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)

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

#skim(datos) # Comentar para generar PDF

#### Veamos algunos ejemplos de los datos

head(datos, 8)

## 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 = 'RETIRADO')  
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")  
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

# 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)

# Entrenamos el modelo con los datos de (train\_set)  
learner\_clasificar$train(task\_clasificar, row\_ids = train\_set)

# Realizamos la prediccion con el objeto learner\_clasificar con los datos de (test\_set)  
prediction <- learner\_clasificar$predict(task\_clasificar, row\_ids = test\_set)

# 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  
## 4 ACTIVO ACTIVO  
## 5 ACTIVO ACTIVO  
## 9 ACTIVO ACTIVO  
## ---   
## 1568 RETIRADO RETIRADO  
## 1570 RETIRADO RETIRADO  
## 1572 RETIRADO RETIRADO

head(as.data.table(prediction))

## row\_id truth response  
## 1: 4 ACTIVO ACTIVO  
## 2: 5 ACTIVO ACTIVO  
## 3: 9 ACTIVO ACTIVO  
## 4: 21 ACTIVO ACTIVO  
## 5: 27 ACTIVO ACTIVO  
## 6: 29 ACTIVO ACTIVO

# Veamos la matriz de confusion  
prediction$confusion

## truth  
## response RETIRADO ACTIVO  
## RETIRADO 65 0  
## ACTIVO 0 251

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

## $row\_ids  
## [1] 4 5 9 21 27 29 45 47 52 53 56 59 68 70 73  
## [16] 85 101 105 106 108 113 114 115 117 123 124 127 128 136 142  
## [31] 157 159 165 166 167 170 175 176 180 182 193 196 202 206 209  
## [46] 211 229 232 238 244 252 261 267 268 287 299 302 305 308 315  
## [61] 317 324 325 337 340 342 355 360 363 367 372 375 381 385 392  
## [76] 393 399 404 405 409 410 417 418 422 423 429 442 454 461 467  
## [91] 484 490 498 502 503 505 513 519 524 526 534 541 542 543 546  
## [106] 552 568 570 571 579 580 590 606 607 610 616 617 624 625 627  
## [121] 630 637 638 641 645 648 649 650 654 656 658 666 671 678 679  
## [136] 680 682 684 689 691 693 695 696 697 705 714 721 723 725 727  
## [151] 736 744 751 752 757 759 763 764 765 769 778 779 781 790 798  
## [166] 800 810 816 818 819 823 827 836 841 847 848 852 853 855 858  
## [181] 866 873 876 882 886 891 894 901 902 903 907 910 911 917 929  
## [196] 933 939 945 947 948 964 968 974 976 977 1002 1004 1006 1007 1009  
## [211] 1010 1024 1025 1027 1033 1036 1059 1063 1065 1066 1074 1078 1080 1087 1088  
## [226] 1089 1105 1107 1108 1109 1113 1116 1124 1128 1130 1131 1134 1139 1146 1154  
## [241] 1157 1158 1165 1167 1170 1175 1179 1181 1187 1189 1203 1206 1209 1210 1220  
## [256] 1222 1229 1232 1239 1247 1251 1256 1267 1273 1281 1282 1299 1303 1311 1315  
## [271] 1318 1321 1328 1331 1334 1343 1352 1356 1359 1380 1381 1383 1384 1391 1394  
## [286] 1396 1401 1403 1405 1409 1415 1417 1420 1433 1435 1437 1440 1461 1467 1468  
## [301] 1469 1483 1488 1489 1490 1503 1508 1509 1546 1555 1556 1557 1558 1568 1570  
## [316] 1572  
##   
## $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 ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO   
## [249] ACTIVO ACTIVO ACTIVO 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 ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO ACTIVO   
## [249] ACTIVO ACTIVO ACTIVO 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

# 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 [23:33:57.255] Applying learner 'classif.rpart' on task 'datos' (iter 1/3)   
## INFO [23:33:57.319] Applying learner 'classif.rpart' on task 'datos' (iter 2/3)   
## INFO [23:33:57.354] Applying learner 'classif.rpart' on task 'datos' (iter 3/3)

