

Memoria Competición Kaggle Preprocesamiento

Francisco Pérez Hernández

17/2/2017

Contents

1	Introducción al problema y a Kaggle	2
1.1	Lectura del dataset accidentes	2
1.2	Primera prueba con un modelo	4
1.3	Creación del archivo de salida y subida a kaggel	5
2	Análisis del dataset	6
2.1	Eliminación de valores perdidos	6
2.2	Prueba del modelo con eliminación de variables	10
3	Vusualización del dataset	11
3.1	Ánalysis de las variables actuales	11
3.2	Ánalysis de variables eliminadas sin valores perdidos	20
4	Visión preeliminar de los datos	26
5	Imputación de valores perdidos	27
5.1	Imputación de variables	28
5.2	Prueba del modelo con imputación de valores perdidos	29
6	Detección de anomalías	30
6.1	Uso del paquete outliers	30
6.2	Paquete mvoutlier	35
6.3	Eliminación de valores anómalos	35
6.4	Prueba del modelo con imputación de valores perdidos	38
7	Transformación de los datos	39
7.1	Transformando los datos	40
7.2	Prueba del modelo con transformación de los datos	40
8	Discretización	40
9	Selección de características	40
9.1	Paquete FSelector	42
9.2	Paquete caret	56
9.3	Paquete Boruta	60
10	Detección de ruido	70
11	Modelo SVM Radial	74
12	SVM (Support Vector Machine)	79
13	Métodos ensamble de construcción de conjuntos de modelos	79
13.1	Bagging	79
13.2	Random forest	83
13.3	Boosting	88

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
## SEÑAL DE STOP : 1750    OTRO TIPO : 327
## SOLO MARCAS VIALES : 1659    UMBRIA : 165
## SEÑAL 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 SUFFICIENTE : 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 predicho. 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)
salida.primer.modelo <- cbind(c(1:(dim(salida.primer.modelo)[1])), salida.primer.modelo)
colnames(salida.primer.modelo) <- c("Id", "Prediction")
write.table(salida.primer.modelo,file="predicciones/PrimeraPrediccion.txt",sep=",",quote = F,row.names =

```

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	fregagel	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 perdidos, 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 qué variables contienen valores perdidos.

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

##	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_IMPPLICADOS
##	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", "OTRA_CIRCUNSTANCIA", "ACERAS", "DENSIDAD_CIRCULACION", "MEDIDAS_ESPECIALES")])
```

```
##      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
## SEÑAL DE STOP              : 1750  OTRA_CAUSA          : 491
## SOLO MARCAS VIALES        : 1659  FACTORES ATMOSFERICOS: 374
## SEÑAL 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 ARGEN: 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 perdidos 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)]
accidentes.test.variables.eliminadas.copia <- accidentes.test.variables.eliminadas

```

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 SUFFICIENTE : 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(1,2,3,4,5,6,7,14,16,17,18,19,20)]
accidentes.train.variables.eliminadas <- cbind(accidentes.train.variables.eliminadas ,accidentes.train.sin.variables.2)
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 realizada.

```

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 predicho. Además el árbol creado tendría la siguiente estructura:

