Memoria Competición Kaggel Preprocesamiento

Francisco Pérez Hernández

17/2/2017

Contents

1 Introducción al problema y a Kaggel

Lo primero que se pretende realizar en este apartado es leer el dataset que nos han dado y realizar una subida a la plataforma Kaggel para obtener una primera puntuación. Mi usuario en Kaggel es "PacoPollos".

1.1 Lectura del dataset accidentes

Vamos a leer tanto los archivos de train como test dados.

```
accidentes.train.original <- read.csv("accidentes-kaggle.csv")
accidentes.test.original <- read.csv("accidentes-kaggle-test.csv")
```

Una vez leídos vamos a realizar un summary para ver como están compuestos los datos.

```
summary(accidentes.train.original)
```

```
##       ANIO              MES             HORA          DIASEMANA
##  Min.   :2008   Julio     : 2757   14     : 1965   DOMINGO  :3597
##  1st Qu.:2009   Junio     : 2649   19     : 1847   JUEVES   :4351
##  Median :2010   Mayo      : 2605   13     : 1823   LUNES    :4349
##  Mean   :2010   Octubre   : 2600   17     : 1749   MARTES   :4343
##  3rd Qu.:2012   Septiembre: 2491   18     : 1726   MIERCOLES:4394
##  Max.   :2013   Diciembre : 2448   12     : 1713   SABADO   :4000
##                 (Other)   :14452   (Other):19179   VIERNES  :4968
##      PROVINCIA                COMUNIDAD_AUTONOMA          ISLA
##  Barcelona: 6238   Cataluna            :8208     NO_ES_ISLA  :28476
##  Madrid   : 4735   Madrid, Comunidad de:4735     MALLORCA    :  608
##  Valencia : 1658   Andalucia           :4412     TENERIFE    :  436
##  Sevilla  :  977   Comunitat Valenciana:2653     GRAN CANARIA:  199
```

```
##  Cadiz    :  887   Pais Vasco          :1594     IBIZA       :  117
##  Girona   :  814   Castilla y Leon     :1505     LANZAROTE   :   53
##  (Other)  :14693   (Other)             :6895     (Other)     :  113
##   TOT_VICTIMAS     TOT_MUERTOS      TOT_HERIDOS_GRAVES TOT_HERIDOS_LEVES
##  Min.   : 1.000   Min.   :0.00000   Min.   :0.0000     Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.:0.00000   1st Qu.:0.0000     1st Qu.: 1.00
##  Median : 1.000   Median :0.00000   Median :0.0000     Median : 1.00
##  Mean   : 1.429   Mean   :0.02447   Mean   :0.1453     Mean   : 1.26
##  3rd Qu.: 2.000   3rd Qu.:0.00000   3rd Qu.:0.0000     3rd Qu.: 1.00
##  Max.   :19.000   Max.   :7.00000   Max.   :9.0000     Max.   :18.00
##
##  TOT_VEHICULOS_IMPLICADOS          ZONA                 ZONA_AGRUPADA
##  Min.   : 1.000           CARRETERA  :13278   VIAS INTERURBANAS:13335
##  1st Qu.: 1.000           TRAVESIA   :  241   VIAS URBANAS     :16667
##  Median : 2.000           VARIANTE   :   57
##  Mean   : 1.738           ZONA URBANA:16426
##  3rd Qu.: 2.000
##  Max.   :21.000
##
##    CARRETERA
##  A-7    :  294
##  A-2    :  278
##  AP-7   :  229
##  N-340  :  229
##  A-4    :  184
##  (Other):12098
##  NA's   :16690
##                                                 RED_CARRETERA
##  OTRAS TITULARIDADES                                   :  318
##  TITULARIDAD AUTONOMICA                                : 3890
##  TITULARIDAD ESTATAL                                   : 4021
##  TITULARIDAD MUNICIPAL                                 :19077
##  TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
##
##              TIPO_VIA
##  OTRO TIPO       :15527
##  VIA CONVENCIONAL:10044
##  AUTOVIA         : 2941
##  AUTOPISTA       :  723
##  CAMINO VECINAL  :  519
##  RAMAL DE ENLACE :  101
##  (Other)         :  147
##                                        TRAZADO_NO_INTERSEC
##  CURVA FUERTE CON MARCA Y SIN VELOCIDAD MARCADA:  559
##  CURVA FUERTE CON MARCA Y VELOCIDAD MARCADA    :  872
##  CURVA FUERTE SIN MARCAR                       :  481
##  CURVA SUAVE                                   : 2875
##  ES_INTERSECCION                               :11038
##  RECTA                                         :14177
##
##             TIPO_INTERSEC
##  EN T O Y          : 3350
##  EN X O +          : 4714
```

```
##  ENLACE DE ENTRADA :  421
##  ENLACE DE SALIDA  :  223
##  GIRATORIA         : 2006
##  NO_ES_INTERSECCION:18983
##  OTROS             :  305
##                                                  ACOND_CALZADA
##  CARRIL CENTRAL DE ESPERA                               :  193
##  NADA ESPECIAL                                          : 4645
##  OTRO TIPO                                              :  791
##  PASO PARA PEATONES O ISLETAS EN CENTRO DE VIA PRINCIPAL:  397
##  RAQUETA DE GIRO IZQUIERDA                              :  109
##  SOLO ISLETAS O PASO PARA PEATONES                      :  168
##  NA's                                                   :23699
##                  PRIORIDAD           SUPERFICIE_CALZADA
##  NINGUNA (SOLO NORMA) :13495   SECA Y LIMPIA  :25236
##  SEMAFORO             : 1778   MOJADA         : 3895
##  SEÆAL DE STOP        : 1750   OTRO TIPO      :  327
##  SOLO MARCAS VIALES   : 1659   UMBRIA         :  165
##  SEÆAL DE CEDA EL PASO: 1629   GRAVILLA SUELTA:  150
##  (Other)              : 1569   ACEITE         :   83
##  NA's                 : 8122   (Other)        :  146
##                          LUMINOSIDAD      FACTORES_ATMOSFERICOS
##  CREPUSCULO                     : 1330   BUEN TIEMPO  :25852
##  NOCHE: ILUMINACION INSUFICIENTE: 1067   LLOVIZNANDO  : 2524
##  NOCHE: ILUMINACION SUFICIENTE  : 4793   OTRO         :  715
##  NOCHE: SIN ILUMINACION         : 1815   LLUVIA FUERTE:  499
##  PLENO DIA                      :20997   VIENTO FUERTE:  156
##                                          NIEBLA LIGERA:   83
##                                          (Other)      :  173
##               VISIBILIDAD_RESTRINGIDA         OTRA_CIRCUNSTANCIA
##  SIN RESTRICCION          :16982      NINGUNA          :24967
##  CONFIGURACION DEL TERRENO:  989      OTRA             :  942
##  OTRA_CAUSA               :  491      OBRAS            :  263
##  FACTORES ATMOSFERICOS    :  374      FUERTE DESCENSO  :  227
##  EDIFICIOS                :  229      CAMBIO DE RASANTE:  100
##  (Other)                  :  252      (Other)          :  264
##  NA's                     :10685      NA's             : 3239
##           ACERAS         DENSIDAD_CIRCULACION          MEDIDAS_ESPECIALES
##  NO HAY ACERA:21416   CONGESTIONADA:  308     CARRIL REVERSIBLE :   17
##  SI HAY ACERA: 5437   DENSA        : 1479     HABILITACION ARCEN:    8
##  NA's        : 3149   FLUIDA       :17505     NINGUNA MEDIDA    :21024
##                       NA's         :10710     OTRA MEDIDA       :  278
##                                               NA's              : 8675
##
##
##             TIPO_ACCIDENTE
##  Atropello         : 3642
##  Colision_Obstaculo:  952
##  Colision_Vehiculos:16520
##  Otro              : 1807
##  Salida_Via        : 6013
##  Vuelco            : 1068
##
```

