Proyecto final, aprendizaje automático supervisado.

Harold Andres Romero Lopez, María Paula Morales Rodríguez, Daniela Alejandra Paternina Avilez

r Sys.Date()

1. Introducción

De acuerdo a la Universidad Veracruzana, diariamente, se crean más 2.5 bytes de datos de diversas fuentes y se espera que para el 2025 se supere el total de 180 zettabytes . Estos grandes volúmenes de datos (Big Data) permiten a empresas, compañías e industrias analizar y procesar dicha información, esto se traduce en una herramienta que permite determinar tendencias, evaluar la reacción del público (Universidad de Alcalá,2018), tomar decisiones de manera más rápida, segura y concisa en diversas áreas de aplicación como por ejemplo la salud, además de otras utilidades.

La ciencia de datos cobra relevancia al combinar herramientas, tecnología y diferentes metodologías para extraer datos y generar información significativa a partir de ellos (Amazon Web Service, s. f.). Es allí donde el aprendizaje supervisado cumple un papel fundamental al momento de poner en práctica estos avances tecnológicos, ya que es una subcategoría del machine learning y la inteligencia artificial que usa conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen o predigan resultados de forma precisa (IBM, s. f.). En la ingeniería de aprendizaje automático, se busca investigar, construir, diseñar y desarrollar sistemas de aprendizaje supervisado que utilicen un conjunto de datos de entrenamiento para enseñar a los modelos a generar salidas deseadas y datos de prueba que determinen la eficacia del modelo creado(De Ceupe, 2022).

2. Marco teórico

El aprendizaje supervisado es una técnica usada en exploración de datos, en la que se genera una función de pronóstico a partir del entrenamiento previo de datos. Se dice que es supervisado porque, antes debe existir una clasificación o etiquetado de los datos que es lo que aporta el conocimiento. El proceso habitual consiste en dividir la muestra en dos conjuntos, uno de entrenamiento y otro de prueba, con los datos de entrenamiento ordenados convenientemente se obtendrá un conjunto de pares de entrada-salida. La salida es la variable dependiente, y las entradas son las variables que usaremos para pronosticar la variable dependiente. Es decir, la salida es lo que se quiere pronosticar. (Villalba. F (s/f))

Existen diversos algoritmos de aprendizaje supervisado, pero de acuerdo al tipo de variable que se maneje se pueden dividir en dos: Cuando la variable sea discreta se llamará clasificación y cuando la variable sea continua se llamará de regresión.

- Clasificación

La clasificación utiliza un algoritmo para asignar con precisión datos de prueba en categorías específicas. Reconoce entidades específicas dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo esas entidades deben etiquetarse o definirse. Los algoritmos de clasificación comunes son clasificadores lineales, máquinas de vectores de soporte, árboles de decisión, k-NN y bosques aleatorios.

- Regresión

La regresión se utiliza para comprender la relación entre variables dependientes e independientes, se utiliza comúnmente para hacer proyecciones, como los ingresos por ventas de un negocio determinado. Regresión lineal, regresión logística y regresión polinomial son algoritmos de regresión popular.

Algoritmos de aprendizaje supervisado

- Regresión

lineal La regresión lineal es utilizada para identificar la relación entre una variable dependiente y una o más variables independientes, y normalmente se aplica para hacer predicciones sobre resultados futuros. Cuando solo hay una variable independiente y una variable dependiente, se conoce como regresión lineal simple. A medida que aumenta el número de variables independientes, se denomina regresión lineal múltiple. Para cada tipo de regresión lineal, esta clasificación busca trazar una línea de mejor ajuste, que se calcula mediante el método de mínimos cuadrados. (IBM.(s/f)).

- Algoritmo k-NN (k-Nearest Neighbour Classification)

El algoritmo k-NN reconoce patrones en los datos sin un aprendizaje específico, el cual consiste en medir la distancia entre grupos de datos. Para crear el modelo es necesario cargar el paquete "class" y usar la función knn() que realiza la clasificación. La idea principal del modelo es que a partir de un conjunto de datos de entrenamiento se pueda deducir el agrupamiento de los datos.

- Regresión logística

Es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica la cual puede adoptar un número limitado de categorías en función de las variables predictoras. Este modelo de pronóstico se usa normalmente en variables que se distribuyen en forma de binomial, es decir que simplemente tratan de decir si algo es 1 que significa SI o 0 que significa NO.

- Árboles de decisión

Un árbol de decisión es una estructura ramificada que muestra las diferentes opciones y sus consecuencias. Los puntos en los que hay que tomar decisiones se muestran como nodos, las ramas unen estos nodos y las últimas decisiones son las hojas.

- Bosques aleatorios de decisión (random forest)

Si se aplica de manera repetitiva el algoritmo de árboles de decisión con diferentes parámetros sobre los mismos datos, se obtendrá un bosque aleatorio de decisión. Este modelo consiste en construir diferentes conjuntos de entrenamiento y prueba sobre los mismos datos, lo que genera diferentes árboles de decisión sobre los mismos datos. La unión de estos árboles de diferentes complejidades y con datos de origen distinto aunque del mismo conjunto resulta un bosque aleatorio. (Villalba. F (s/f))

- Máquina de soporte vectorial (SVM)

Una máquina de vectores de soporte se utiliza tanto para la clasificación como para la regresión de datos, el modelo se basa en la construcción de un hiperplano donde la distancia entre dos clases de puntos de datos es máxima. Este hiperplano se conoce como el límite de decisión, que separa las clases de puntos de datos en ambos lados del plano.(IBM.(s/f)).