```
#ct2
```

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

```

salida.segundo.modelo <- as.matrix(testPred2)
salida.segundo.modelo <- cbind(c(1:(dim(salida.segundo.modelo)[1])), salida.segundo.modelo)
colnames(salida.segundo.modelo) <- c("Id", "Prediction")
write.table(salida.segundo.modelo, file="predicciones/SegundaPrediccion.txt", sep=",", quote = F, row.names

```

El resultado de este modelo para la competición de Kaggle, 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 Kaggle

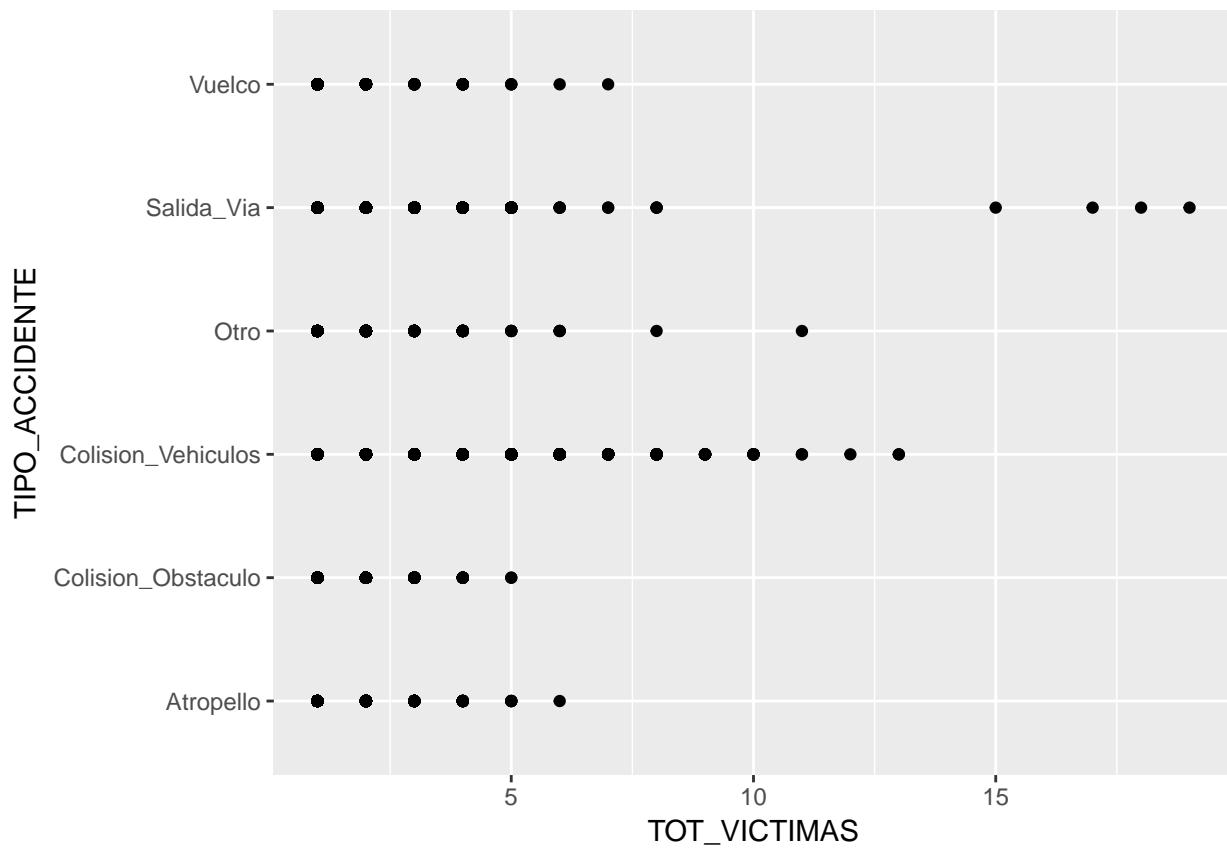
3 Vvisualización del dataset

Como no se ha hecho antes, y debería ser uno de los primeros pasos a realizar, vamos a realizar una visualización del dataset.

3.1 Análisis de las variables actuales

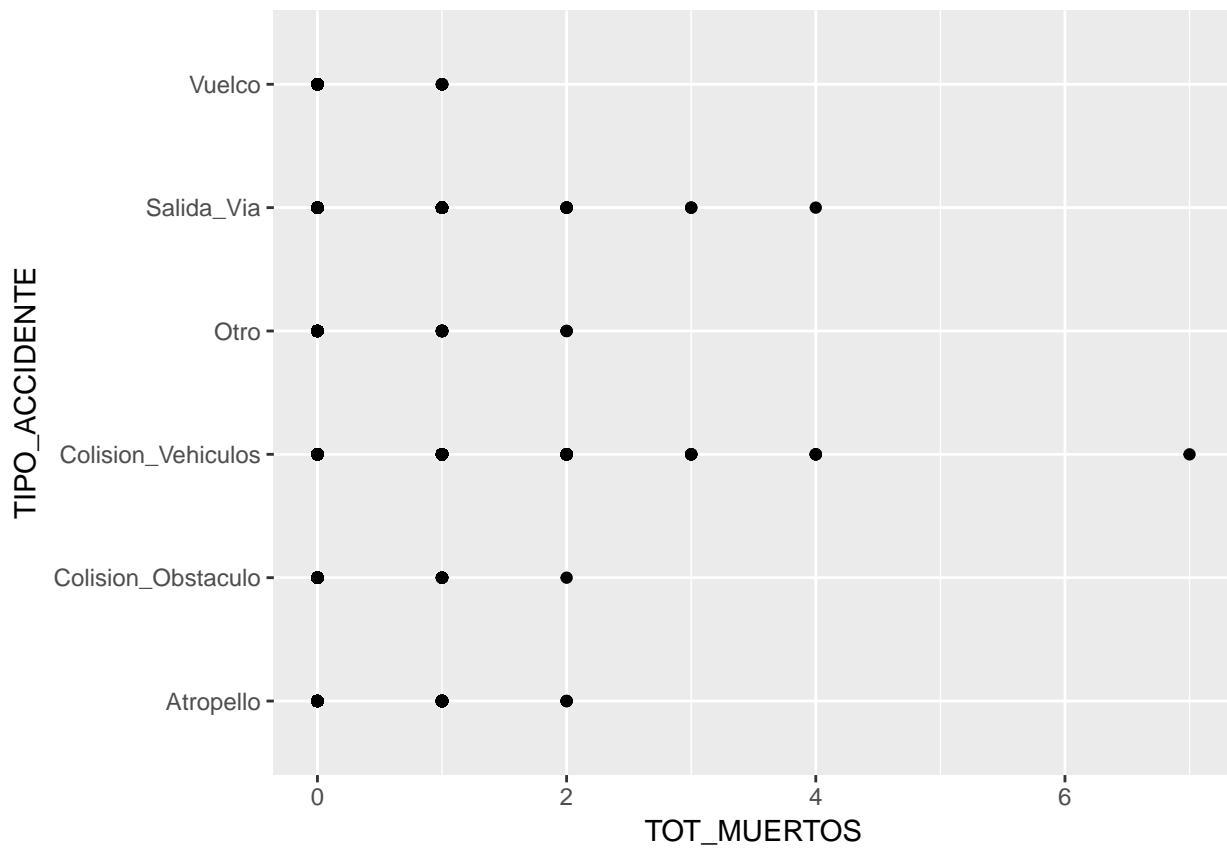
Vamos a ver el comportamiento de nuestras variables con respecto al TIPO_ACCIDENTE, a ver que relación pueden tener.

