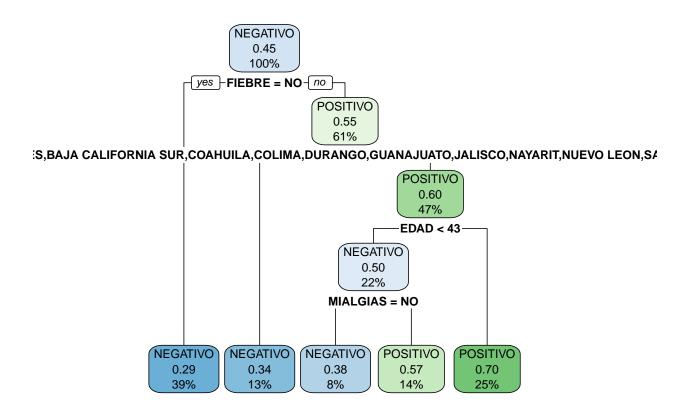
Positivos a Covid

Al intentar explicar nuestra variable de Resultado definitivo utilizaremos las variables de origen, sector, cveentuni, entidad, sexo, tipo de paciente, edad, ocupación, si es o habla indígena, vacunación y todos los síntomas.

Además eliminamos todos los registros donde nuestra variables de interés tenía datos faltantes.

Primero explicaremos con los métodos de regresión logística y un árbol de decisión, ya que estos nos pueden dar un nivel de importancia a las variables.

El árbol resultante es el siguiente:



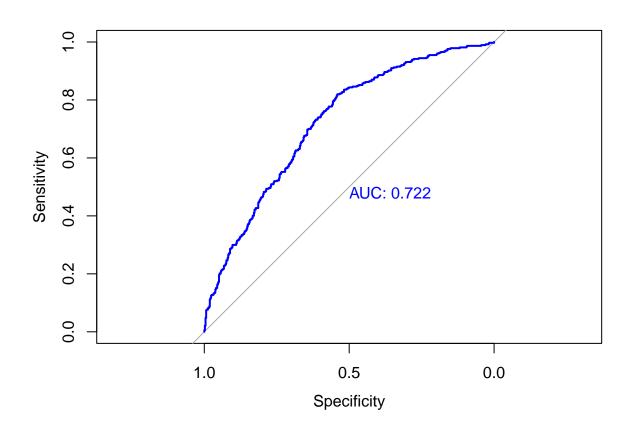
Consideraremos las seis variables más importantes para el árbol y aquellas en el mejor modelo con el método de Akaike (AIC) para la regresión logística, luego usaremos la unión de estos dos conjuntos de variables importantes para hacer nuestra selección.

Estas variables son:

```
[1] "FIEBRE"
                      "ENTIDAD"
                                   "EDAD"
##
                                                "MIALGIAS"
                                                             "TOS"
                                                                          "ARTRAL"
    [7]
        "ORIGEN"
                      "SECTOR"
                                   "SEXO"
                                                "TIPACIEN"
                                                             "DIGCLINE"
                                                                          "ESINDIGE"
        "HABLEIND"
                      "OCUPACIO"
                                   "ODINOGIA"
                                                "DISNEA"
                                                                          "DIARREA"
   [13]
                                                             "IRRITABI"
   [19]
        "DOTORACI"
                      "CALOFRIOS"
                                  "CEFALEA"
                                                "ATAEDOGE"
                                                             "RINORREA"
                                                                          "POLIPNEA"
##
   [25]
        "VOMITO"
                      "DOLABDO"
                                   "CONJUN"
                                                "CIANOSIS"
                                                                          "ANOSMIA"
##
                                                             "INISUBIS"
       "DISGEUSIA"
                     "VACUNADO"
   Г317
```

Así con nuestras variables reducidas y haciendo 5- Folds probaremos los métodos de regresión logística, máquinas de soporte vectorial, bosques aleatorios y el clasificador bayesiano ingenuo cuyos resultados se

mostrarán posteriormente, pero se puede observar la curva ROC de sensibiliad/especificidad con el método de regresión logística