Training data vs. Testing data

La principal diferencia entre los datos de entrenamiento y los datos de prueba es que los datos de entrenamiento son el subconjunto de datos originales que se utiliza para entrenar el modelo de aprendizaje automático, mientras que los datos de prueba se utilizan para verificar la precisión del modelo. El conjunto de datos de entrenamiento es generalmente de mayor tamaño en comparación con el conjunto de datos de prueba. Las proporciones generales de división de conjuntos de datos de entrenamiento y prueba son 80-20, 70-30 o 90-10. (JavaTpoint (s/f))

Overfitting y Underfitting

El overfitting o sobreajuste es un fenómeno que hace que un algoritmo predictivo presente un bajo porcentaje de acierto en sus resultados, ofreciendo previsiones con una alta varianza. Esto sucede si la muestra utilizada en el entrenamiento del modelo:

- Es poco representativa de la realidad con la que se tendrá que enfrentar después el algoritmo.

- Incluye demasiadas variables, e incluso variables irrelevantes, que confunden al modelo y le impiden identificar la tendencia subyacente.
- Se ha sobrepasado el umbral óptimo de épocas (número de veces que el modelo procesa los mismos datos de entrada en el training).

Por oposición al overfitting se tiene a el underfitting o el desajuste, el cual genera una escasa fiabilidad en las predicciones del modelo. El underfitting o desajuste quiere decir que los datos de entrada son insuficientes o escasa información para lo que se pretende deducir. (BETWEEN. (2020))

Cross-Validation

El cross-validation o validación cruzada es un método que permite probar el rendimiento de un modelo predictivo, después de entrenar un modelo de Machine Learning con datos etiquetados, se supone que tiene que funcionar con nuevos datos, sin embargo es importante garantizar la exactitud de las predicciones del modelo en producción. para poder determinar si aún falta por ajustarlo, se ha ajustado de más o está "bien generalizado". Para probar la eficacia de un modelo se utiliza el "cross-validation" o validación cruzada. Este método también es un procedimiento de remuestreo que permite evaluar un modelo incluso con datos limitados. (Datascientest (2022)).

3. Metodológia

Metodologia

Para el desarrollo del presente trabajo utilizamos el paquete `DynamicCancerDriverKM` , para ello realizamos la instalación del paquete en el repositorio en github `https://github.com/AndresMCB/DynamicCancerDriverKM`en su interior se encuentran las instrucciones de instalación (por favor referirse al repositorio para mas información)

Posteriomente se construye una matriz de expresión genetica unificada, donde combinamos las matrices con pacientes con `Primary Tumor` (Tumor detectado) y pacientes con `Solid Tissue Normal` (Tejido normal), manteniendo la varibale `sample_type` la cual sera nuestra variable clase, se eliminan las variables que no contienen la variable clase y los genes a analizar, estos pasos fueron realizados con el codigo:

```
Matriz_Normal_Tumor <-
  merge(DynamicCancerDriverKM::BRCA_normal,
                              DynamicCancerDriverKM::BRCA_PT,all = TRUE)

Matriz_Normal_Tumor <-
  Matriz_Normal_Tumor[,-c(1:3)]
Matriz_Normal_Tumor <-
  Matriz_Normal_Tumor[,-c(2:4)]
```

Se filtra el dataset verificando cuales de los genes (variables predictoras) estan activas por cada una de las muestras, para determinar el umbral que define si la muestra esta activa, se hallo el valor maximo en el dataset, resultando un valor de 7032374, sobre este valor se definio como umbral el 0.02% del valor maximo, para concluir en dicho porcentaje, se tomo como base que al realizar el primer filtro, debian quedar entre 7.000 y 14.000 genes, se realiza con el codigo:

```
Valor_Maximo <- max(Matriz_Normal_Tumor[,2:23688], na.rm=FALSE)*0.0002
```

Binarizamos el valor que se expresa en cada Gen por observacion, donde se deja un 0 para aquellos genes que no se encuetren activos (que no superaron el umbral) se deja un 1 para los genes que superan el umbral y que se encuentran activos.

Eliminamos los genes que no se encuentren expresados en almenos el 20% de las observaciones, resultando 9122 genes que se encuentran expresados en almenos el 20% de las muestras

En la matriz PPI, clasificamos los 100 genes que mas se repiten, sumando la cantidad de veces que se repite cada Gen tanto en la columna `Input-node Gene Symbol`como en la columna `Output-node Gene Symbol`, se clasifican de manera descendente (de mayor a menor) y se toma el top 100 de muestras, esto se realiza con el codigo, estos genes seran los predictores de los modelos a implementar:

```
Agrupacion <- PPI %>% group_by(`Input-node Gene Symbol`) %>%
  summarise(NN = n()) %>%
  rename(`Gen`=`Input-node Gene Symbol`)

Agrupacion2 <- PPI %>% group_by(`Output-node Gene Symbol`) %>%
  summarise(NN = n()) %>%
  rename(`Gen`=`Output-node Gene Symbol`)

tabla_combinada <- bind_rows(Agrupacion, Agrupacion2) %>%
  group_by(`Gen`) %>%
  summarise(NN = sum(NN)) %>%
  arrange(desc(NN)) %>%
  top_n(101, NN)
```

Posteriormente eliminamos de la Matriz que incluye pacientes con tumor y sin tumor, aquellos genes que no se encuentran incluidos en el top 100 de genes de la Matriz PPI, resultando la `Matriz_final` con la que implementaremos nuestros modelos

4. Resultados y discusión

Modelo k-NN

Se observa que el modelo KNN presenta mayor rendimiento al implementar la normalización de mínimos y máximos (Fig. 1) que al implementar la estandarización z score .

```
confusionMatrix(knnPredict, test.data$sample_type)
```

```
## Confusion Matrix and Statistics
##
##                      Reference
## Prediction            Primary Tumor Solid Tissue Normal
##   Primary Tumor                 332                   5
##   Solid Tissue Normal             5                  24
##
##                Accuracy : 0.9727
##                  95% CI : (0.9503, 0.9868)
##     No Information Rate : 0.9208
##     P-Value [Acc > NIR] : 2.566e-05
##
##                   Kappa : 0.8127
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9852
##             Specificity : 0.8276
##          Pos Pred Value : 0.9852
##          Neg Pred Value : 0.8276
##              Prevalence : 0.9208
##          Detection Rate : 0.9071
##    Detection Prevalence : 0.9208
##       Balanced Accuracy : 0.9064
```

```
##
##        'Positive' Class : Primary Tumor
##
```