```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_VICTIMAS , y = TIPO_
```

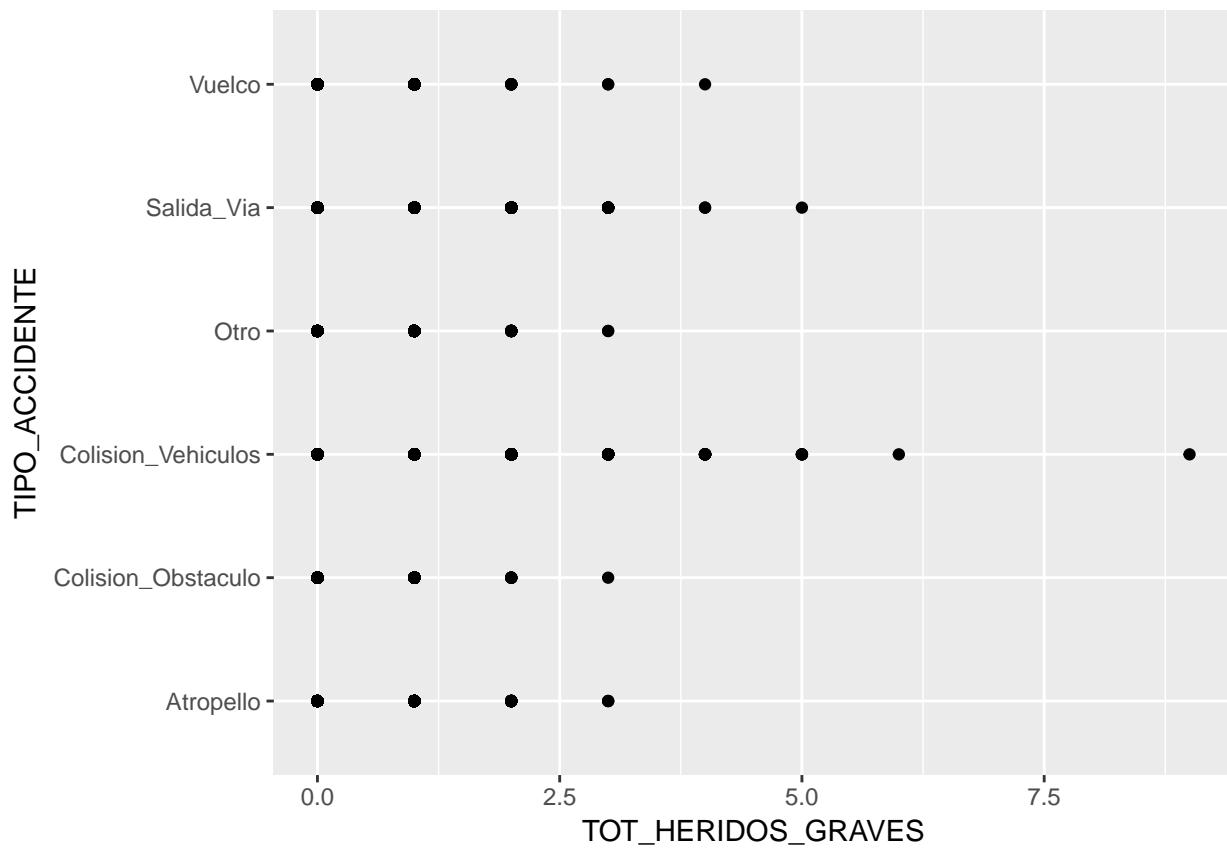


Podemos ver como para a partir de 10 victimas, el accidente suele ser o una colisión de vehículos, salida de vía, o muy pocas veces otro tipo de accidente. Por lo que puede ser una relación interesante.

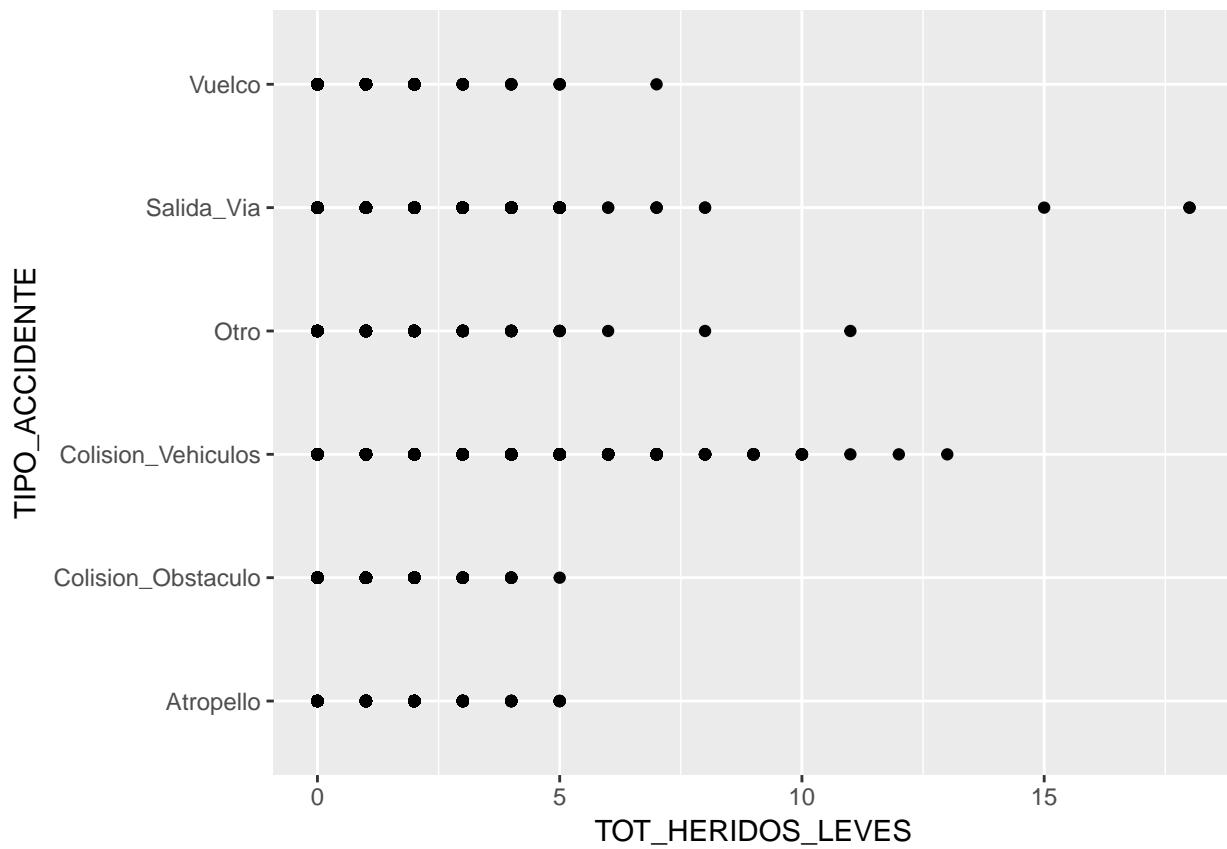
```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_MUERTOS , y = TIPO_A
```



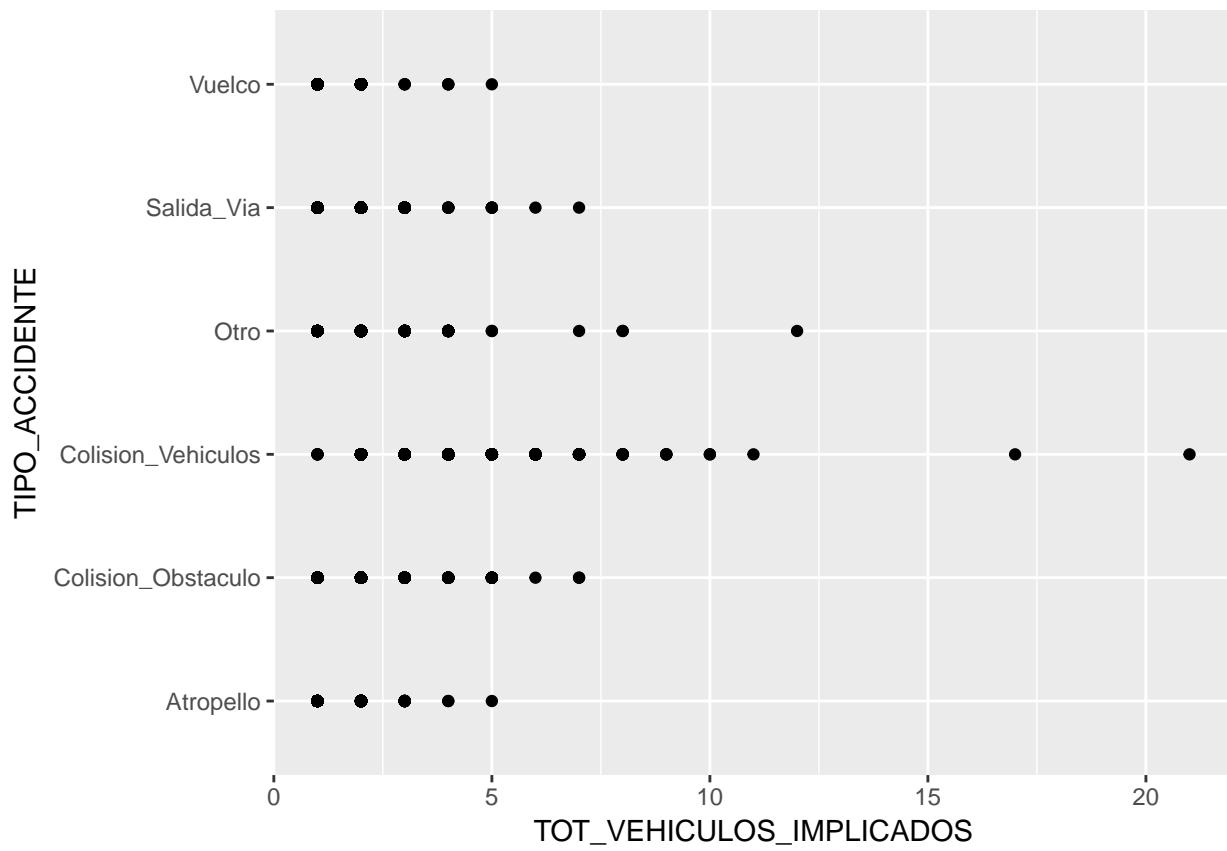
```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_HERIDOS_GRAVES , y =
```