Vemos como las variables TTO_VICTIMAS, TOT_MUERTOS, TOT_HERIDOS_GRAVES, TOT_HERIDOS_LEVES y TOT_VEHICULOS_IMPLICADOS son las únicas variables numéricas, por lo que nos quedaremos con ellas para la primera prueba, junto con la variable clasificadora TIPO_ACCIDENTE.

```
accidentes.train.solo.numericos <- accidentes.train.original[,c(8,9,10,11,12,30)]
accidentes.test.solo.numericos <- accidentes.test.original[,c(8,9,10,11,12)]
```

1.2 Primera prueba con un modelo

Lo primero es, con las variables numéricas únicamente, voy a realizar un primer modelo, que será un árbol, para predecir la clase del conjunto de test y comprobar el funcionamiento de Kaggel al no tener experiencia anterior.

```
set.seed(1234)
ct1 <- ctree(TIPO_ACCIDENTE ~., accidentes.train.solo.numericos)
testPred1 <- predict(ct1, newdata = accidentes.test.solo.numericos)
```

Por lo que ya tenemos el conjunto de test predecido. Además el árbol creado tendría la siguiente estructura:

```
ct1
```

```
##
##   Conditional inference tree with 14 terminal nodes
##
## Response:  TIPO_ACCIDENTE
## Inputs:  TOT_VICTIMAS, TOT_MUERTOS, TOT_HERIDOS_GRAVES, TOT_HERIDOS_LEVES, TOT_VEHICULOS_IMPLICADOS
## Number of observations:  30002
##
## 1) TOT_VEHICULOS_IMPLICADOS <= 1; criterion = 1, statistic = 14488.658
##   2) TOT_VICTIMAS <= 1; criterion = 1, statistic = 329.362
##     3) TOT_HERIDOS_GRAVES <= 0; criterion = 1, statistic = 38.228
##       4) TOT_HERIDOS_LEVES <= 0; criterion = 0.996, statistic = 21.181
##         5)*  weights = 256
##       4) TOT_HERIDOS_LEVES > 0
##         6)*  weights = 7696
##     3) TOT_HERIDOS_GRAVES > 0
##       7)*  weights = 1476
##   2) TOT_VICTIMAS > 1
##     8) TOT_VICTIMAS <= 2; criterion = 1, statistic = 47.735
##       9)*  weights = 1605
##     8) TOT_VICTIMAS > 2
##       10)*  weights = 550
## 1) TOT_VEHICULOS_IMPLICADOS > 1
##   11) TOT_HERIDOS_LEVES <= 1; criterion = 1, statistic = 99.886
##     12) TOT_HERIDOS_LEVES <= 0; criterion = 1, statistic = 49.242
##       13)*  weights = 1276
##     12) TOT_HERIDOS_LEVES > 0
##       14) TOT_VICTIMAS <= 1; criterion = 1, statistic = 34.382
##         15) TOT_VEHICULOS_IMPLICADOS <= 3; criterion = 1, statistic = 28.319
##           16) TOT_VEHICULOS_IMPLICADOS <= 2; criterion = 0.999, statistic = 24.207
##             17)*  weights = 10133
```

```
##           16) TOT_VEHICULOS_IMPLICADOS > 2
##             18)*  weights = 924
##         15) TOT_VEHICULOS_IMPLICADOS > 3
##           19)*  weights = 254
##       14) TOT_VICTIMAS > 1
##         20) TOT_VEHICULOS_IMPLICADOS <= 3; criterion = 0.965, statistic = 15.891
##           21)*  weights = 370
##         20) TOT_VEHICULOS_IMPLICADOS > 3
##           22)*  weights = 21
##   11) TOT_HERIDOS_LEVES > 1
##     23) TOT_VEHICULOS_IMPLICADOS <= 4; criterion = 0.994, statistic = 20.095
##       24) TOT_VEHICULOS_IMPLICADOS <= 2; criterion = 0.998, statistic = 22.592
##         25)*  weights = 4183
##       24) TOT_VEHICULOS_IMPLICADOS > 2
##         26)*  weights = 1124
##     23) TOT_VEHICULOS_IMPLICADOS > 4
##       27)*  weights = 134
```

1.3 Creación del archivo de salida y subida a kaggel

Vamos a escribir la salida del primer modelo para ver su puntuación en Kaggel.

```
salida.primer.modelo <- as.matrix(testPred1)
write.table(salida.primer.modelo,file="predicciones/PrimeraPrediccion.txt",sep=",",quote = F)
```

Por lo que ya tenemos un fichero con la salida del conjunto de test. Lo único que tendremos que modificar es la primera linea del archivo para añadir "Id, Prediction". El resultado de este primer modelo para la competición de Kaggel, subido el 11/02/2017 a las 19:54, con un total de 5 personas entregadas, se ha quedado en la posición 3 con una puntuación del 0.73246.

#	Δ3d	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Luis Suárez	0.82948	2	Fri, 10 Feb 2017 19:54:58
2	↓1	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
3	new	**PacoPollos**	**0.73246**	**1**	**Sat, 11 Feb 2017 18:51:32**
4	↓1	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
5	↓1	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10

Figure 1: Primera puntuación obtenida en Kaggel

2 Análisis del dataset

Una vez realizada la primera prueba en Kaggel, vamos a analizar con detalle el dataset que nos han dado.

2.1 Eliminación de valores perdidos

Anteriormente en el summary, hemos visto que hay variables con valores perdidios, ya que por ejemplo, en la variable CARRETERA uno de los valores que más se repite es NA's. Por lo tanto, vamos a analizar que