De acuerdo a la Fig. 1 se observa que el modelo se entrenó con 853 muestras, contaba con 101 variables predictoras y la variable clase presentó 2 niveles: “Primary Tumor” y “ Solid Tissue Normal”. El modelo presentó una precisión del 97% y su valor kappa fue de 0.84, este modelo detecta de mejor manera los verdaderos positivos teniendo en cuenta que su sensibilidad es del 99%. Se predijo que 334 casos tendrían un gen activo para un tumor primario, el cual se tiene activo y predijo que 3 tendrían un tumor primario pero no lo tenía. Por otro lado, predijo que 5 casos no tendrían el gen activo (no tendrían tumor) pero si tienen un tumor primario y 24 casos no tendrían tumor y efectivamente no lo tienen.

Modelo Regresión Lineal

Para el modelo se utilizan los mismos datos del modelo anterior, solamente se genera el cambio de los valores de los coeficientes a numéricos. En la primera etapa se obtienen los coeficientes de regresión y la predicción de cada uno de los genes.

Dada la significancia estadística, se realiza nuevamente el modelo con los genes más representativos, menores a 0.01, los cuales, EP300, AR, ESR1, RB1, CSNK2A1, MAPK1, HDAC1, PRKCA, EGFR, SMAD1, MAPK3, CSNK2B, YWHAB, TBP, RELA, SMAD9, PTK2 , JAK2, MYC,HCK, VCL, SKIL, SRF, APP, PDPK1. Donde finalmente se describen 58% de los datos de la data , en la que se pueden explicar con el modelo y se obtiene un error de predicción promedio mínimo de y -0.84658 máximo de 0.75576, lo que indica que se subestima un 0.755 ya que es un valor positivo.

```
ins_model_1
```

```
##
## Call:
## lm(formula = as.numeric(sample_type) ~ ., data = train.data)
##
## Coefficients:
## (Intercept)         TP53       CREBBP        EP300        YWHAG        SMAD3
##   1.048e+00           NA    7.940e-02   -1.455e-01    1.050e-01   -9.237e-02
##        GRB2          SRC           AR         ESR1          RB1      CSNK2A1
##  -2.558e-02   -3.527e-02   -5.367e-02   -2.302e-01    5.864e-02    7.621e-02
##       SMAD2       CDKN1A        MAPK1          FYN        HDAC1        PRKCA
##   5.264e-02   -4.922e-02    1.346e-01   -8.042e-02   -7.055e-02   -4.974e-02
##         TK1         EGFR        SMAD4          JUN      CCDC85B        MAPK6
##   9.232e-03   -1.224e-01    1.946e-02   -4.728e-02   -3.431e-04    5.702e-04
##       GSK3B       PIK3R1        SMAD1         SHC1        TRAF2        YWHAZ
##  -1.654e-01    2.209e-02    8.294e-02    2.792e-02   -1.343e-02   -1.287e-02
##       CASP3        UBE2I          SP1          VIM        ATXN1         SMN1
##   6.805e-05   -1.719e-02    4.545e-02    3.368e-02    1.060e-01   -1.702e-01
##      UBQLN4        MAPK3       PRKACA       TGFBR1       CSNK2B        CALM1
##   4.817e-01    7.090e-02   -3.174e-01    2.284e-02   -1.002e-01   -2.016e-02
##      SETDB1        YWHAB          TBP        BRCA1         RELA       CTNNB1
##   3.741e-02    1.351e-01    3.530e-01    1.219e-02   -6.297e-02    1.793e-02
##         LCK          LYN         RXRA       EEF1A1         AKT1        SMAD9
##   2.886e-01    4.522e-02   -1.965e-02   -2.545e-02   -3.319e-02    3.719e-01
##       ANXA7        STAT3       PTPN11        NCOA1        PLCG1         ACTB
##          NA    5.052e-02   -6.834e-03    4.043e-02    3.603e-02    2.405e-02
##        MDFI        EWSR1         PTK2         RAC1        NFKB1        NR3C1
##          NA    4.546e-02   -3.679e-02   -1.313e-01    3.155e-02   -7.209e-02
##      UNC119         ABL1         DLG4         ATN1        NCOR2         CDK2
##   2.071e-02   -1.972e-02   -7.955e-03   -1.058e-02   -1.571e-02    3.789e-02
##        CHD3        PRKCD         JAK2       MAPK14         TLE1         XRCC6
```

```
##  -9.215e-02   -3.238e-02    1.386e-01    2.951e-02   -4.870e-02           NA
##         CBL         INSR          MYC          PTN       ZBTB16          HCK
##   1.467e-02   -3.797e-03   -6.978e-02    5.812e-02   -2.443e-02   -6.186e-02
##        KAT5          VCL         CAV1         RAF1        STAT1        COPS6
##   3.939e-02    7.414e-02    5.429e-02   -1.767e-02   -3.367e-03    1.383e-02
##       KAT2B        PTPN6         SKIL          SRF        MAPK8          PXN
##  -4.313e-02           NA   -8.640e-02    9.805e-02    2.805e-02   -4.433e-02
##       ACTA1        NCOR1        PDPK1         PIN1        TRAF6
##  -2.675e-03    3.577e-02    6.669e-02   -6.757e-02    3.328e-02
```