```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_HERIDOS_LEVES , y = TOT_HERIDOS_GRAVES))
```

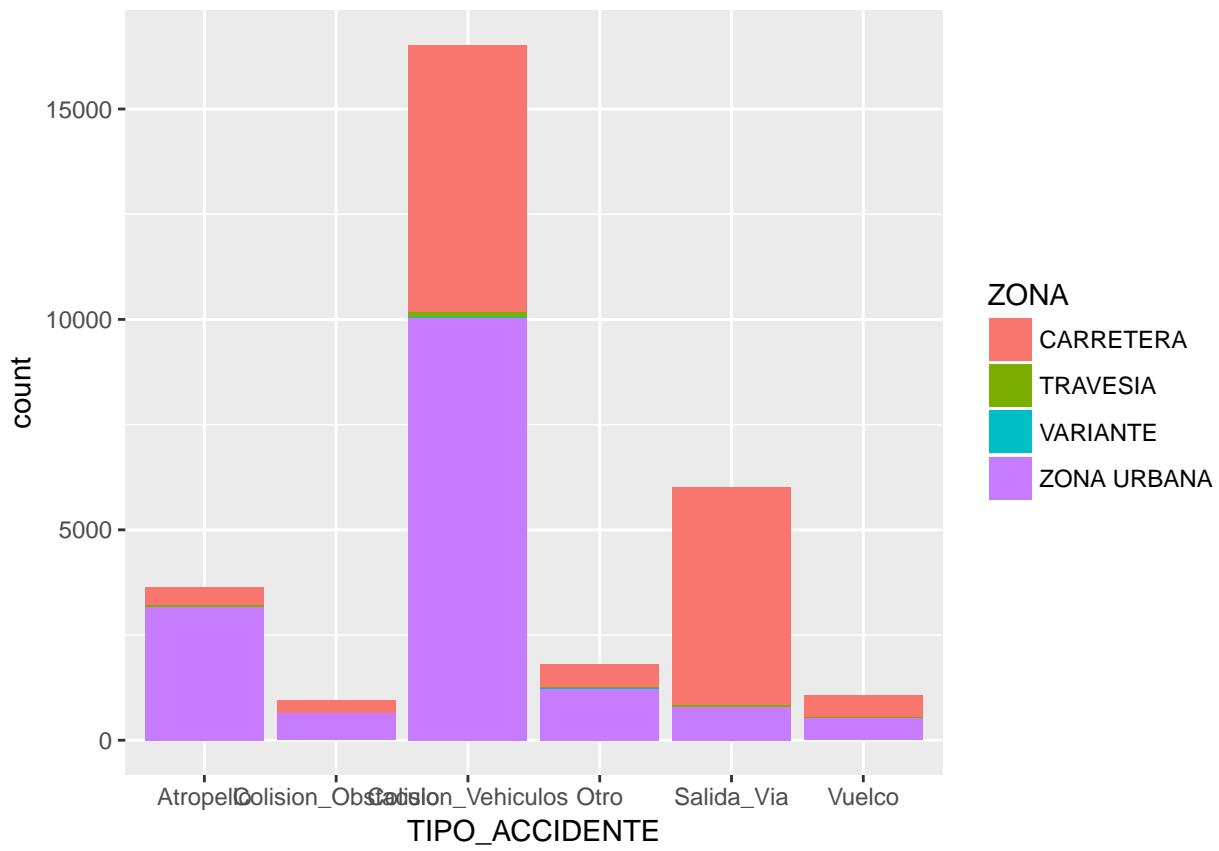


