Parcial 1 - Machine Learning

Norma Yuliana Cala, Lady Alexandra Duarte, Jhon Freddy Puentes

12/12/2020

Etapa 1: Definición del problema

¿Qué se pretende predecir?

Se desea predecir la probabilidad de que un empleado se retire o no de la empresa.

¿De qué datos se dispone?

Disponemos de datos históricos de retiros de empleados en una empresa.

¿Cuál es la métrica de éxito?

Consideraremos como métrica de exito la creación del un modelo que sea capaz de p#predecir minimo el 90% de los casos de retiro.

Etapa 2: Datos

el set de datos `UNION_EMPLEADOS_RETIROS.csv` contiene información sobre los empleados, datos de contratos y observaciones de retiros de empleados en una compañia. Además de esto, incluye las siguientes varibles adicionales:

- `FECHA`:
- `CODIGO_EMPRESA`: código unico de la compañia.
- `ID_GENERO`: Identificador del genero de la persona.
- `ID_ESTADO_CIVIL`: Identificador del estado civil.
- `FECHA_NACIMIENTO`: Fecha de nacimiento el empleado.
- `CODIGO_LOCALIDAD`: Identificador la localidad del empleado.
- `TIPO_NOMINA`: Tipo de vinculación de la persona con la empresa.
- `FECHA_INGRESO`: Fecha de inicio de labores en la compañia.
- `FECHA_INICIO_CONTRATO`: Fecha de inicio de labores en la compañia.
- `ID_CARGO`: Identificador del cargo del empleado.
- `TIPO_CARGO`: Descripción del tipo de cargo.
- `NIVEL_CARGO`: Identificador del nivel de cargo.
- `CATEGORIA_CARGO`: Identificador de la categoría del cargo.
- `TIPO_AREA`: Área de la compañia a la que pertenece o perteneció el empleado.
- `ID_CENTRO_COSTO`: Identificador del centro del costo.
- `SENAL_SINDICALIZADO`: Carácter que indica si es o era de un sindicato.
- `CLASE_EMPLEADO`: Descripción que indica la clase de empleado.
- `CLASE_NOMINA`: Identificador del tipo de vinculación de la persona con la empresa.

- `SUBSIDIO_TRANSPORTE`: Identificador que indica si tiene o tuvo subsidio de transporte.
- `PORCENTAJE_RIESGO`: Porcentaje de riesgo del cargo.
- `SALARIO`: Salario base del empleado.
- `COMPENSACION_VARIABLE`: Valor de compensación variable en caso de que reciba.
- `PORCENT_COMP_VARIABLE`: Valor en % de compensación variable en caso de que reciba.
- `TOTAL_SALARIO`: Total recibido por el empleado. salario + compensación variable.
- `CEDULA_ANONIMIZADA`: Identificador único del empleado en la base de datos.
- `UBICACION_ANONIMIZADA`: Identificador de la ubicación del empleado.
- `CAUSA_NOMBRE`: Descripción de la causa de retiro.
- `TIPO_DE_RETIRO`: Descripción del tipo de retiro del empleado.
- `FECHA_FIN_CONTRATO`: Fecha fin del contrato del empleado.

Etapa Depuración Datos

Para el conjunto de datos de Empleados y Retiros se realizó la siguiente depuración de registros con el consentimiento de que las descripciones siguientes no aportarían al modelo de predicción porque interfieren en las predicciones por ser datos que llevarían a una ppredicción direfente. Total datos = 40.813únicos de toda la base = 1.872* Datos a Eliminar Edad menores de 18 años = 14 Nota: Registros con 2 y 11 años.* Causa nombre Eliminar Fallecidos = 151 Eliminar Pensionados = 411* Tipo Nomina Eliminar Aprendiz Sena = 2.072 Jubilados por la empresa = 297Por último se quitan todos los duplicados 36.289 y se dejamos el registro con el mayor salario para quedarnos con 1.579 registros únicos por cliente.

En la `descripción del procesamiento y arreglo de la información` se describen los procesos relizados a cada uno de los campos donde fue necesario realzar actividades como: homologaciones, cruces de datos entre vacios y los que contenían información para traer los datos correspondientes a cada uno de los registros necesarios para completar la información en las variables, asignación de registros DUMMY entre otros, a continuación se desglozan las actividades realizadas: Tipo texto = NO DEFINIDO Tipo número = 999999

- `Fecha` Se personaliza el formato de las fechas.

  FECHAS

  FORMATO

  2017-12-19 00:00:00 UTC

  dd/MM/yyyy

- `Género` Se encuentran 1.218 registros sin género, se procede a realizar el cruce de datos para encontrar el valor correspondiente para el registro.

- `Estado civil`: Las observaciones para el estado civil vienen de diferentes maneras por lo que se unifica la descripción y se homologan algunos registros. Para valores 0, vacios, ND, no definido se procede a realizar el cruce de datos para encontrar el valor correspondiente para el registro. Los que aún no se encontraron con valor se homologan con NO DEFINIDO. Algunas cédulas tenían asignados los valores de M y F por tal motivo se asignó el valor con mayor número según el conteo.