```
summary(ins_model_1)
```

```
##
## Call:
## lm(formula = as.numeric(sample_type) ~ ., data = train.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71331 -0.11827 -0.00765  0.08769  0.83807
##
## Coefficients: (5 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.048e+00  4.269e-01   2.456 0.014285 *
## TP53                NA         NA      NA       NA
## CREBBP       7.940e-02  1.485e-01   0.535 0.593145
## EP300       -1.455e-01  2.769e-02  -5.253 1.94e-07 ***
## YWHAG        1.050e-01  6.217e-02   1.689 0.091638 .
## SMAD3       -9.237e-02  1.019e-01  -0.907 0.364847
## GRB2        -2.558e-02  2.489e-02  -1.027 0.304543
## SRC         -3.527e-02  2.568e-02  -1.373 0.170137
## AR          -5.367e-02  2.459e-02  -2.183 0.029361 *
## ESR1        -2.302e-01  5.560e-02  -4.140 3.86e-05 ***
## RB1          5.864e-02  1.962e-02   2.988 0.002898 **
## CSNK2A1      7.621e-02  2.242e-02   3.399 0.000710 ***
## SMAD2        5.264e-02  9.697e-02   0.543 0.587415
## CDKN1A      -4.922e-02  6.652e-02  -0.740 0.459586
## MAPK1        1.346e-01  2.628e-02   5.122 3.85e-07 ***
## FYN         -8.042e-02  4.240e-02  -1.897 0.058213 .
## HDAC1       -7.055e-02  2.255e-02  -3.129 0.001823 **
## PRKCA       -4.974e-02  1.910e-02  -2.604 0.009401 **
## TK1          9.232e-03  2.651e-02   0.348 0.727762
## EGFR        -1.224e-01  2.023e-02  -6.048 2.30e-09 ***
## SMAD4        1.946e-02  1.744e-02   1.116 0.264856
## JUN         -4.728e-02  1.211e-01  -0.391 0.696233
## CCDC85B     -3.431e-04  8.494e-02  -0.004 0.996779
## MAPK6        5.702e-04  2.227e-02   0.026 0.979574
## GSK3B       -1.654e-01  1.652e-01  -1.001 0.317028
## PIK3R1       2.209e-02  1.530e-01   0.144 0.885285
## SMAD1        8.294e-02  1.730e-02   4.793 1.98e-06 ***
## SHC1         2.792e-02  7.145e-02   0.391 0.696094
## TRAF2       -1.343e-02  1.843e-02  -0.729 0.466204
## YWHAZ       -1.287e-02  2.966e-02  -0.434 0.664568
## CASP3        6.805e-05  3.669e-02   0.002 0.998521
## UBE2I       -1.719e-02  7.294e-02  -0.236 0.813729
## SP1          4.545e-02  9.150e-02   0.497 0.619530
```

```
## VIM          3.368e-02  3.410e-02   0.988 0.323655
## ATXN1        1.060e-01  1.386e-01   0.765 0.444788
## SMN1        -1.702e-01  1.061e-01  -1.604 0.109167
## UBQLN4       4.817e-01  2.437e-01   1.977 0.048420 *
## MAPK3        7.090e-02  2.134e-02   3.322 0.000937 ***
## PRKACA      -3.174e-01  2.281e-01  -1.392 0.164388
## TGFBR1       2.284e-02  1.853e-02   1.233 0.218143
## CSNK2B      -1.002e-01  3.523e-02  -2.845 0.004558 **
## CALM1       -2.016e-02  1.825e-02  -1.104 0.269811
## SETDB1       3.741e-02  2.031e-02   1.842 0.065832 .
## YWHAB        1.351e-01  2.211e-02   6.108 1.62e-09 ***
## TBP          3.530e-01  1.259e-01   2.804 0.005173 **
## BRCA1        1.219e-02  2.731e-02   0.446 0.655517
## RELA        -6.297e-02  1.718e-02  -3.666 0.000263 ***
## CTNNB1       1.793e-02  2.288e-02   0.784 0.433532
## LCK          2.886e-01  1.743e-01   1.655 0.098303 .
## LYN          4.522e-02  2.214e-02   2.043 0.041412 *
## RXRA        -1.965e-02  1.354e-01  -0.145 0.884641
## EEF1A1      -2.545e-02  1.865e-02  -1.365 0.172621
## AKT1        -3.319e-02  5.250e-02  -0.632 0.527400
## SMAD9        3.719e-01  1.709e-01   2.176 0.029882 *
## ANXA7               NA         NA      NA       NA
## STAT3        5.052e-02  1.349e-01   0.374 0.708217
## PTPN11      -6.834e-03  2.036e-02  -0.336 0.737232
## NCOA1        4.043e-02  2.918e-02   1.385 0.166336
## PLCG1        3.603e-02  2.134e-02   1.688 0.091745 .
## ACTB         2.405e-02  1.848e-02   1.301 0.193623
## MDFI                NA         NA      NA       NA
## EWSR1        4.546e-02  2.414e-01   0.188 0.850672
## PTK2        -3.679e-02  2.081e-02  -1.768 0.077395 .
## RAC1        -1.313e-01  1.848e-01  -0.710 0.477707
## NFKB1        3.155e-02  1.724e-02   1.830 0.067653 .