```
ggplot(data = accidentes.train.sin.variables.2) + geom_point(mapping = aes(x = TOT_VEHICULOS_IMPLICADOS
```



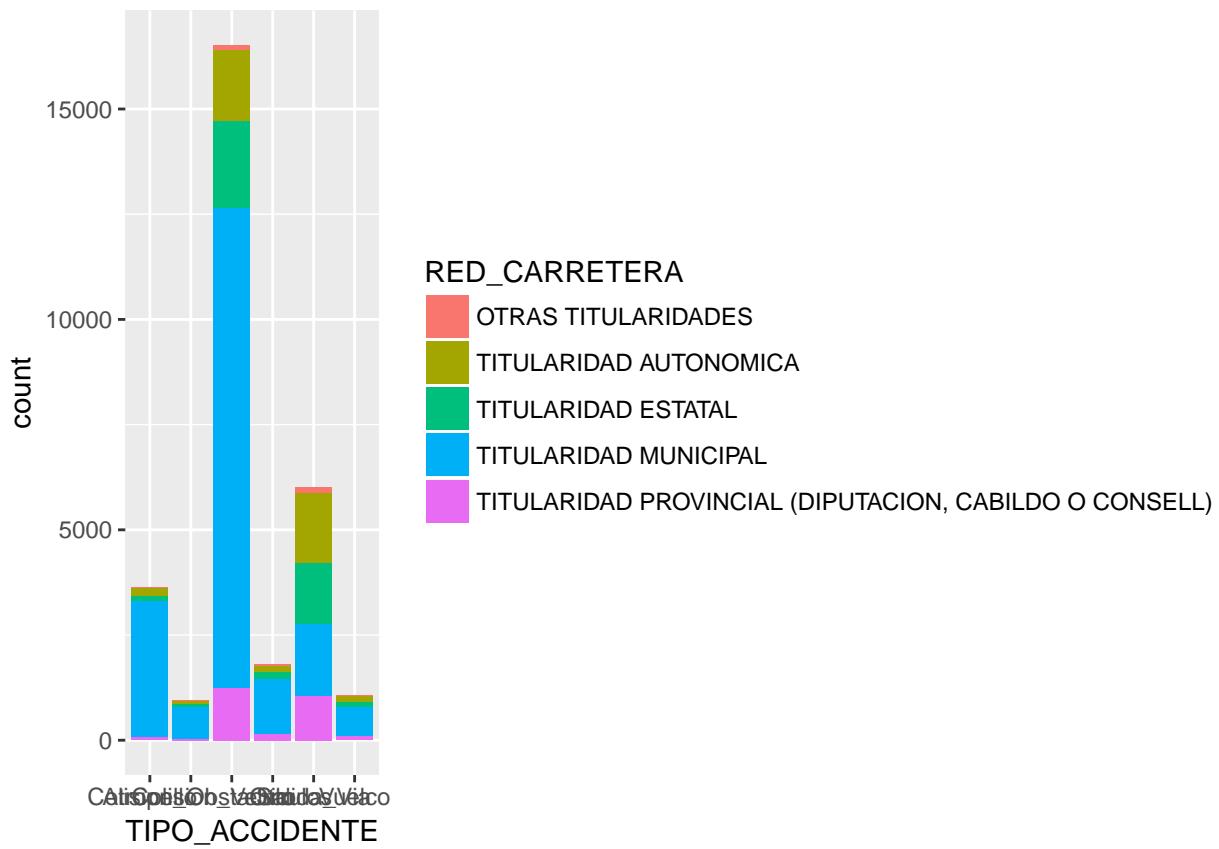
Normalmente a partir de 3 muertos, el accidente es una colisión de vehículos o una salida de vía. Si hay más de 3 heridos graves, suele ser colisión de vehículos, salida de vía o vuelco. A partir de 6 heridos leves el accidente es una colisión, una salida de vía, un vuelco o otro accidente. A partir de 6 vehículos implicados, los accidentes suelen ser colisiones, salida de vía u otro tipo. Por lo que ya tenemos varias relaciones que podrían ser representadas en un árbol.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=ZONA))
```