variables contienen valores perdidos.

```
porcentaje.de.valores.perdidos.por.columna.train <- apply(accidentes.train.original,2,function(x) sum(is
columnas.train.con.valores.perdidios <- (porcentaje.de.valores.perdidos.por.columna.train > 0)
columnas.train.con.valores.perdidios
```

```
##                      ANIO                       MES                      HORA
##                     FALSE                     FALSE                     FALSE
##                 DIASEMANA                 PROVINCIA       COMUNIDAD_AUTONOMA
##                     FALSE                     FALSE                     FALSE
##                      ISLA             TOT_VICTIMAS              TOT_MUERTOS
##                     FALSE                     FALSE                     FALSE
##       TOT_HERIDOS_GRAVES        TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##                     FALSE                     FALSE                     FALSE
##                      ZONA             ZONA_AGRUPADA                 CARRETERA
##                     FALSE                     FALSE                      TRUE
##            RED_CARRETERA                  TIPO_VIA      TRAZADO_NO_INTERSEC
##                     FALSE                     FALSE                     FALSE
##            TIPO_INTERSEC             ACOND_CALZADA                 PRIORIDAD
##                     FALSE                      TRUE                      TRUE
##       SUPERFICIE_CALZADA               LUMINOSIDAD    FACTORES_ATMOSFERICOS
##                     FALSE                     FALSE                     FALSE
##  VISIBILIDAD_RESTRINGIDA       OTRA_CIRCUNSTANCIA                    ACERAS
##                      TRUE                      TRUE                      TRUE
##     DENSIDAD_CIRCULACION       MEDIDAS_ESPECIALES           TIPO_ACCIDENTE
##                      TRUE                      TRUE                     FALSE
```

Por lo que tenemos que las variables con valores perdidos son: CARRETERA, ACOND_CALZADA, PRIORIDAD, VISIBILIDAD_RESTRINGIDA, OTRA_CIRCUNSTANCIA, ACERAS, DENSIDAD_CIRCULACION y MEDIDAS_ESPECIALES. Veamos el resumen para esas variables.

```
summary(accidentes.train.original[c("CARRETERA","ACOND_CALZADA","PRIORIDAD", "VISIBILIDAD_RESTRINGIDA",
```

```
##    CARRETERA
##  A-7    :  294
##  A-2    :  278
##  AP-7   :  229
##  N-340  :  229
##  A-4    :  184
##  (Other):12098
##  NA's   :16690
##                                                ACOND_CALZADA
##  CARRIL CENTRAL DE ESPERA                                :  193
##  NADA ESPECIAL                                           : 4645
##  OTRO TIPO                                               :  791
##  PASO PARA PEATONES O ISLETAS EN CENTRO DE VIA PRINCIPAL:  397
##  RAQUETA DE GIRO IZQUIERDA                               :  109
##  SOLO ISLETAS O PASO PARA PEATONES                       :  168
##  NA's                                                    :23699
##                  PRIORIDAD                  VISIBILIDAD_RESTRINGIDA
##  NINGUNA (SOLO NORMA) :13495   SIN RESTRICCION          :16982
##  SEMAFORO             : 1778   CONFIGURACION DEL TERRENO:  989
##  SEÆAL DE STOP        : 1750   OTRA_CAUSA               :  491
```

```
##  SOLO MARCAS VIALES   : 1659   FACTORES ATMOSFERICOS    :  374
##  SEÆAL DE CEDA EL PASO: 1629   EDIFICIOS               :  229
##  (Other)              : 1569   (Other)                  :  252
##  NA's                 : 8122   NA's                     :10685
##          OTRA_CIRCUNSTANCIA          ACERAS         DENSIDAD_CIRCULACION
##  NINGUNA          :24967    NO HAY ACERA:21416   CONGESTIONADA:  308
##  OTRA             :  942    SI HAY ACERA: 5437   DENSA        : 1479
##  OBRAS            :  263    NA's        : 3149   FLUIDA       :17505
##  FUERTE DESCENSO  :  227                         NA's         :10710
##  CAMBIO DE RASANTE:  100
##  (Other)          :  264
##  NA's             : 3239
##           MEDIDAS_ESPECIALES
##  CARRIL REVERSIBLE :   17
##  HABILITACION ARCEN:    8
##  NINGUNA MEDIDA    :21024
##  OTRA MEDIDA       :  278
##  NA's              : 8675
##
##
```

Donde podemos ver que el valor más pequeño de NA's es para la variable ACERAS con 3149 instancias con valores perdidos, lo que sería un 10,49% de los datos. Un 25% de los datos de este train serían unas 7500 instancias, por lo que las variables que tienen más del 25% de valores perdidios son: CARRETERA, ACOND_CALZADA, PRIORIDAD, VISIBILIDAD_RESTRINGIDA, DENSIDAD_CIRCULACION y MEDIDAS_ESPECIALES. O lo que es lo mismo, me quedo con las variables OTRA_CIRCUNSTANCIA y ACERAS, del anterior grupo. Pero además voy a comenzar eliminando esas variables ya que a mi juicio pueden no tener demasiada importancia.

```
primeras.variables.eliminadas <- c("CARRETERA","ACOND_CALZADA","PRIORIDAD", "VISIBILIDAD_RESTRINGIDA",
accidentes.train.sin.variables.1 <- accidentes.train.original[,-c(15,20,21,25,26,27,28,29)]
accidentes.train.variables.eliminadas <- accidentes.train.original[,c(15,20,21,25,26,27,28,29)]
```