## NR3C1       -7.209e-02  4.088e-02  -1.764 0.078214 .
## UNC119       2.071e-02  2.205e-02   0.939 0.348107
## ABL1        -1.972e-02  1.691e-02  -1.166 0.243910
## DLG4        -7.955e-03  4.165e-02  -0.191 0.848575
## ATN1        -1.058e-02  1.639e-02  -0.646 0.518691
## NCOR2       -1.571e-02  1.769e-02  -0.888 0.374930
## CDK2         3.789e-02  5.778e-02   0.656 0.512149
## CHD3        -9.215e-02  5.681e-02  -1.622 0.105160
## PRKCD       -3.238e-02  3.003e-02  -1.078 0.281258
## JAK2         1.386e-01  3.855e-02   3.595 0.000346 ***
## MAPK14       2.951e-02  1.917e-02   1.540 0.124086
## TLE1        -4.870e-02  2.157e-01  -0.226 0.821462
## XRCC6               NA         NA      NA       NA
## CBL          1.467e-02  2.025e-02   0.724 0.469011
## INSR        -3.797e-03  3.930e-02  -0.097 0.923047
## MYC         -6.978e-02  1.758e-02  -3.969 7.90e-05 ***
## PTN          5.812e-02  8.019e-02   0.725 0.468782
## ZBTB16      -2.443e-02  2.083e-02  -1.173 0.241267
## HCK         -6.186e-02  1.913e-02  -3.234 0.001273 **
## KAT5         3.939e-02  1.729e-02   2.278 0.023004 *
## VCL          7.414e-02  2.112e-02   3.511 0.000473 ***
## CAV1         5.429e-02  8.704e-02   0.624 0.532975
```

```
## RAF1        -1.767e-02  1.686e-02  -1.048 0.294847
## STAT1       -3.367e-03  1.889e-02  -0.178 0.858588
## COPS6        1.383e-02  3.061e-02   0.452 0.651527
## KAT2B       -4.313e-02  1.121e-01  -0.385 0.700634
## PTPN6               NA         NA      NA       NA
## SKIL        -8.640e-02  3.047e-02  -2.835 0.004698 **
## SRF          9.805e-02  3.993e-02   2.456 0.014291 *
## MAPK8        2.805e-02  2.003e-02   1.400 0.161774
## PXN         -4.433e-02  1.907e-02  -2.325 0.020335 *
## ACTA1       -2.675e-03  2.124e-01  -0.013 0.989954
## NCOR1        3.577e-02  3.124e-02   1.145 0.252489
## PDPK1        6.669e-02  3.224e-02   2.069 0.038912 *
## PIN1        -6.757e-02  1.067e-01  -0.633 0.526651
## TRAF6        3.328e-02  3.897e-02   0.854 0.393276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2005 on 757 degrees of freedom
## Multiple R-squared:  0.5979, Adjusted R-squared:  0.5475
## F-statistic: 11.85 on 95 and 757 DF,  p-value: < 2.2e-16
```

```
summary(model_excl_Rl)
```

```
##
## Call:
## lm(formula = as.numeric(sample_type) ~ ., data = train.data_Rl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65573 -0.12053 -0.01662  0.08230  0.86663
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98981    0.08004  12.366  < 2e-16 ***
## EP300       -0.11962    0.02548  -4.694 3.13e-06 ***
## AR          -0.03884    0.02090  -1.859 0.063437 .
## ESR1        -0.25679    0.04715  -5.446 6.81e-08 ***
## RB1          0.08298    0.01642   5.053 5.37e-07 ***
## CSNK2A1      0.08081    0.02108   3.833 0.000136 ***
## MAPK1        0.14359    0.02489   5.769 1.13e-08 ***
## HDAC1       -0.06925    0.02096  -3.304 0.000995 ***
## PRKCA       -0.06539    0.01732  -3.776 0.000171 ***
## EGFR        -0.13421    0.01866  -7.190 1.45e-12 ***
## SMAD1        0.07720    0.01558   4.956 8.75e-07 ***
## MAPK3        0.11127    0.01854   6.000 2.94e-09 ***
## CSNK2B      -0.09501    0.03268  -2.908 0.003738 **
## YWHAB        0.14900    0.02006   7.426 2.79e-13 ***
## TBP          0.38660    0.10491   3.685 0.000244 ***
## RELA        -0.06341    0.01561  -4.062 5.32e-05 ***
## SMAD9        0.30732    0.09054   3.394 0.000720 ***
## PTK2        -0.03707    0.01860  -1.992 0.046647 *
## JAK2         0.16078    0.03699   4.346 1.56e-05 ***
## MYC         -0.08854    0.01597  -5.545 3.95e-08 ***
## HCK         -0.06340    0.01707  -3.714 0.000218 ***
## VCL          0.07864    0.01974   3.985 7.35e-05 ***
```

```
## SKIL        -0.09405    0.02810  -3.348 0.000852 ***
## SRF          0.08553    0.03353   2.551 0.010919 *
## PDPK1        0.07974    0.02831   2.817 0.004964 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2033 on 828 degrees of freedom
## Multiple R-squared:  0.548,  Adjusted R-squared:  0.5349
## F-statistic: 41.82 on 24 and 828 DF,  p-value: < 2.2e-16
```