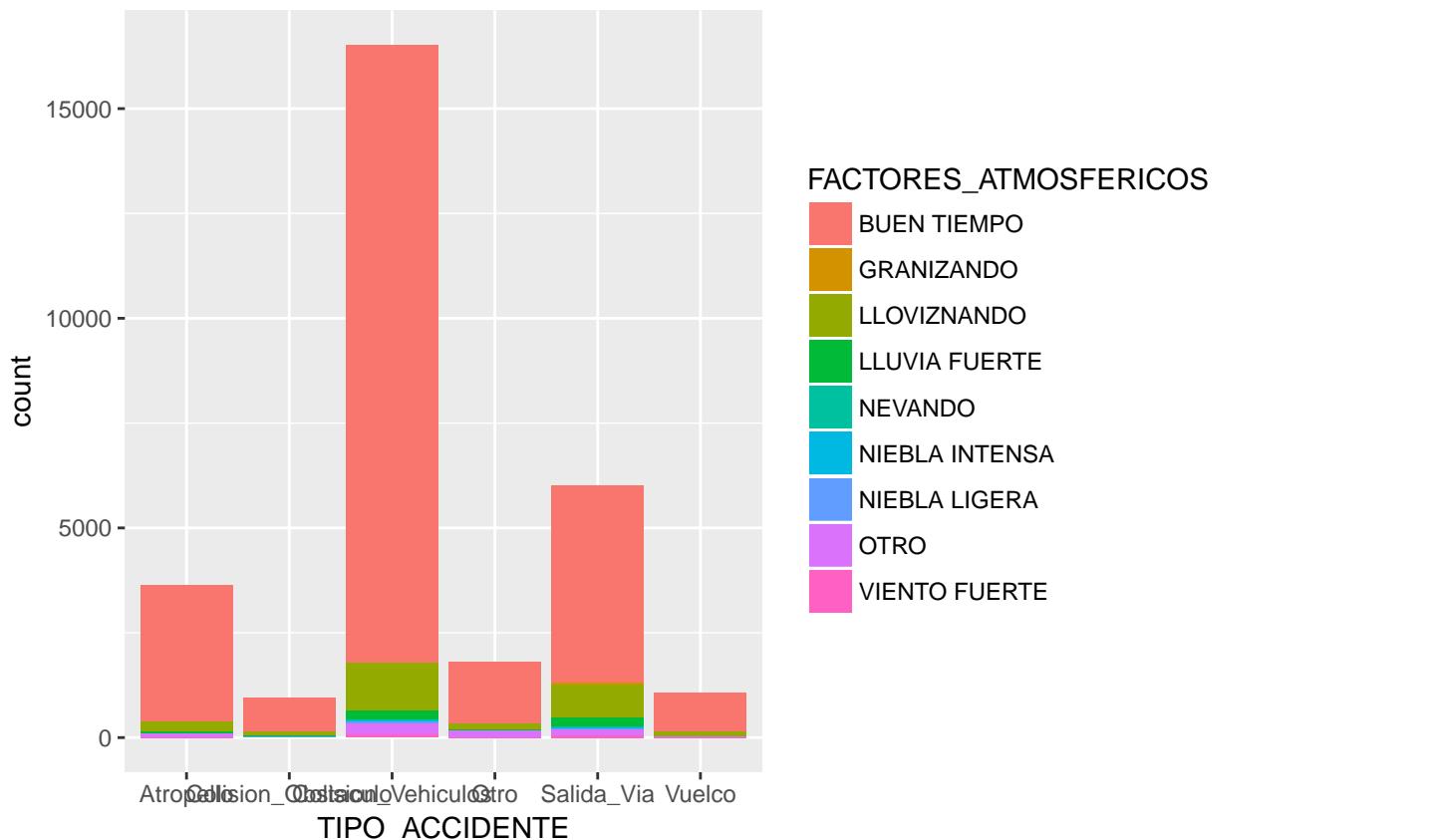
Podemos ver como las zonas predominantes son carretera y zona urbana, pero no parece que esta variable pueda ser influyente a la hora de decir que tipo de accidente se produce por lo que eliminaré esta variable para futuras pruebas.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=RED_CAR))
```



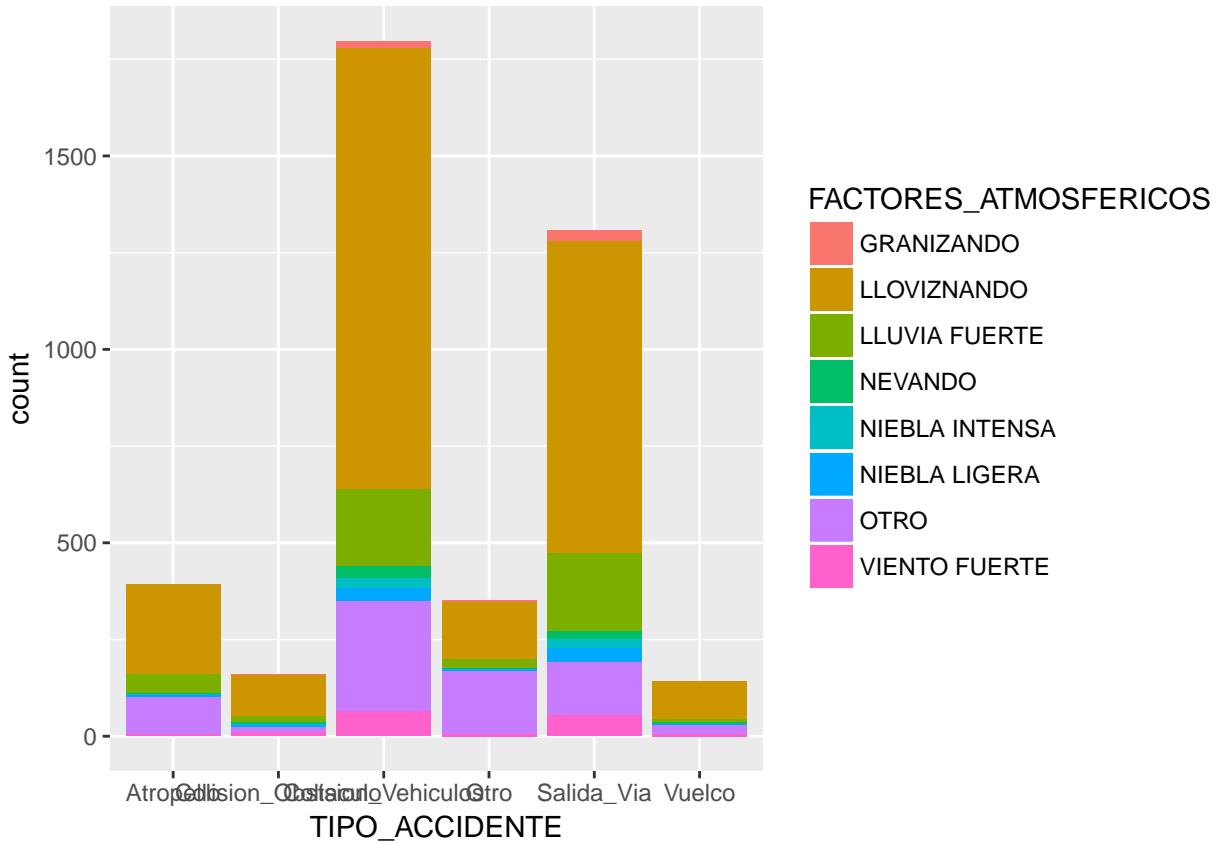
Puede parecer que esta variable no tiene demasiado que ver con la variable que queremos predecir por lo que puede ser que la descartemos.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=FACTORES))
```