Por lo que guardo en una variable las variables que he eliminado, y creo mi dataset sin variables con valores NA. Hago lo mismo para el test:

```
accidentes.test.sin.variables.1 <- accidentes.test.original[,-c(15,20,21,25,26,27,28,29)]
accidentes.test.variables.eliminadas <- accidentes.test.original[,c(15,20,21,25,26,27,28,29)]
```

Pensemos ahora que variables restantes pueden ser no interesantes.

```
summary(accidentes.train.sin.variables.1)
```

```
##       ANIO              MES             HORA           DIASEMANA
##  Min.   :2008   Julio     : 2757   14     : 1965   DOMINGO  :3597
##  1st Qu.:2009   Junio     : 2649   19     : 1847   JUEVES   :4351
##  Median :2010   Mayo      : 2605   13     : 1823   LUNES    :4349
##  Mean   :2010   Octubre   : 2600   17     : 1749   MARTES   :4343
##  3rd Qu.:2012   Septiembre: 2491   18     : 1726   MIERCOLES:4394
##  Max.   :2013   Diciembre : 2448   12     : 1713   SABADO   :4000
##                 (Other)   :14452   (Other):19179   VIERNES  :4968
##      PROVINCIA               COMUNIDAD_AUTONOMA          ISLA
```

```
##  Barcelona: 6238   Cataluna            :8208     NO_ES_ISLA  :28476
##  Madrid   : 4735   Madrid, Comunidad de:4735     MALLORCA    :  608
##  Valencia : 1658   Andalucia           :4412     TENERIFE    :  436
##  Sevilla  :  977   Comunitat Valenciana:2653     GRAN CANARIA:  199
##  Cadiz    :  887   Pais Vasco          :1594     IBIZA       :  117
##  Girona   :  814   Castilla y Leon     :1505     LANZAROTE   :   53
##  (Other)  :14693   (Other)             :6895     (Other)     :  113
##   TOT_VICTIMAS     TOT_MUERTOS      TOT_HERIDOS_GRAVES TOT_HERIDOS_LEVES
##  Min.   : 1.000   Min.   :0.00000   Min.   :0.0000     Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.:0.00000   1st Qu.:0.0000     1st Qu.: 1.00
##  Median : 1.000   Median :0.00000   Median :0.0000     Median : 1.00
##  Mean   : 1.429   Mean   :0.02447   Mean   :0.1453     Mean   : 1.26
##  3rd Qu.: 2.000   3rd Qu.:0.00000   3rd Qu.:0.0000     3rd Qu.: 1.00
##  Max.   :19.000   Max.   :7.00000   Max.   :9.0000     Max.   :18.00
##
##  TOT_VEHICULOS_IMPLICADOS          ZONA                 ZONA_AGRUPADA
##  Min.   : 1.000           CARRETERA  :13278   VIAS INTERURBANAS:13335
##  1st Qu.: 1.000           TRAVESIA   :  241   VIAS URBANAS     :16667
##  Median : 2.000           VARIANTE   :   57
##  Mean   : 1.738           ZONA URBANA:16426
##  3rd Qu.: 2.000
##  Max.   :21.000
##
##                                                RED_CARRETERA
##  OTRAS TITULARIDADES                                  :  318
##  TITULARIDAD AUTONOMICA                               : 3890
##  TITULARIDAD ESTATAL                                  : 4021
##  TITULARIDAD MUNICIPAL                                :19077
##  TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
##
##              TIPO_VIA
##  OTRO TIPO       :15527
##  VIA CONVENCIONAL:10044
##  AUTOVIA         : 2941
##  AUTOPISTA       :  723
##  CAMINO VECINAL  :  519
##  RAMAL DE ENLACE :  101
##  (Other)         :  147
##                                       TRAZADO_NO_INTERSEC
##  CURVA FUERTE CON MARCA Y SIN VELOCIDAD MARCADA:  559
##  CURVA FUERTE CON MARCA Y VELOCIDAD MARCADA    :  872
##  CURVA FUERTE SIN MARCAR                       :  481
##  CURVA SUAVE                                   : 2875
##  ES_INTERSECCION                               :11038
##  RECTA                                         :14177
##
##             TIPO_INTERSEC         SUPERFICIE_CALZADA
##  EN T O Y          : 3350   SECA Y LIMPIA  :25236
##  EN X O +          : 4714   MOJADA         : 3895
##  ENLACE DE ENTRADA :  421   OTRO TIPO      :  327
##  ENLACE DE SALIDA  :  223   UMBRIA         :  165
##  GIRATORIA         : 2006   GRAVILLA SUELTA:  150
##  NO_ES_INTERSECCION:18983   ACEITE         :   83
```

```
##  OTROS             :  305   (Other)        :  146
##                            LUMINOSIDAD      FACTORES_ATMOSFERICOS
##  CREPUSCULO                    : 1330   BUEN TIEMPO  :25852
##  NOCHE: ILUMINACION INSUFICIENTE: 1067   LLOVIZNANDO  : 2524
##  NOCHE: ILUMINACION SUFICIENTE  : 4793   OTRO         :  715
##  NOCHE: SIN ILUMINACION         : 1815   LLUVIA FUERTE:  499
##  PLENO DIA                     :20997   VIENTO FUERTE:  156
##                                         NIEBLA LIGERA:   83
##                                         (Other)      :  173
##             TIPO_ACCIDENTE
##  Atropello         : 3642
##  Colision_Obstaculo:  952
##  Colision_Vehiculos:16520
##  Otro              : 1807
##  Salida_Via        : 6013
##  Vuelco            : 1068
##
```