Modelo Regresión Logística

```
Lpredict
```

```
##            3            4           12           20           23           24
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##           28           32           36           43           45           46
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##           48           49           50           53           56           58
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##           61           66           68           72           79           82
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##           83           90           94           97           98          101
## 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00 9.907309e-01
##          104          105          107          111          112          113
## 1.000000e+00 1.000000e+00 1.000000e+00 8.031367e-11 1.000000e+00 1.000000e+00
##          116          120          124          126          130          136
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.877175e-10 1.000000e+00
##          140          141          144          145          150          153
## 9.850803e-01 2.220446e-16 2.220446e-16 1.000000e+00 2.220446e-16 1.000000e+00
##          156          162          163          166          168          171
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          172          184          186          189          192          198
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          199          203          209          213          220          221
## 1.000000e+00 1.000000e+00 1.000000e+00 9.952845e-01 1.000000e+00 1.000000e+00
##          222          227          234          237          238          240
## 1.107266e-09 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          242          246          254          257          258          260
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          262          263          266          267          272          276
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          282          285          296          298          304          305
## 2.220446e-16 3.342329e-03 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          311          312          313          319          321          327
## 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00
##          328          336          343          345          350          351
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          355          356          357          362          363          368
## 1.000000e+00 3.900357e-01 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          369          370          379          382          386          390
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          395          396          399          400          401          403
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00
##          408          411          412          421          423          436
```

```
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          442          443          446          452          453          457
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          458          465          467          468          470          471
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          473          474          477          479          485          490
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          492          494          495          496          502          503
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          504          505          507          508          509          510
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          511          513          514          517          519          525
## 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00 9.935430e-01
##          529          530          531          539          541          542
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          545          547          548          552          553          556
## 1.000000e+00 1.000000e+00 2.220446e-16 9.999406e-01 1.000000e+00 1.000000e+00
##          557          559          564          565          566          569
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00
##          578          579          587          589          592          595
## 1.000000e+00 1.000000e+00 2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00
##          598          605          608          609          620          621
## 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 2.220446e-16 1.000000e+00
##          623          634          645          646          657          658
## 2.220446e-16 1.000000e+00 1.000000e+00 4.926075e-10 1.649425e-11 1.000000e+00
##          661          663          665          667          671          672
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16
##          673          677          679          683          684          685
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00
##          691          692          693          699          711          719
## 1.000000e+00 1.000000e+00 2.220446e-16 2.220446e-16 1.000000e+00 2.220446e-16
##          720          723          724          730          734          735
## 1.000000e+00 6.129846e-03 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          738          748          749          751          760          762
## 2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          763          765          766          775          776          779
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16
##          784          797          799          806          815          816
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          818          819          821          823          827          828
## 1.000000e+00 1.000000e+00 9.999986e-01 1.000000e+00 1.000000e+00 1.000000e+00
##          830          831          833          834          837          842
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          843          845          847          852          856          860
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          867          868          874          885          890          892
## 1.000000e+00 1.000000e+00 1.000000e+00 1.905913e-07 1.000000e+00 1.000000e+00
##          894          897          902          904          907          908
## 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00 1.198667e-04
##          909          910          911          913          917          922
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          931          932          933          936          938          944
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          946          950          957          958          959          962
```

```
## 9.999989e-01 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00
##          967          968          969          972          977          979
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          983          987          990         1008         1014         1018
## 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1022         1023         1025         1029         1031         1033
## 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16 2.220446e-16 1.000000e+00
##         1034         1043         1045         1046         1048         1053
## 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1059         1061         1062         1065         1067         1069
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1074         1077         1078         1079         1085         1088
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1092         1093         1095         1096         1099         1100
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1103         1105         1110         1114         1115         1118
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1122         1125         1126         1132         1135         1138
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1143         1150         1152         1160         1169         1172
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1174         1175         1176         1178         1179         1181
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##         1186         1188         1191         1198         1199         1201
## 1.000000e+00 1.000000e+00 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00
##         1204         1210         1211         1212         1214         1215
## 1.000000e+00 1.000000e+00 1.000000e+00 9.994911e-01 1.000000e+00 2.220446e-16
## attr(,"non-estim")
##    3    4   12   20   23   24   28   32   36   43   45   46   48   49   50   53
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##   56   58   61   66   68   72   79   82   83   90   94   97   98  101  104  105
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##  107  111  112  113  116  120  124  126  130  136  140  141  144  145  150  153
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
##  156  162  163  166  168  171  172  184  186  189  192  198  199  203  209  213
##   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64
##  220  221  222  227  234  237  238  240  242  246  254  257  258  260  262  263
##   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80
##  266  267  272  276  282  285  296  298  304  305  311  312  313  319  321  327
##   81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96
##  328  336  343  345  350  351  355  356  357  362  363  368  369  370  379  382
##   97   98   99  100  101  102  103  104  105  106  107  108  109  110  111  112
##  386  390  395  396  399  400  401  403  408  411  412  421  423  436  442  443
##  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128
##  446  452  453  457  458  465  467  468  470  471  473  474  477  479  485  490
##  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144
##  492  494  495  496  502  503  504  505  507  508  509  510  511  513  514  517
##  145  146  147  148  149  150  151  152  153  154  155  156  157  158  159  160
##  519  525  529  530  531  539  541  542  545  547  548  552  553  556  557  559
##  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176
##  564  565  566  569  578  579  587  589  592  595  598  605  608  609  620  621
##  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191  192
##  623  634  645  646  657  658  661  663  665  667  671  672  673  677  679  683
##  193  194  195  196  197  198  199  200  201  202  203  204  205  206  207  208
```

```
##  684  685  691  692  693  699  711  719  720  723  724  730  734  735  738  748
##  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224
##  749  751  760  762  763  765  766  775  776  779  784  797  799  806  815  816
##  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240
##  818  819  821  823  827  828  830  831  833  834  837  842  843  845  847  852
##  241  242  243  244  245  246  247  248  249  250  251  252  253  254  255  256
##  856  860  867  868  874  885  890  892  894  897  902  904  907  908  909  910
##  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272
##  911  913  917  922  931  932  933  936  938  944  946  950  957  958  959  962
##  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288
##  967  968  969  972  977  979  983  987  990 1008 1014 1018 1022 1023 1025 1029
##  289  290  291  292  293  294  295  296  297  298  299  300  301  302  303  304
## 1031 1033 1034 1043 1045 1046 1048 1053 1059 1061 1062 1065 1067 1069 1074 1077
##  305  306  307  308  309  310  311  312  313  314  315  316  317  318  319  320
## 1078 1079 1085 1088 1092 1093 1095 1096 1099 1100 1103 1105 1110 1114 1115 1118
##  321  322  323  324  325  326  327  328  329  330  331  332  333  334  335  336
## 1122 1125 1126 1132 1135 1138 1143 1150 1152 1160 1169 1172 1174 1175 1176 1178
##  337  338  339  340  341  342  343  344  345  346  347  348  349  350  351  352
## 1179 1181 1186 1188 1191 1198 1199 1201 1204 1210 1211 1212 1214 1215
##  353  354  355  356  357  358  359  360  361  362  363  364  365  366
```

```
print(confusion_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  24  17
##          1   3 322
##
##                Accuracy : 0.9454
##                  95% CI : (0.9169, 0.9663)
##     No Information Rate : 0.9262
##     P-Value [Acc > NIR] : 0.09285
##
##                   Kappa : 0.6772
##
##  Mcnemar's Test P-Value : 0.00365
##
##             Sensitivity : 0.88889
##             Specificity : 0.94985
##          Pos Pred Value : 0.58537
##          Neg Pred Value : 0.99077
##              Prevalence : 0.07377
##          Detection Rate : 0.06557
##    Detection Prevalence : 0.11202
##       Balanced Accuracy : 0.91937
##
##        'Positive' Class : 0
##
```