  VALOR

  ASIGNACIÓN

  cas

  CASADO

  div

  DIVORCIADO

  sep

SEPARADO

sol

SOLTERO

uni

UNION LIBRE

viu

VIUDO

vacios

NO DEFINIDO

- `Código localidad` Para los 2.293 registros con valores 0 y vacíos se realiza la búsqueda de información cruzando los campos vacíos con los que sí tienen información para para traer el valor. Si no hay se dejan como 999999.
- `Fecha Nacimiento` Se saca una nueva variable llamada Edad, restando el año actual con la fecha de nacimiento.
- `Tipo de Nómina` Esta columan de informacióon viene mezclados tanto códigos como descripción del tipo de nomina que tiene. Se hace necesario realizar cruces y trer información para completar la información que está como codficada con números. Lo que no se encuentre en la base se le asigna el valor DUMMY.

  VALOR

  ASIGNACIÓN

  EST UNIVERSITARIO EN PRACTICA

  APRENDIZ SENA

  s,f, I, 1,2,3,4,6,9,15

  NO DEFINIDO

  vacías

  NO DEFINIDO

- `Tipo Cargo` Valores O, N/A, Vacías se realiza la búsqueda de información cruzando los campos vacíos con los que sí tienen información para para traer el valor

  VALOR

  ASIGNACIÓN

  OPERATIVA

  OPERATIVO

  DIRECCIÓN

  DIRECCION

  DIRECTOR

  DIRECTIVO

- `ID Cargo` Se homologan así:

  VALOR

ASIGNACIÓN

0

999999

vacíos

999999

- `Causa Nombre` Algunos nombres se referencian a lo mismo por tal motivo se hacen homologaciones.

  VALOR

  ASIGNACIÓN

  PENSIÒN DE VEJEZ

  PENSION DE VEJEZ

  PENSION DE VEJEZ

  PENSION

  RENUNCIA

  VOLUNTARIA

  TERMINACIÓN UNILATERAL

  TERMINACION UNILATERAL

- `Fecha inicio de contraro` Algunas columnas se encuentran con registros vacios y en la columna de inicio de contrato si registra datos, dicho de esta manera se traen los datos de esta columna para rellenar la columna faltante.

- `Tipo de retiro` Algunos campos se les realiza la homologación.

  VALOR

  ASIGNACIÓN

  DESEADA

  DESEADO

  No Deseada

  NO DESEADO

- `Id Cargo` Valor 0 se asigna el DUMMY de valor 999999

- `Nivel Cargo`

  VALOR

  ASIGNACIÓN

  0

  Vacíos

  -1

  Vacíos

  Vacios

  999999

- `Categoría Cargo`

  VALOR

  ASIGNACIÓN

  0

  Vacíos

  -1

  Vacíos

  Vacios

  999999

- `Tipo Area` Columna de información en donde se realizaron más homologaciones para las observaciones de los registros incluidos.

  VALOR

  ASIGNACIÓN

  -

  Vacíos

  0

  Vacíos

  N/A

  Vacíos

  ADMINISTRAVIVO

  ADMINISTRATIVO

  OPERATIVA

  OPERATIVO

  PRODUCCIÓN

  PRODUCCION

  ADMINISTRACION PLANTA

  PLANTA ADMINISTRATIVO

  MARCAS

  MARCAS DE CANAL

  N/A

  Vacíos

  N/A

  Vacíos

  ADMINISTRATIVA PLANTA

  ADMINISTRATIVO PLANTA

  ADMINISTRACION PLANTA

ADMINISTRATIVO PLANTA

Vacíos

NO DEFINIDO

- `Clase Empleado`

  VALOR

  ASIGNACIÓN

  0,3,4,5

  Vacíos

  Vacios

  999999

- `Porcentaje Riesgo` Para los valores que no tienen datos se les asigna el valor DUMMY correspondiente al tipo de dato en este caso 999999.
- `Total Salario` Se calcula el total salario con el calculo se define sumando las columnas de Salario más Compensación variable.
- `Retiro` Para los valores que se encontraron como vaciós se realiza la validación de que no tienen fecha de retiro ni causa de retiro y se les asigna el valor 0.

Lectura de datos

```
datos <- read.csv(file = 'UNION_EMPLEADOS_RETIROS.csv', sep = ';')
```

Análisis exploratorio

```
library(skimr)
```

```
#skim(datos) # Comentar para generar PDF
```

En el siguiente resultado podrá observar las principales metricas de los datos, de una forma estructurada y compacta.

```
head(datos, 8)
```