Por el conocimiento que tenemos, seguramente esta variable no sea demasiado importante para el tipo de accidente. Veamos que le ocurre si eliminamos los elementos que tienen buen tiempo.

```
vector.buen.tiempo <- accidentes.train.sin.variables.1$FACTORES_ATMOSFERICOS == "BUEN TIEMPO"
valores.sin.buen.tiempo <- accidentes.train.sin.variables.1[!vector.buen.tiempo,]
ggplot(data = valores.sin.buen.tiempo) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=FACTORES_ATMOSFERICOS))
```



Pero seguimos viendo que no se puede sacar ninguna conclusión de esta visualización.

3.2 Análisis de variables eliminadas sin valores perdidos

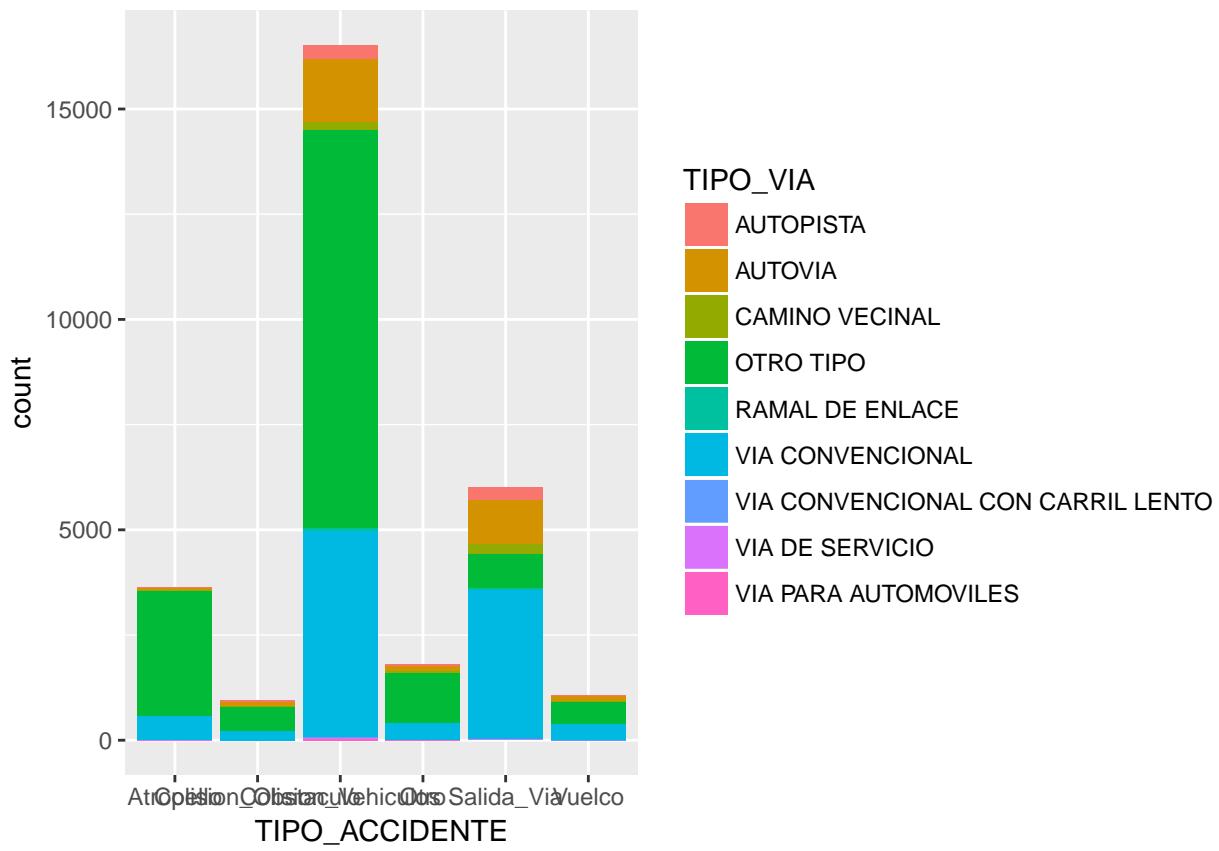
Recordemos las variables que eliminamos sin tener valores perdidos.

```
segundas.variables.eliminadas
```

```
## [1] "ANIO"                 "MES"                  "HORA"
## [4] "DIASEMANA"            "PROVINCIA"             "COMUNIDAD_AUTONOMA"
## [7] "ISLA"                  "ZONA_AGRUPADA"        "TIPO_VIA"
## [10] "TRAZADO_NO_INTERSEC"  "TIPO_INTERSEC"         "SUPERFICIE_CALZADA"
## [13] "LUMINOSIDAD"
```