El algoritmo en su predicción calculó la probabilidad entre 0 y 1, donde 1 hace referencia a que el gen está activo y 0 el gen no está activo, de acuerdo a la figura anterior, las muestras 2 3 4 5 8 11 13 17 18 22 23 25 26 40 51 52 54 55 56 60 61 63 64 tienen el gen activo mientras que la muestra 66 tiene el gen no activo.

Arbol de decisión

```
head (Matriz_Final)
```

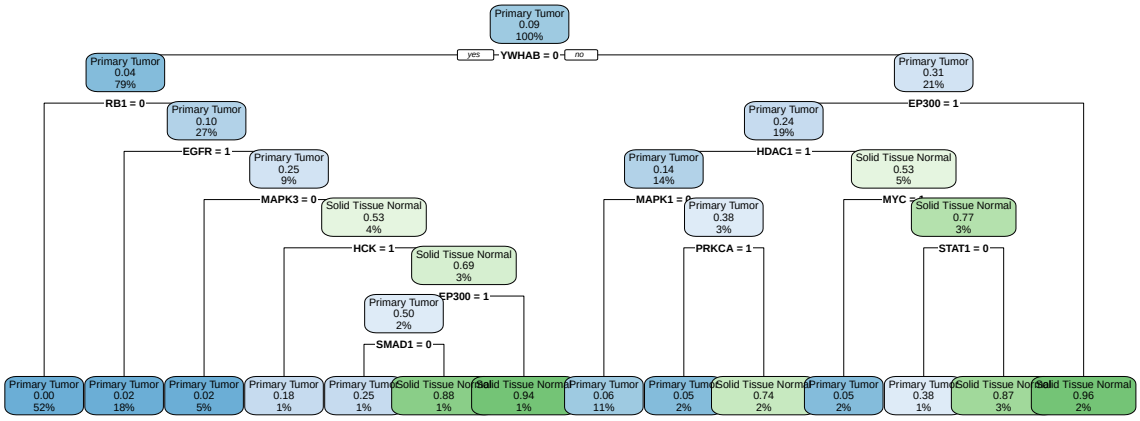
```
##     sample_type TP53 CREBBP EP300 YWHAG SMAD3 GRB2 SRC AR ESR1 RB1 CSNK2A1
## 1 Primary Tumor    0      0     1     1     1    1   0  1    1   0       0
## 2 Primary Tumor    0      0     1     1     1    1   0  0    1   0       1
## 3 Primary Tumor    0      0     0     0     1    0   0  0    1   0       0
## 4 Primary Tumor    0      0     1     1     1    0   0  0    1   0       1
## 5 Primary Tumor    0      0     1     1     1    0   0  0    1   0       1
## 6 Primary Tumor    0      0     1     1     1    0   0  0    1   0       1
##   SMAD2 CDKN1A MAPK1 FYN HDAC1 PRKCA TK1 EGFR SMAD4 JUN CCDC85B MAPK6 GSK3B
## 1     1      1     0   0     0     1   0    0     1   0       1     0     0
## 2     1      1     0   0     1     0   1    0     0   0       1     0     0
## 3     1      1     0   0     0     0   0    0     0   0       1     0     0
## 4     1      1     0   0     1     0   0    0     1   0       1     0     0
## 5     1      1     0   0     1     0   0    1     1   0       1     0     0
## 6     1      1     0   0     1     0   0    1     1   0       1     0     0
##   PIK3R1 SMAD1 SHC1 TRAF2 YWHAZ CASP3 UBE2I SP1 VIM ATXN1 SMN1 UBQLN4 MAPK3
## 1      0     1    0     1     0     1     1   0   1     0    0       0     0
## 2      0     0    0     0     1     1     0   0   1     0    0       0     0
## 3      0     0    0     0     0     1     0   0   1     0    0       0     0
## 4      0     0    0     0     1     1     0   0   1     0    0       0     0
## 5      0     0    0     0     1     1     0   0   1     0    0       0     1
## 6      0     0    0     0     1     1     0   0   1     0    0       0     0
##   PRKACA TGFBR1 CSNK2B CALM1 SETDB1 YWHAB TBP BRCA1 RELA CTNNB1 LCK LYN RXRA
## 1      1      0      1     1      1     1   0     1    1       0   1   1    1
## 2      1      0      1     0      0     0   0     0    0       0   1   1    1
## 3      1      0      0     0      0     0   0     0    0       0   1   0    1
## 4      1      0      1     0      0     0   0     1    1       0   1   0    1
## 5      1      0      1     0      1     0   0     1    1       0   1   1    1
## 6      1      0      1     0      0     0   0     0    0       0   1   1    1
##   EEF1A1 AKT1 SMAD9 ANXA7 STAT3 PTPN11 NCOA1 PLCG1 ACTB MDFI EWSR1 PTK2 RAC1
## 1      0    0     1     1     1      1     0     1    1    0     1    0    1
## 2      0    0     1     1     1      0     0     0    0    0     1    0    1
## 3      0    0     1     1     1      0     0     0    0    0     1    0    1
## 4      0    0     1     1     1      1     0     0    1    0     1    0    1
## 5      1    0     1     1     1      0     0     0    0    0     1    0    1
## 6      0    0     1     1     1      0     0     0    0    0     1    0    1
##   NFKB1 NR3C1 UNC119 ABL1 DLG4 ATN1 NCOR2 CDK2 CHD3 PRKCD JAK2 MAPK14 TLE1
## 1     1     1      1    0    1    1     1    0    0     1    1      1    1
## 2     1     0      0    0    1    0     0    0    0     1    1      0    1
## 3     0     0      0    0    0    0     0    0    0     0    0      0    1
## 4     1     1      1    0    1    0     1    0    0     1    1      1    1
## 5     1     1      1    1    1    0     1    0    0     1    1      1    1
## 6     0     1      1    0    1    0     0    0    0     1    0      1    1
##   XRCC6 CBL INSR MYC PTN ZBTB16 HCK KAT5 VCL CAV1 RAF1 STAT1 COPS6 KAT2B PTPN6
## 1     0   1    0   1   0      1   0    0   0    0    0     0     0     1     1
## 2     0   0    1   1   0      0   1    0   0    0    0     0     0     1     1
## 3     0   0    0   0   0      0   0    0   0    0    0     0     0     1     1
## 4     0   1    0   0   0      0   1    0   0    0    0     0     0     1     1
## 5     0   1    0   0   0      0   1    0   0    0    0     0     0     1     1
## 6     0   0    0   0   0      0   0    0   0    0    1     0     0     1     1
##   SKIL SRF MAPK8 PXN ACTA1 NCOR1 PDPK1 PIN1 TRAF6
## 1    0   1     1   1     0     1     1    0     1
## 2    0   1     1   1     0     1     0    0     1
```

```
## 3    0   0     0   0     0     1     0    0     0
## 4    0   1     0   0     0     1     1    0     1
## 5    0   1     0   0     0     1     1    0     1
## 6    0   1     0   0     0     1     1    0     1
```

```
Matriz_Final[, "sample_type"] <- as.factor(Matriz_Final[, "sample_type"])

modelo_arbol <- rpart(sample_type ~ ., data = Matriz_Final, na.action = na.omit)

rpart.plot::rpart.plot(modelo_arbol, tweak=1.5)
```




```
modelo_arbol$cptable
```

```
##           CP nsplit rel error    xerror       xstd
## 1 0.10619469      0 1.0000000 1.0000000 0.08960588
## 2 0.09292035      2 0.7876106 0.9026549 0.08555519
## 3 0.03982301      4 0.6017699 0.7522124 0.07869293
## 4 0.03097345      6 0.5221239 0.7079646 0.07651147
## 5 0.02654867     10 0.3982301 0.7345133 0.07783016
## 6 0.01769912     12 0.3451327 0.7168142 0.07695438
## 7 0.01000000     13 0.3274336 0.6548673 0.07377994
```

```
Etiquetas_Modelo_Arbol <- predict(modelo_arbol, Matriz_Final[,-1], type="class")
confusionMatrix(Etiquetas_Modelo_Arbol, Matriz_Final[,1])
```

```
## Confusion Matrix and Statistics
##
##                      Reference
```

```
## Prediction            Primary Tumor Solid Tissue Normal
##   Primary Tumor                1094                   25
##   Solid Tissue Normal            12                   88
##
##                Accuracy : 0.9696
##                  95% CI : (0.9584, 0.9785)
##     No Information Rate : 0.9073
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8097
##
##  Mcnemar's Test P-Value : 0.04852
##
##             Sensitivity : 0.9892
##             Specificity : 0.7788
##          Pos Pred Value : 0.9777
##          Neg Pred Value : 0.8800
##              Prevalence : 0.9073
##          Detection Rate : 0.8975
##    Detection Prevalence : 0.9180
##       Balanced Accuracy : 0.8840
##
##        'Positive' Class : Primary Tumor
##
```