Veamos algunos ejemplos de los datos

```
##   CEDULA_ANONIMIZADA      FECHA CODIGO_EMPRESA ID_GENERO ID_ESTADO_CIVIL
## 1                 50 31/12/2017             21         F         SOLTERO
## 2                158 31/12/2017             21         F         SOLTERO
## 3                215 31/12/2017             21         M         SOLTERO
## 4                279 31/12/2017             21         F         SOLTERO
## 5                319 31/12/2017             21         F         SOLTERO
## 6                376 31/12/2017             21         F         SOLTERO
## 7                382 31/12/2017             21         M     UNION LIBRE
```

```
## 8               390 31/12/2017            21         F          CASADO
##   FECHA_NACIMIENTO EDAD CODIGO_LOCALIDAD TIPO_NOMINA FECHA_INGRESO
## 1       14/08/1998   22           999999 NO DEFINIDO     1/06/2017
## 2       21/08/1997   23           999999 NO DEFINIDO    16/06/2017
## 3       12/04/1999   21           999999 NO DEFINIDO    19/04/2017
## 4       15/07/1995   25           999999 NO DEFINIDO     4/07/2017
## 5       21/01/1994   26           999999 NO DEFINIDO    16/01/2017
## 6       13/12/1998   22           999999 NO DEFINIDO     3/04/2017
## 7       26/03/1988   32           999999 NO DEFINIDO     1/08/2017
## 8       29/05/1992   28           999999 NO DEFINIDO    16/06/2017
##   FECHA_INICIO_CONTRATO ID_CARGO TIPO_CARGO NIVEL_CARGO CATEGORIA_CARGO
## 1             1/06/2017   999999    SOPORTE           7               7
## 2            16/06/2017   999999    SOPORTE           7               7
## 3            19/04/2017   999999    SOPORTE           7               7
## 4             4/07/2017   999999    SOPORTE           7               7
## 5            16/01/2017   999999    SOPORTE           7               7
## 6             3/04/2017   999999    SOPORTE           7               7
## 7             1/08/2017   999999    SOPORTE           7               7
## 8            16/06/2017   999999    SOPORTE           7               7
##      TIPO_AREA ID_CENTRO_COSTO SENAL_SINDICALIZADO CLASE_EMPLEADO CLASE_NOMINA
## 1 PRACTICANTES          110109                   N    NO DEFINIDO            0
## 2 PRACTICANTES          110131                   N    NO DEFINIDO            0
## 3 PRACTICANTES          110109                   N    NO DEFINIDO            0
## 4 PRACTICANTES          110148                   N    NO DEFINIDO            0
## 5 PRACTICANTES          310111                   N    NO DEFINIDO            0
## 6 PRACTICANTES          110109                   N    NO DEFINIDO            0
## 7 PRACTICANTES          110109                   N    NO DEFINIDO            0
## 8 PRACTICANTES          110401                   N    NO DEFINIDO            0
##   SUBSIDIO_TRANSPORTE PORCENTAJE_RIESGO SALARIO COMPENSACION_VARIABLE
## 1                   N            999999  737717                    0
## 2                   N            999999  737717                    0
## 3                   N            999999  737717                    0
## 4                   N            999999 1475434                    0
## 5                   N            999999 1475434                    0
## 6                   N            999999  737717                    0
## 7                   N            999999  737717                    0
## 8                   N            999999  737717                    0
##   PORCENT_COMP_VARIABLE TOTAL_SALARIO UBICACION_ANONIMIZADA FECHA.FIN.CONTRATO
## 1                     0        737717                     7         31/12/2040
## 2                     0        737717                     7         31/12/2040
## 3                     0        737717                     7         31/12/2040
## 4                     0       1475434                     5         31/12/2040
## 5                     0       1475434                     5         31/12/2040
## 6                     0        737717                     7         31/12/2040
## 7                     0        737717                     7         31/12/2040
## 8                     0        737717                     7         31/12/2040
##   CAUSA.NOMBRE TIPO.DE.RETIRO RETIRO
## 1   NO APLICA      NO APLICA      0
## 2   NO APLICA      NO APLICA      0
## 3   NO APLICA      NO APLICA      0
## 4   NO APLICA      NO APLICA      0
## 5   NO APLICA      NO APLICA      0
## 6   NO APLICA      NO APLICA      0
## 7   NO APLICA      NO APLICA      0
```

```
## 8    NO APLICA      NO APLICA      0
```

Número de observaciones y valores ausentes

```
# datos %>% map_dbl(.f = function(x){ sum(is.na(x)) })
```

La base de datos no tiene valores ausentes en ninguna observación.

Etapa 3: Pre-procesamiento

Decidimos estudiar y aprender a usar la libreria mlr3, nos gustó muchisimo lo potente y sencilla que es. Asi que la aplicamos al proyecto una vez estudiada.

```
# Imporamos la libreria base.
library(mlr3)
```

Para definir la tarea del clsificador, lo primero que haremos es convertir la variable objetivo de tipo numeric a factor. Ya que asi lo requiere la instancia de task.

```
datos$GRUPO <- ifelse(datos$RETIRO==1, 'RETIRADO', 'ACTIVO')
```