Podemos pensar que otras de las variables que puede que no nos sean de mucha utilidad pueden ser: ANIO, MES, HORA, DIASEMANA, PROVINCIA, COMUNIDAD_AUTONOMA, ISLA, ZONA_AGRUPADA, TIPO_VIA, TRAZADO_NO_INTERSEC, TIPO_INTERSEC, SUPERFICIE_CALZADA y LUMINOSIDAD. Ya que muchas de estas variables podrían no ser de vital importancia, de primera mano, para la obtención de la predicción del tipo de accidente. Por lo tanto, vamos a eliminarlas de momento para agilizar los modelos primeros.

```
segundas.variables.eliminadas <- c("ANIO", "MES", "HORA", "DIASEMANA", "PROVINCIA", "COMUNIDAD_AUTONOMA
accidentes.train.sin.variables.2 <- accidentes.train.sin.variables.1[,-c(1,2,3,4,5,6,7,14,16,17,18,19,2
accidentes.train.variables.eliminadasa <- cbind(accidentes.train.variables.eliminadas ,accidentes.train
accidentes.test.sin.variables.2 <- accidentes.test.sin.variables.1[,-c(1,2,3,4,5,6,7,14,16,17,18,19,20)]
accidentes.test.variables.eliminadas <- cbind(accidentes.test.sin.variables.2 ,accidentes.test.original
```

Donde podemos ver ahora el resumen del dataset resultante:

```
summary(accidentes.train.sin.variables.2)
```

```
##   TOT_VICTIMAS     TOT_MUERTOS      TOT_HERIDOS_GRAVES TOT_HERIDOS_LEVES
##  Min.   : 1.000   Min.   :0.00000   Min.   :0.0000     Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.:0.00000   1st Qu.:0.0000     1st Qu.: 1.00
##  Median : 1.000   Median :0.00000   Median :0.0000     Median : 1.00
##  Mean   : 1.429   Mean   :0.02447   Mean   :0.1453     Mean   : 1.26
##  3rd Qu.: 2.000   3rd Qu.:0.00000   3rd Qu.:0.0000     3rd Qu.: 1.00
##  Max.   :19.000   Max.   :7.00000   Max.   :9.0000     Max.   :18.00
##
##  TOT_VEHICULOS_IMPLICADOS          ZONA
##  Min.   : 1.000           CARRETERA  :13278
##  1st Qu.: 1.000           TRAVESIA   :  241
##  Median : 2.000           VARIANTE   :   57
##  Mean   : 1.738           ZONA URBANA:16426
##  3rd Qu.: 2.000
##  Max.   :21.000
##
##                                              RED_CARRETERA
##  OTRAS TITULARIDADES                                :  318
```

```
##  TITULARIDAD AUTONOMICA                                : 3890
##  TITULARIDAD ESTATAL                                   : 4021
##  TITULARIDAD MUNICIPAL                                 :19077
##  TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
##
##    FACTORES_ATMOSFERICOS            TIPO_ACCIDENTE
##  BUEN TIEMPO  :25852     Atropello         : 3642
##  LLOVIZNANDO  : 2524     Colision_Obstaculo:  952
##  OTRO         :  715     Colision_Vehiculos:16520
##  LLUVIA FUERTE:  499     Otro              : 1807
##  VIENTO FUERTE:  156     Salida_Via        : 6013
##  NIEBLA LIGERA:   83     Vuelco            : 1068
##  (Other)      :  173
```

2.2 Prueba del modelo con eliminación de variables

Hagamos por lo tanto una prueba de como afecta la inclusión de estas variables con respecto a la primera prueba realziada.

```
set.seed(1234)
ct2 <- ctree(TIPO_ACCIDENTE ~., accidentes.train.sin.variables.2)
testPred2 <- predict(ct2, newdata = accidentes.test.sin.variables.2)
```