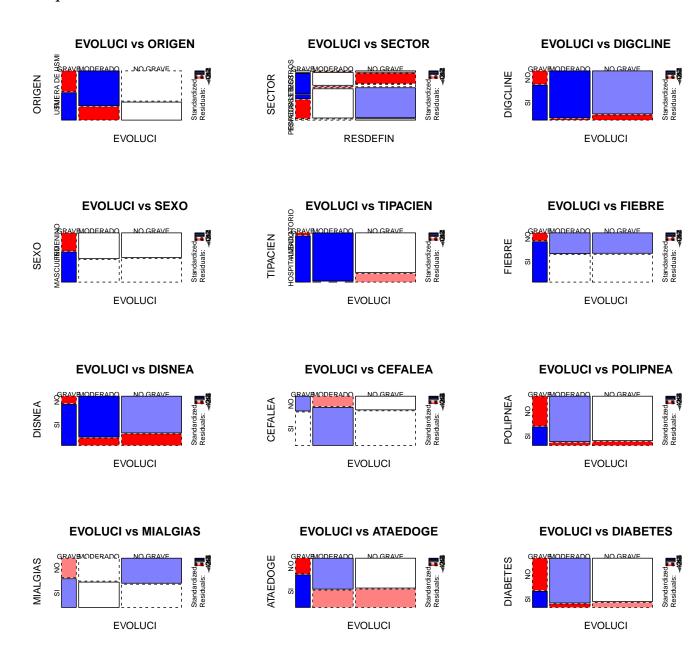


Evolución

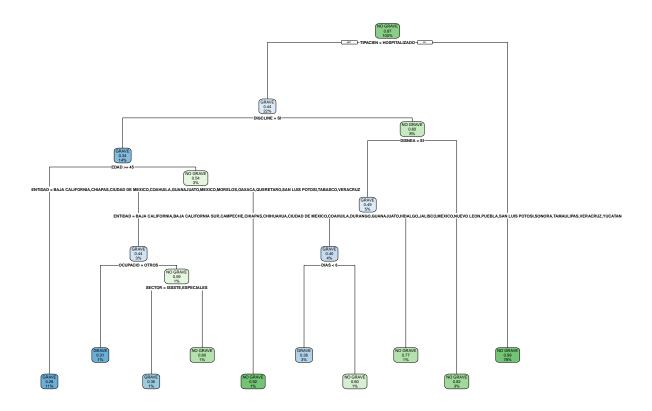
Por otro lado, también trataremos de explicar la variable evolución, la cual la dividiremos en 2 y 3 categorías, GRAVE y NO GRAVE, y GRAVE, MODERADO y NO GRAVE respectivamente.

Las variables explicativas son las mismas que para el caso anterior además de las comorbilidades Al usar sólo dos categorías haremos los mismo que con la variable **RESDEFIN**, primero usar regresión logística y un árbol, seleccionar variables y posteriormente usar los cuatro métodos.

Descriptivo



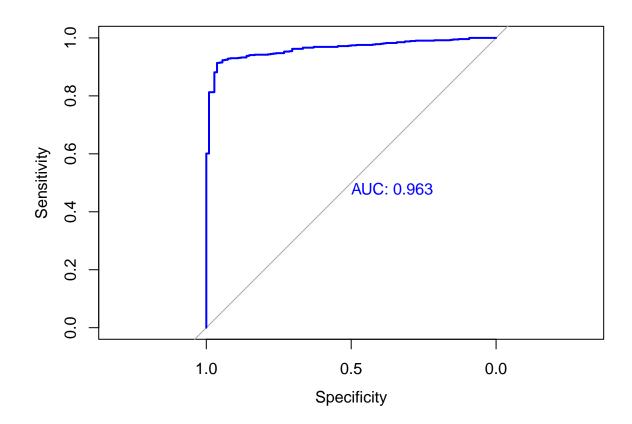
##Metodología



Las variables seleccionadas son:

```
## [1] "TIPACIEN" "DIGCLINE" "DISNEA" "POLIPNEA" "EDAD" "ENTIDAD" ## [7] "SEXO" "OCUPACIO" "ANOSMIA" "VACUNADO"
```

Podemos ver la curva ROC para la regresión logística con las variables ya seleccionadas



Y al tener tres categorías usaremos las mismas variables seleccionadas anteriormente con los cuatro clasificadores

Resultados

```
## RESDEFIN 0.6690224 0.6822069 0.6605470 0.6647838
## EVOLUCION BIN 0.9218441 0.9296127 0.9140771 0.9213732
## EVOLUCION TRI 0.6007425 0.6181709 0.5899832 0.5863895
```

Finalmente, habiendo analizado que en general los métodos se comportan de manera similar, mostraremos una matríz de confusión para cada problema de clasificación con diferentes métodos: logístico para resultado definitivo, SVM para evolución con dos categorías y NB para evolución con tres categorías.

```
## Reference
## Prediction NEGATIVO POSITIVO
```

```
##
     NEGATIVO
                     347
                               169
##
     POSITIVO
                     126
                               208
##
              Reference
## Prediction GRAVE NO GRAVE
##
     GRAVE
                  79
     NO GRAVE
                  29
                           707
##
##
              Reference
## Prediction GRAVE MODERADO NO GRAVE
##
     GRAVE
                  95
                             1
                                       45
##
     MODERADO
                   0
                           166
                                     166
     NO GRAVE
                           139
                                     225
##
                  13
```

Donde la matriz de confusión para resultado definitivo tiene una tasa de error importante en ambas clases, pero la mayoría sigue cayendo en donde supondríamos. En la de tres clases de evolución se confunden mucho entre moderados y graves. En la de dos clases las matrices de confusión parecen comportarse bien. Para todos los métodos las matrices de confusión tienen un comportamiento similar.

Conclusión

De acuerdo a los resultados anteriores, podemos observar que tenemos una mejor precisión de clasificación de los métodos usados para el modelo Evolución con dos categorías que para el modelo con 3 categorías, esto puede deberse al desequilibrio en la baja cantidad de casos graves en comparación con los no graves, por lo que el acierto de clasificación aumentaría.

Debido a los buenos resultados al menos para un modelo, entonces hemos encontrado bastante útil la aplicación de los diferentes métodos de clasificación supervisada para un problema con una muestra relativamente grande y con datos reales.