Validemos que en efecto sea de tipo factor.

```
datos$GRUPO <- as.factor(datos$GRUPO)
class(datos$GRUPO)
```

```
## [1] "factor"
```

```
# Convertir las columnas de datos categoricos como factores.
datos$ID_GENERO <- as.factor(datos$ID_GENERO)
datos$ID_ESTADO_CIVIL <- as.factor(datos$ID_ESTADO_CIVIL)
datos$TIPO_NOMINA <- as.factor(datos$TIPO_NOMINA)
datos$TIPO_CARGO <- as.factor(datos$TIPO_CARGO)
datos$TIPO_AREA <- as.factor(datos$TIPO_AREA)
datos$SENAL_SINDICALIZADO <- as.factor(datos$SENAL_SINDICALIZADO)
datos$CLASE_EMPLEADO <- as.factor(datos$CLASE_EMPLEADO)
datos$SUBSIDIO_TRANSPORTE <- as.factor(datos$SUBSIDIO_TRANSPORTE)
datos$CAUSA.NOMBRE <- as.factor(datos$CAUSA.NOMBRE)
datos$TIPO.DE.RETIRO <- as.factor(datos$TIPO.DE.RETIRO)

datos$FECHA  <- as.factor(datos$FECHA)
datos$FECHA.FIN.CONTRATO <- as.factor(datos$FECHA.FIN.CONTRATO)
datos$FECHA_INGRESO <- as.factor(datos$FECHA_INGRESO)
datos$FECHA_INICIO_CONTRATO <- as.factor(datos$FECHA_INICIO_CONTRATO)
datos$FECHA_NACIMIENTO <- as.factor(datos$FECHA_NACIMIENTO)
```

```
# Convertir las columnnas con datos numericos a tipos numericos.
datos$COMPENSACION_VARIABLE <- as.numeric(datos$COMPENSACION_VARIABLE)
```

```
## Warning: NAs introduced by coercion
```

```
datos$PORCENTAJE_RIESGO  <- as.numeric(datos$PORCENTAJE_RIESGO)
```

```
## Warning: NAs introduced by coercion
```

```
datos$TOTAL_SALARIO <- as.numeric(datos$TOTAL_SALARIO)
```

```
## Warning: NAs introduced by coercion
```

Elección de variables

Justificación desde el punto de vista de la elección de las variables predictoras. De forma manual creemos que las variables mas importantes seran: * edad * causa de retiro * fecha retiro * tipo de cargo * salario Ya que pueden influir en la desición de que una persona se vaya o no. Y las primeras variables porque estan directamente relacionadas con la variable objetivo. Es decir, si tiene causa de retiro es porque se retiró, de ahi viene la etiqueta. Sin embargo, usaremos la potencia de la libreria MLR3 para que indicarle al objeto learner y el task que hagan el feature selection automaticamente.

Etapa 4: Modelado, pruebas, evaluación y optimizacion

```
# Definimos una tarea para clasificar, configuramos los datos, la variable objetivo y la clase positiva
task_clasificar <- TaskClassif$new(id = "datos", backend = datos, target = "GRUPO", positive = 'RETIRAD
task_clasificar
```

```
## <TaskClassif:datos> (1579 x 32)
## * Target: GRUPO
## * Properties: twoclass
## * Features (31):
##   - fct (15): CAUSA.NOMBRE, CLASE_EMPLEADO, FECHA, FECHA.FIN.CONTRATO,
##     FECHA_INGRESO, FECHA_INICIO_CONTRATO, FECHA_NACIMIENTO,
##     ID_ESTADO_CIVIL, ID_GENERO, SENAL_SINDICALIZADO,
##     SUBSIDIO_TRANSPORTE, TIPO.DE.RETIRO, TIPO_AREA, TIPO_CARGO,
##     TIPO_NOMINA
##   - int (13): CATEGORIA_CARGO, CEDULA_ANONIMIZADA, CLASE_NOMINA,
##     CODIGO_EMPRESA, CODIGO_LOCALIDAD, EDAD, ID_CARGO, ID_CENTRO_COSTO,
##     NIVEL_CARGO, PORCENT_COMP_VARIABLE, RETIRO, SALARIO,
##     UBICACION_ANONIMIZADA
##   - dbl (3): COMPENSACION_VARIABLE, PORCENTAJE_RIESGO, TOTAL_SALARIO
```