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 311 ACTIVO (0.2462391 0.7537609)   
## 2) CAUSA.NOMBRE=DESPIDO CON JUSTA CAUSA,DESPIDO SIN JUSTA CAUSA,MUTUO ACUERDO,PERIODO DE PRUEBA,RENUNCIA VOLUNTARIA,RETIRO VOLUNTARIO,TERMINACION PERIODO DE PRUEBA,TERMINACION UNILATERAL,VENCIMIENTO DE CONTRATO 311 0 RETIRADO (1.0000000 0.0000000) \*  
## 3) CAUSA.NOMBRE=NO APLICA 952 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: 4 ACTIVO ACTIVO 0 1  
## 2: 5 ACTIVO ACTIVO 0 1  
## 3: 9 ACTIVO ACTIVO 0 1  
## 4: 21 ACTIVO ACTIVO 0 1  
## 5: 27 ACTIVO ACTIVO 0 1  
## 6: 29 ACTIVO ACTIVO 0 1

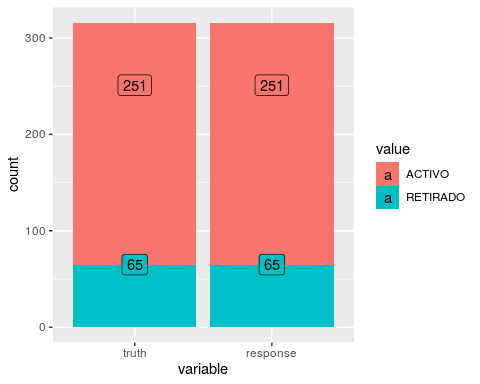
head(prediction$response)

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

head(prediction$prob)

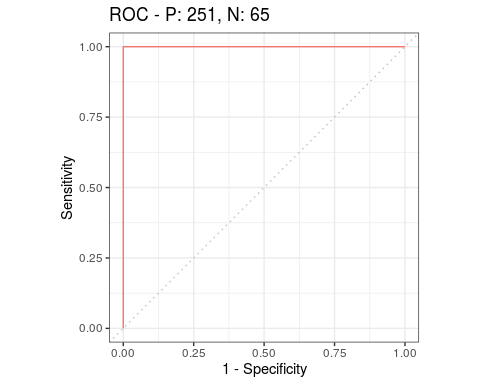
## RETIRADO ACTIVO  
## [1,] 0 1  
## [2,] 0 1  
## [3,] 0 1  
## [4,] 0 1  
## [5,] 0 1  
## [6,] 0 1

library("mlr3viz")  
  
autoplot(prediction)

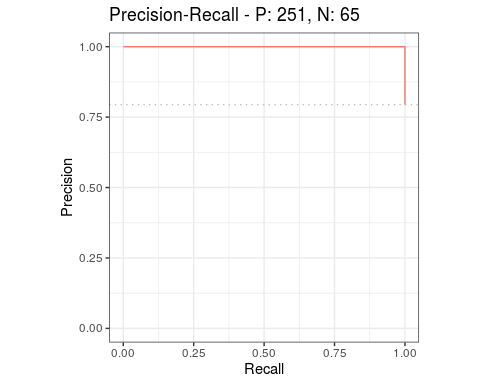


# Curva ROC

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 = FALSE, ratio = 1 / 3) # Comentar para generar PDF  
# 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, ratio = 3) # Comentar para generar PDF  
# 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 311 ACTIVO (0.2462391 0.7537609)   
## 2) CAUSA.NOMBRE=DESPIDO CON JUSTA CAUSA,DESPIDO SIN JUSTA CAUSA,MUTUO ACUERDO,PERIODO DE PRUEBA,RENUNCIA VOLUNTARIA,RETIRO VOLUNTARIO,TERMINACION PERIODO DE PRUEBA,TERMINACION UNILATERAL,VENCIMIENTO DE CONTRATO 311 0 RETIRADO (1.0000000 0.0000000) \*  
## 3) CAUSA.NOMBRE=NO APLICA 952 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.