Por lo que ya tenemos el conjunto de test predecido. Además el árbol creado tendría la siguiente estructura:

```
ct2
```

```
##
##   Conditional inference tree with 36 terminal nodes
##
## Response:  TIPO_ACCIDENTE
## Inputs:  TOT_VICTIMAS, TOT_MUERTOS, TOT_HERIDOS_GRAVES, TOT_HERIDOS_LEVES, TOT_VEHICULOS_IMPLICADOS,
## Number of observations:  30002
##
## 1) TOT_VEHICULOS_IMPLICADOS <= 1; criterion = 1, statistic = 14488.658
##   2) ZONA == {CARRETERA, VARIANTE}; criterion = 1, statistic = 5782.443
##     3) RED_CARRETERA == {TITULARIDAD AUTONOMICA, TITULARIDAD ESTATAL, TITULARIDAD PROVINCIAL (DIPUTA
##       4) TOT_HERIDOS_LEVES <= 1; criterion = 1, statistic = 85.21
##         5) TOT_HERIDOS_LEVES <= 0; criterion = 1, statistic = 78.662
##           6) TOT_HERIDOS_GRAVES <= 0; criterion = 0.998, statistic = 29.773
##             7)*  weights = 163
##           6) TOT_HERIDOS_GRAVES > 0
##             8) TOT_VICTIMAS <= 1; criterion = 0.984, statistic = 36.399
##               9)*  weights = 695
##             8) TOT_VICTIMAS > 1
##               10)*  weights = 91
##         5) TOT_HERIDOS_LEVES > 0
##           11) RED_CARRETERA == {TITULARIDAD AUTONOMICA, TITULARIDAD ESTATAL}; criterion = 1, statisti
##             12) RED_CARRETERA == {TITULARIDAD AUTONOMICA}; criterion = 0.954, statistic = 44.317
##               13)*  weights = 1232
##             12) RED_CARRETERA == {TITULARIDAD ESTATAL}
```

```
##               14)*  weights = 1027
##           11) RED_CARRETERA == {TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL)}
##             15)*  weights = 809
##       4) TOT_HERIDOS_LEVES > 1
##         16)*  weights = 912
##     3) RED_CARRETERA == {OTRAS TITULARIDADES, TITULARIDAD MUNICIPAL}
##       17) TOT_HERIDOS_GRAVES <= 0; criterion = 1, statistic = 64.01
##         18) RED_CARRETERA == {TITULARIDAD MUNICIPAL}; criterion = 0.969, statistic = 59.443
##           19)*  weights = 1053
##         18) RED_CARRETERA == {OTRAS TITULARIDADES}
##           20)*  weights = 134
##       17) TOT_HERIDOS_GRAVES > 0
##         21)*  weights = 130
##   2) ZONA == {TRAVESIA, ZONA URBANA}
##     22) FACTORES_ATMOSFERICOS == {GRANIZANDO, LLOVIZNANDO, NEVANDO, NIEBLA INTENSA, NIEBLA LIGERA, V
##       23) TOT_VICTIMAS <= 1; criterion = 1, statistic = 36.573
##         24)*  weights = 433
##       23) TOT_VICTIMAS > 1
##         25)*  weights = 65
##     22) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, LLUVIA FUERTE, OTRO}
##       26) TOT_VICTIMAS <= 2; criterion = 1, statistic = 78.751
##         27) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, LLUVIA FUERTE}; criterion = 1, statistic = 38.418
##           28) TOT_VICTIMAS <= 1; criterion = 0.997, statistic = 25.281
##             29) ZONA == {ZONA URBANA}; criterion = 0.976, statistic = 25.971
##               30)*  weights = 3975
##             29) ZONA == {TRAVESIA}
##               31)*  weights = 64
##           28) TOT_VICTIMAS > 1
##             32)*  weights = 516
##         27) FACTORES_ATMOSFERICOS == {OTRO}
##           33)*  weights = 172
##       26) TOT_VICTIMAS > 2
##         34)*  weights = 112
## 1) TOT_VEHICULOS_IMPLICADOS > 1
##   35) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, GRANIZANDO, LLOVIZNANDO, LLUVIA FUERTE, NEVANDO, NIEBLA