Al imprimir el objeto `task_clasificar` podemos observar que nuestros target es la variable `GrUPO`, apreciamos que nuestro modelo posee 2 clases para clasificar: `RETIRADO` o `ACTIVO`; Finalmente, observamos que tenemos 31 features.

```
# Definimos un objeto learner y lo inicializamos con el metodo de aprendizaje de clasificacion
learner_clasificar <- lrn("classif.rpart", cp = .01)
learner_clasificar
```

```
## <LearnerClassifRpart:classif.rpart>
## * Model: -
## * Parameters: xval=0, cp=0.01
## * Packages: rpart
## * Predict Type: response
## * Feature types: logical, integer, numeric, factor, ordered
## * Properties: importance, missings, multiclass, selected_features,
##   twoclass, weights
```

```
# Veamos las variables mas importantes para el modelo
library("mlr3filters")
```

```r
filter = flt("importance", learner = learner_clasificar)
filter$calculate(task_clasificar)
head(as.data.table(filter), 3)
```

```
##           feature    score
## 1:         RETIRO 572.9297
## 2:   CAUSA.NOMBRE 572.9297
## 3: TIPO.DE.RETIRO 534.8360
```

```r
# Division de datos para entrenamiento y pruebas desde el objeto (task_clasificar)
train_set <- sample(task_clasificar$nrow, 0.8 * task_clasificar$nrow)
test_set <- setdiff(seq_len(task_clasificar$nrow), train_set)
```

```r
# Entrenamos el modelo con los datos de (train_set)
learner_clasificar$train(task_clasificar, row_ids = train_set)
```

```r
# Realizamos la prediccion con el objeto learner_clasificar con los datos de (test_set)
prediction <- learner_clasificar$predict(task_clasificar, row_ids = test_set)
```

```r
# Veamos los resultados del modelo y como le fué con la prediccion en el set de pruebas.
print(prediction)
```

```
## <PredictionClassif> for 316 observations:
##     row_id    truth response
##          1   ACTIVO   ACTIVO
##         11   ACTIVO   ACTIVO
##         19   ACTIVO   ACTIVO
## ---
##       1561 RETIRADO RETIRADO
##       1576 RETIRADO RETIRADO
##       1578 RETIRADO RETIRADO
```

```r
head(as.data.table(prediction))
```

```
##    row_id  truth response
## 1:      1 ACTIVO   ACTIVO
## 2:     11 ACTIVO   ACTIVO
## 3:     19 ACTIVO   ACTIVO
## 4:     22 ACTIVO   ACTIVO
## 5:     25 ACTIVO   ACTIVO
## 6:     28 ACTIVO   ACTIVO
```

```r
# Veamos la matriz de confusion
prediction$confusion
```

```
##           truth
## response   RETIRADO ACTIVO
##   RETIRADO       75      0
##   ACTIVO          0    241
```

```r
# Veamos algunos datos de la predicción
head(prediction$data)
```

```
## $row_ids
##   [1]    1   11   19   22   25   28   32   34   35   37   45   48   55   60   65
##  [16]   66   74   77   81   82  101  109  111  123  124  131  133  134  138  148
##  [31]  150  155  156  161  163  166  171  177  182  184  194  197  208  209  218
##  [46]  221  229  230  234  240  241  247  248  264  265  275  278  281  282  283
```

```
##  [61]  294  306  309  316  323  333  335  340  345  346  348  351  352  355  356
##  [76]  358  368  370  375  389  398  400  407  409  415  418  420  430  431  434
##  [91]  442  443  450  453  454  458  468  474  480  487  489  490  492  493  506
## [106]  508  511  513  515  516  521  533  535  538  539  540  549  565  566  568
## [121]  569  581  589  592  594  604  616  623  626  630  632  633  640  644  654
## [136]  656  665  669  670  672  675  683  688  689  702  704  712  715  717  722
## [151]  728  729  730  753  754  760  764  767  779  787  788  789  804  813  817
## [166]  821  823  837  840  843  845  854  860  864  866  867  871  878  880  884
## [181]  889  892  893  894  897  899  909  913  914  916  927  932  938  941  945
## [196]  950  952  958  961  974  975  978  979  980  982  988 1008 1023 1028 1046
## [211] 1052 1054 1055 1056 1061 1063 1065 1072 1087 1092 1094 1095 1099 1113 1126
## [226] 1128 1129 1131 1143 1148 1152 1161 1166 1167 1168 1174 1177 1188 1197 1199
## [241] 1203 1210 1213 1215 1218 1219 1220 1224 1235 1246 1247 1248 1253 1261 1265
## [256] 1266 1267 1270 1272 1273 1276 1285 1286 1293 1296 1300 1307 1308 1312 1315
## [271] 1319 1335 1337 1339 1340 1341 1348 1357 1360 1376 1379 1382 1384 1393 1401
## [286] 1402 1407 1414 1422 1426 1427 1445 1447 1448 1449 1455 1468 1471 1493 1505
## [301] 1508 1515 1516 1521 1530 1537 1544 1547 1548 1551 1552 1555 1560 1561 1576
## [316] 1578
##
## $truth
##   [1] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##   [9] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [17] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [25] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [33] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [41] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [49] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [57] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [65] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [73] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [81] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [89] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [97] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [105] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [113] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [121] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [129] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [137] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [145] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [153] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [161] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [169] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [177] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [185] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [193] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [201] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [209] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [217] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [225] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [233] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [241] ACTIVO   RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [249] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [257] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [265] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
```

```
## [273] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [281] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [289] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [297] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [305] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [313] RETIRADO RETIRADO RETIRADO RETIRADO
## Levels: RETIRADO ACTIVO
##
## $response
##   [1] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##   [9] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [17] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [25] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [33] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [41] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [49] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [57] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [65] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [73] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [81] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [89] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
##  [97] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [105] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [113] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [121] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [129] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [137] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [145] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [153] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [161] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [169] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [177] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [185] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [193] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [201] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [209] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [217] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [225] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [233] ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO   ACTIVO
## [241] ACTIVO   RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [249] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [257] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [265] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [273] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [281] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [289] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [297] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [305] RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO RETIRADO
## [313] RETIRADO RETIRADO RETIRADO RETIRADO
## Levels: RETIRADO ACTIVO
```

```r
# Veamos la metrica de exactitud de la prediccion
measure <- msr("classif.acc")
prediction$score(measure)
```

```
## classif.acc
##           1
```