##5. Conclusión

Se puede concluir que ambos tipos de modelos tienen sus ventajas y desventajas, los modelos de clasificación son buenos para problemas en los que se necesita predecir una categoría, y los modelos de regresión son buenos para problemas en los que se necesita predecir un valor continuo.

En este caso, el objetivo era predecir si un paciente tiene cáncer o no. Por lo tanto, cualquier modelo de clasificación es un buen candidato para dar una solución. Sin embargo, un modelo de regresión se puede contemplar, de este modo, utilizar un modelo de regresión logística puede predecir la probabilidad de que un paciente tenga o no cáncer.

##6. Referencias

IBM.(s/f) ¿Qué es el aprendizaje supervisado?. Ibm.com. Recuperado el 22 de noviembre de 2023, de https://www.ibm.com/mx-es/topics/supervised-learning

Villalba (), F. (s/f). Aprendizaje supervisado en R. Github.io. Recuperado el 22 de noviembre de 2023, de https://fervilber.github.io/Aprendizaje-supervisado-en-R/

Capítulo 10 Aprendizaje Supervisado. (2020, junio 26). Bookdown.org. https://bookdown.org/dparedesi/data-science-con-r/aprendizaje-supervisado.html

Train and Test datasets in Machine Learning. (s/f). Www.javatpoint.com. Recuperado el 22 de noviembre de 2023, de https://www.javatpoint.com/train-and-test-datasets-in-machine-learning

IT Solutions de BETWEEN. (2020). ¿Qué es el overfitting en machine learning? Between.tech. Recuperado el 22 de noviembre de 2023, de https://impulsate.between.tech/overfitting-machine-learning

Cross-Validation : definición e importancia en Machine Learning. (2022, mayo 13). Formation Data Science | Datascientest.com. https://datascientest.com/es/cross-validation-definicion-e-importancia

Amazon Web Service. (s. f.). ¿Qué es la ciencia de datos? - Explicación de la ciencia de datos - AWS. Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/data-science/

De Ceupe, B. (2022, 28 marzo). Ceupe. Ceupe. https://www.ceupe.com/blog/aprendizaje-supervisado.html

Importancia del Data Science - Máster en Data Science. (2018, 17 junio). Universidad de Alcalá. https://www.master-data-scientist.com/importancia-data-science/

¿Qué es el aprendizaje supervisado? | IBM. (s. f.). https://www.ibm.com/mx-es/topics/supervised-learning

Universidad Veracruzana. (s. f.). Conocimientos generales: ¿Sabes cuántos datos se generan en un minuto? – Seguridad de la información. https://www.uv.mx/infosegura/general/conocimientos_datos/