##     36) TOT_HERIDOS_LEVES <= 1; criterion = 1, statistic = 130.164
##       37) TOT_HERIDOS_LEVES <= 0; criterion = 1, statistic = 77.217
##         38) RED_CARRETERA == {TITULARIDAD AUTONOMICA, TITULARIDAD ESTATAL, TITULARIDAD MUNICIPAL, TI
##           39)*  weights = 1223
##         38) RED_CARRETERA == {OTRAS TITULARIDADES}
##           40)*  weights = 15
##       37) TOT_HERIDOS_LEVES > 0
##         41) TOT_VICTIMAS <= 1; criterion = 1, statistic = 77.397
##           42) ZONA == {VARIANTE, ZONA URBANA}; criterion = 1, statistic = 77.906
##             43) TOT_VEHICULOS_IMPLICADOS <= 2; criterion = 1, statistic = 107.916
##               44) FACTORES_ATMOSFERICOS == {LLOVIZNANDO}; criterion = 1, statistic = 107.204
##                 45)*  weights = 436
##               44) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, GRANIZANDO, LLUVIA FUERTE, NEVANDO, NIEBLA
##                 46)*  weights = 6610
##             43) TOT_VEHICULOS_IMPLICADOS > 2
##               47)*  weights = 591
##           42) ZONA == {CARRETERA, TRAVESIA}
##             48) RED_CARRETERA == {OTRAS TITULARIDADES, TITULARIDAD AUTONOMICA, TITULARIDAD ESTATAL, T
##               49)*  weights = 2514
```

```
##             48) RED_CARRETERA == {TITULARIDAD MUNICIPAL}
##               50)*  weights = 905
##         41) TOT_VICTIMAS > 1
##           51) FACTORES_ATMOSFERICOS == {GRANIZANDO, NEVANDO, NIEBLA INTENSA, NIEBLA LIGERA, VIENTO FU
##             52)*  weights = 12
##           51) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, LLOVIZNANDO, LLUVIA FUERTE}
##             53) ZONA == {TRAVESIA, ZONA URBANA}; criterion = 1, statistic = 37.374
##               54)*  weights = 104
##             53) ZONA == {CARRETERA}
##               55)*  weights = 270
##     36) TOT_HERIDOS_LEVES > 1
##       56) ZONA == {CARRETERA, VARIANTE}; criterion = 1, statistic = 92.374
##         57) RED_CARRETERA == {TITULARIDAD MUNICIPAL}; criterion = 1, statistic = 75.983
##           58) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, GRANIZANDO, LLOVIZNANDO, NEVANDO, VIENTO FUERTE
##             59) FACTORES_ATMOSFERICOS == {GRANIZANDO, LLOVIZNANDO}; criterion = 0.981, statistic = 3
##               60)*  weights = 48
##             59) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, NEVANDO, VIENTO FUERTE}
##               61)*  weights = 421
##           58) FACTORES_ATMOSFERICOS == {LLUVIA FUERTE}
##             62)*  weights = 20
##         57) RED_CARRETERA == {OTRAS TITULARIDADES, TITULARIDAD AUTONOMICA, TITULARIDAD ESTATAL, TITU
##           63) FACTORES_ATMOSFERICOS == {BUEN TIEMPO, NIEBLA INTENSA, NIEBLA LIGERA, VIENTO FUERTE}; 
##             64)*  weights = 1877
##           63) FACTORES_ATMOSFERICOS == {GRANIZANDO, LLOVIZNANDO, LLUVIA FUERTE, NEVANDO}
##             65)*  weights = 283
##       56) ZONA == {TRAVESIA, ZONA URBANA}
##         66)*  weights = 2693
##   35) FACTORES_ATMOSFERICOS == {OTRO}
##     67) TOT_HERIDOS_GRAVES <= 1; criterion = 1, statistic = 55.801
##       68) ZONA == {CARRETERA, TRAVESIA}; criterion = 1, statistic = 50.682
##         69)*  weights = 122
##       68) ZONA == {ZONA URBANA}
##         70)*  weights = 264
##     67) TOT_HERIDOS_GRAVES > 1
##       71)*  weights = 11
```