Re-construyamos la instancia del modelo usando Validacion Cruzada

```
resampling_cross_validation <- rsmp("cv", folds = 3L)
resampling <- resample(task_clasificar, learner_clasificar, resampling_cross_validation)
```

```
## INFO  [23:33:18.822] Applying learner 'classif.rpart' on task 'datos' (iter 1/3)
## INFO  [23:33:18.871] Applying learner 'classif.rpart' on task 'datos' (iter 3/3)
## INFO  [23:33:18.897] Applying learner 'classif.rpart' on task 'datos' (iter 2/3)
```

```
resampling$score(measure)
```

```
##                 task task_id                   learner    learner_id
## 1: <TaskClassif[45]>   datos <LearnerClassifRpart[33]> classif.rpart
## 2: <TaskClassif[45]>   datos <LearnerClassifRpart[33]> classif.rpart
## 3: <TaskClassif[45]>   datos <LearnerClassifRpart[33]> classif.rpart
##            resampling resampling_id iteration              prediction
## 1: <ResamplingCV[19]>            cv         1 <PredictionClassif[19]>
## 2: <ResamplingCV[19]>            cv         2 <PredictionClassif[19]>
## 3: <ResamplingCV[19]>            cv         3 <PredictionClassif[19]>
##    classif.acc
## 1:           1
## 2:           1
## 3:           1
```

```
# Veamos la metrica de accuracy
resampling$aggregate(measure)
```

```
## classif.acc
##           1
```

```
# Veamos la matriz de confusion.
resampling$prediction()$confusion
```

```
##           truth
## response   RETIRADO ACTIVO
##   RETIRADO      376      0
##   ACTIVO          0   1203
```

```
# Cambiemos el tipo de prediccion del modelo para ver las probabilidades.
learner_clasificar$predict_type = "prob"
```

```
# re-entrenamos el modelo
learner_clasificar$train(task_clasificar, row_ids = train_set)
```

```
# Veamos que tiene el objeto modelo por dentro
learner_clasificar$model
```

```
## n= 1263
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
```

```
##
## 1) root 1263 301 ACTIVO (0.2383215 0.7616785)
##   2) CAUSA.NOMBRE=DESPIDO CON JUSTA CAUSA,DESPIDO SIN JUSTA CAUSA,MUTUO ACUERDO,RENUNCIA VOLUNTARIA,
##   3) CAUSA.NOMBRE=NO APLICA 962   0 ACTIVO (0.0000000 1.0000000) *
```

```
# Hacemos de nuevo una prediccion con los datos de pruebas.
prediction = learner_clasificar$predict(task_clasificar, row_ids = test_set)
```

```
# Veamos respuestas y probabilidades juntas.
head(as.data.table(prediction))
```

```
##    row_id  truth response prob.RETIRADO prob.ACTIVO
## 1:      1 ACTIVO   ACTIVO             0           1
## 2:     11 ACTIVO   ACTIVO             0           1
## 3:     19 ACTIVO   ACTIVO             0           1
## 4:     22 ACTIVO   ACTIVO             0           1
## 5:     25 ACTIVO   ACTIVO             0           1
## 6:     28 ACTIVO   ACTIVO             0           1
```

```
head(prediction$response)
```

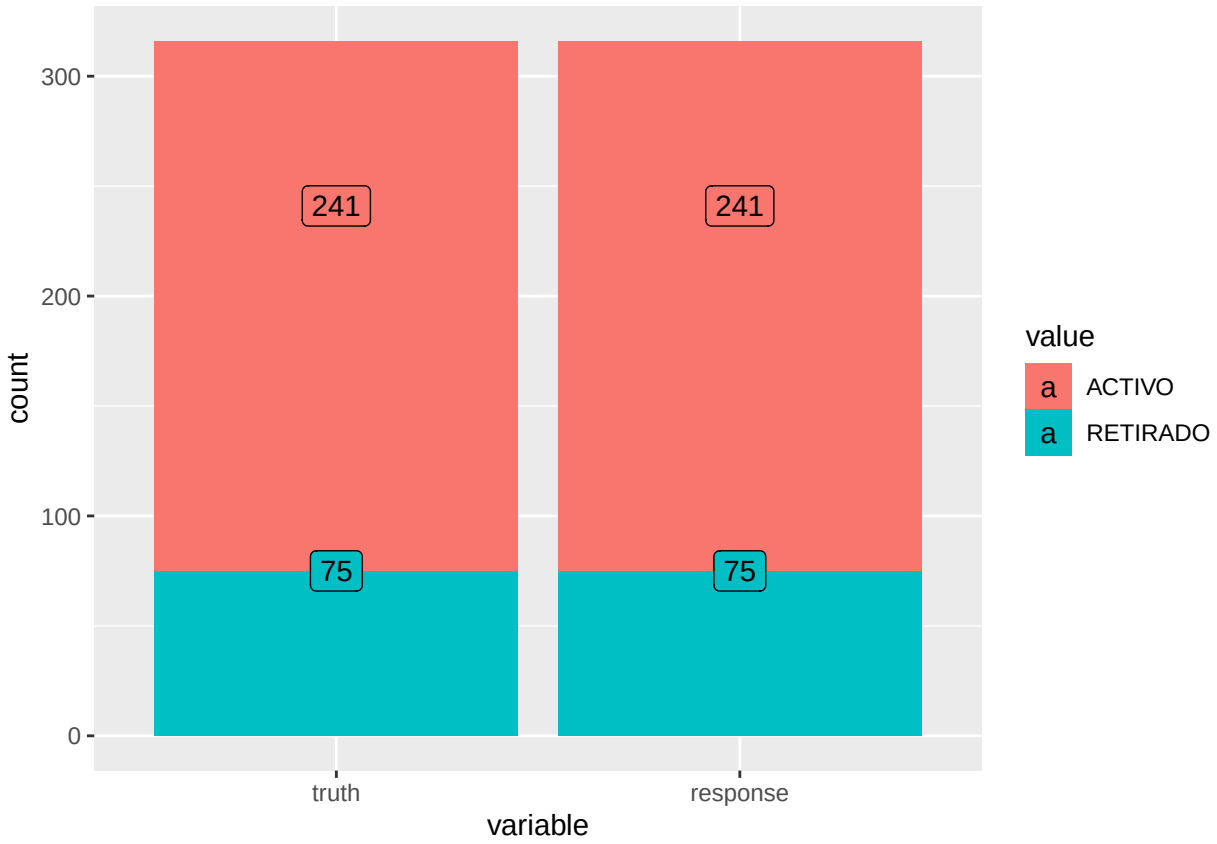
```
## [1] ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO
## Levels: RETIRADO ACTIVO
```

```
head(prediction$prob)
```

```
##      RETIRADO ACTIVO
## [1,]        0      1
## [2,]        0      1
## [3,]        0      1
## [4,]        0      1
## [5,]        0      1
## [6,]        0      1
```

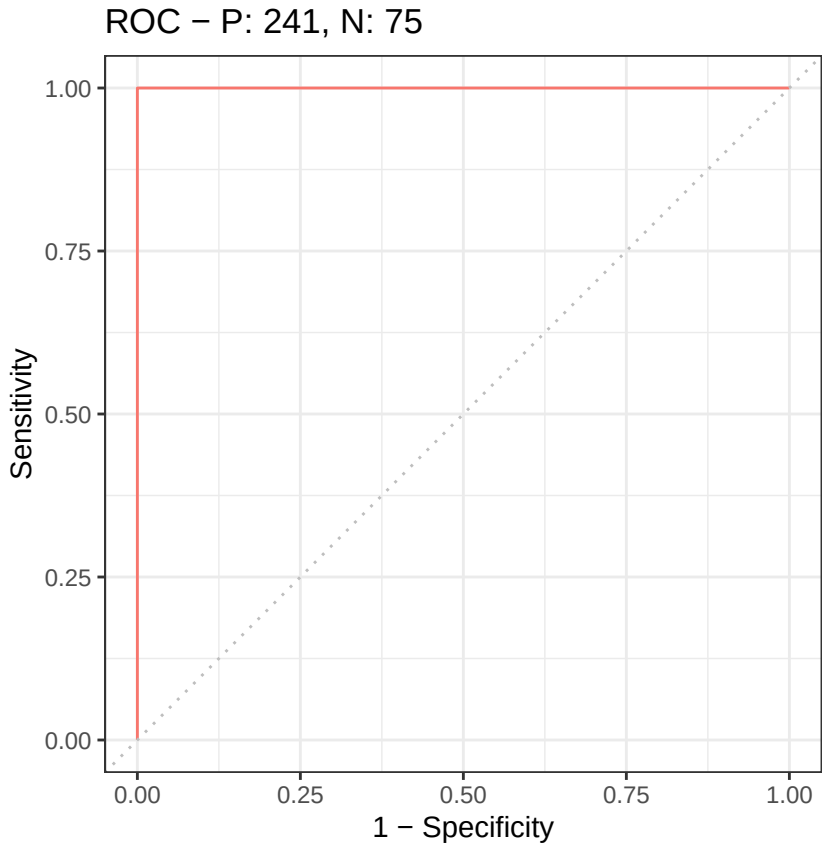
```
library("mlr3viz")

autoplot(prediction)
```



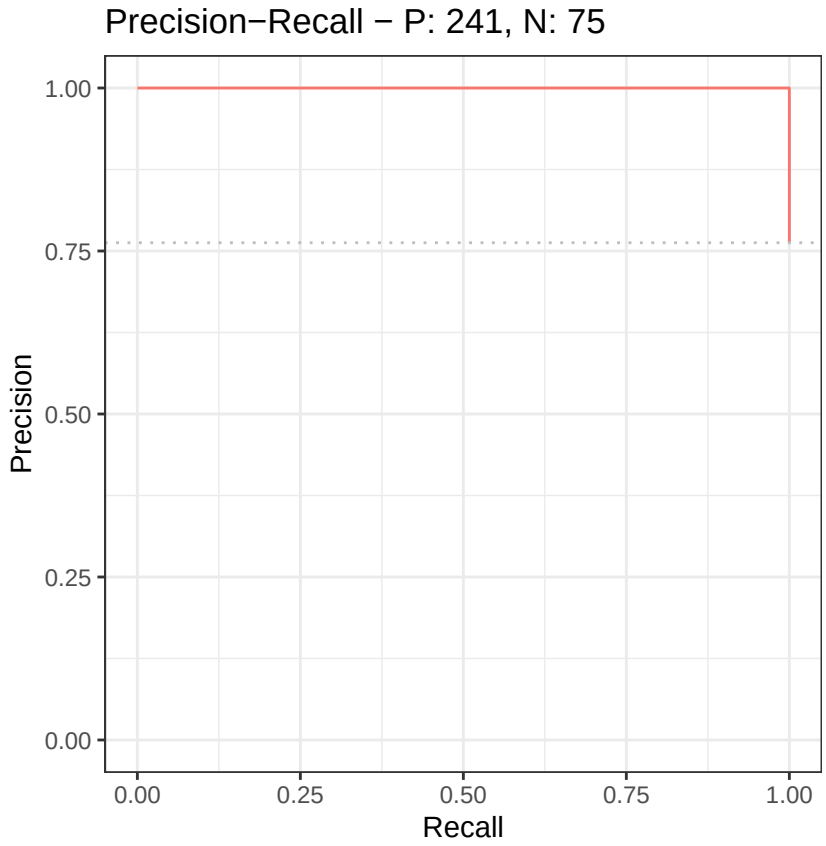
Curva ROC

```
ggplot2::autoplot(prediction, type = "roc")
```