Vamos a escribir la salida del modelo para ver su puntuación en Kaggel.

```
salida.segundo.modelo <- as.matrix(testPred2)
write.table(salida.segundo.modelo,file="predicciones/SegundaPrediccion.txt",sep=",",quote = F)
```

El resultado de este primer modelo para la competición de Kaggel, subido el 17/02/2017 a las 17:51, con un total de 14 personas entregadas, se ha quedado en la posición 9 con una puntuación del 0.81891.

#	Δ5d	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	new	Anabel Gómez	0.83175	9	Fri, 17 Feb 2017 11:34:17 (-19.6h)
2	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
3	new	Jonathan Espinosa	0.82671	8	Thu, 16 Feb 2017 12:28:22
4	new	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
5	new	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
6	new	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
7	new	WhiteShadow	0.82247	3	Thu, 16 Feb 2017 13:06:30
8	**↓5**	**PacoPollos**	**0.81891**	**2**	**Fri, 17 Feb 2017 16:50:29**

Your Best Entry ↑

Top Ten!

You made the top ten by improving your score by 0.08645.

You just moved up 1 position on the leaderboard. Tweet this!

9	↓7	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
10	new	Jorge Jimena	0.73246	1	Fri, 17 Feb 2017 02:57:20
11	↓7	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
12	↓7	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
13	new	Xisco Fauli	0.48735	2	Wed, 15 Feb 2017 23:16:45
14	new	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 2: Segunda puntuación obtenida en Kaggel