Precision vs Recall

```
ggplot2::autoplot(prediction, type = "prc")
```




```
measure = msr("classif.acc")
prediction$score(measure)
```

```
## classif.acc
##           1
```

```
measure_1 = msr("classif.bacc")
prediction$score(measure_1)
```

```
## classif.bacc
##            1
```

```
measure_2 = msr("classif.ce")
prediction$score(measure_2)
```

```
## classif.ce
##          0
```

Veamos ahora como manejamos el imbalance de clases con los metodos de undersample, oversample y smote

```
# Veamos el tamaño de cada clase
table(task_clasificar$truth())
```

```
##
## RETIRADO   ACTIVO
##      376     1203
```

Se nota que estan bastante desbalanceadas las clases. Vamos a revisarlo.

Vamos a balancear por los tres metodos para ver como nos va.

Vamos a disminuir la clase mayoritaria en 1/3

```
# under = po("classbalancing", id = "undersample", adjust = "major", reference = "major", shuffle = FAL
# table(under$train(list(task_clasificar))$output$truth()) # Comentar para generar PDF
```

Vamos a aumentar la clase minitoria a 3 veces.

```
# over = po("classbalancing", id = "oversample", adjust = "minor", reference = "minor", shuffle = FALSE
# table(over$train(list(task_clasificar))$output$truth()) # Comentar para generar PDF
```

Dado que no todas las variables son numericas, no podremos usar SMOTE por esta ocasion. Aunque podriamos pasar todo a numeros, pero no es un costo que asumiremos esta vez.

Veamos las propiedades del modelo:

```
learner_clasificar$model
```

```
## n= 1263
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 1263 301 ACTIVO (0.2383215 0.7616785)
##   2) CAUSA.NOMBRE=DESPIDO CON JUSTA CAUSA,DESPIDO SIN JUSTA CAUSA,MUTUO ACUERDO,RENUNCIA VOLUNTARIA,
##   3) CAUSA.NOMBRE=NO APLICA 962   0 ACTIVO (0.0000000 1.0000000) *
```

Etapa 5: Extracción de resultados para su uso en producción

```
prediccion_todos_los_datos <- predict(learner_clasificar, datos, predict_type = "<Prediction>")


# Agregamos a los datos los valores de la prediccion
datos$prediccion <- prediccion_todos_los_datos$data$response
```

```
# Seleccionamos solo las columnas de interes
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
resultado_final <- datos %>% select(CEDULA_ANONIMIZADA, GRUPO, prediccion )
head(resultado_final)
```

```
##   CEDULA_ANONIMIZADA  GRUPO prediccion
## 1                 50 ACTIVO     ACTIVO
## 2                158 ACTIVO     ACTIVO
## 3                215 ACTIVO     ACTIVO
## 4                279 ACTIVO     ACTIVO
## 5                319 ACTIVO     ACTIVO
## 6                376 ACTIVO     ACTIVO
```

```
# install.packages("writexl")
library("writexl")
write_xlsx(resultado_final,"resultado_prediccion.xlsx")
```

Etapa 6: Conclusiones y recomendaciones

- Los datos estaban muy sucios. Lo que nos puso un gran reto para entenderlos, limpiarlos, unirlos y prepararlos para empezar a usarlos.
- La cantidad de empleados retirados y activos son bastantes diferentes. Lo que causa que las clases esten desbalamceadas.
- Entender y usar librerias potentes nos permiten ahorrar tiempo y dinero al poder enfocarnos en procesos de limpieza de datos y entendimiento del problema y del negocio.
- Nos faltó comunicación con el cliente (profesor) para resolver dudas y dejar de suponer cosas. Tambien faltó hacer más preguntas.
- Los resultados del modelo parece ser bastante buenos. Creemos que se puede usar en un contexto serio para ponerlo a prueba.
- El modelo terminó usando solo 3 de las 29 variables. Lo que nos hace suponer que si creamos un modelo manualmente entonces, seria mas complejo (mas variables).
- Las personas que se retiran más son las peronas que por lo general gana menos en la compañia. La mayoria de personas que tenian salarios buenos, se encuentra en el el GRUPO de ACTIVO.