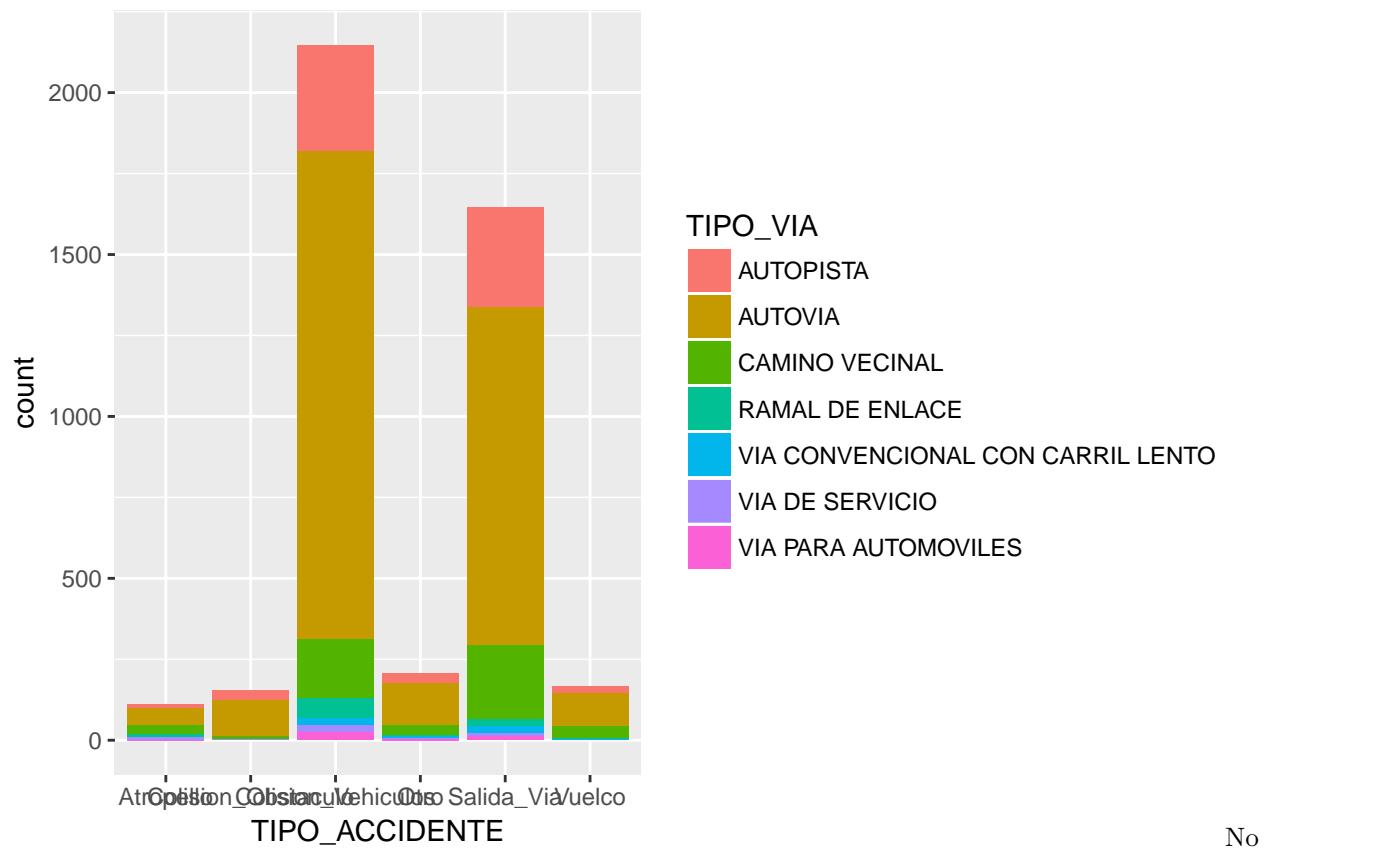
Una de la variables que podrían ser interesantes es TIPO_VIA, TRAZADO_NO_INTERSEC, TIPO_INTERSEC, SUPERFICIE_CALZADA y LUMINOSIDAD. Veamos visualizaciones de estas variables.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TIPO_VIA))
```



Eliminemos las instancias con OTRO TIPO o VIA CONVENCIONAL

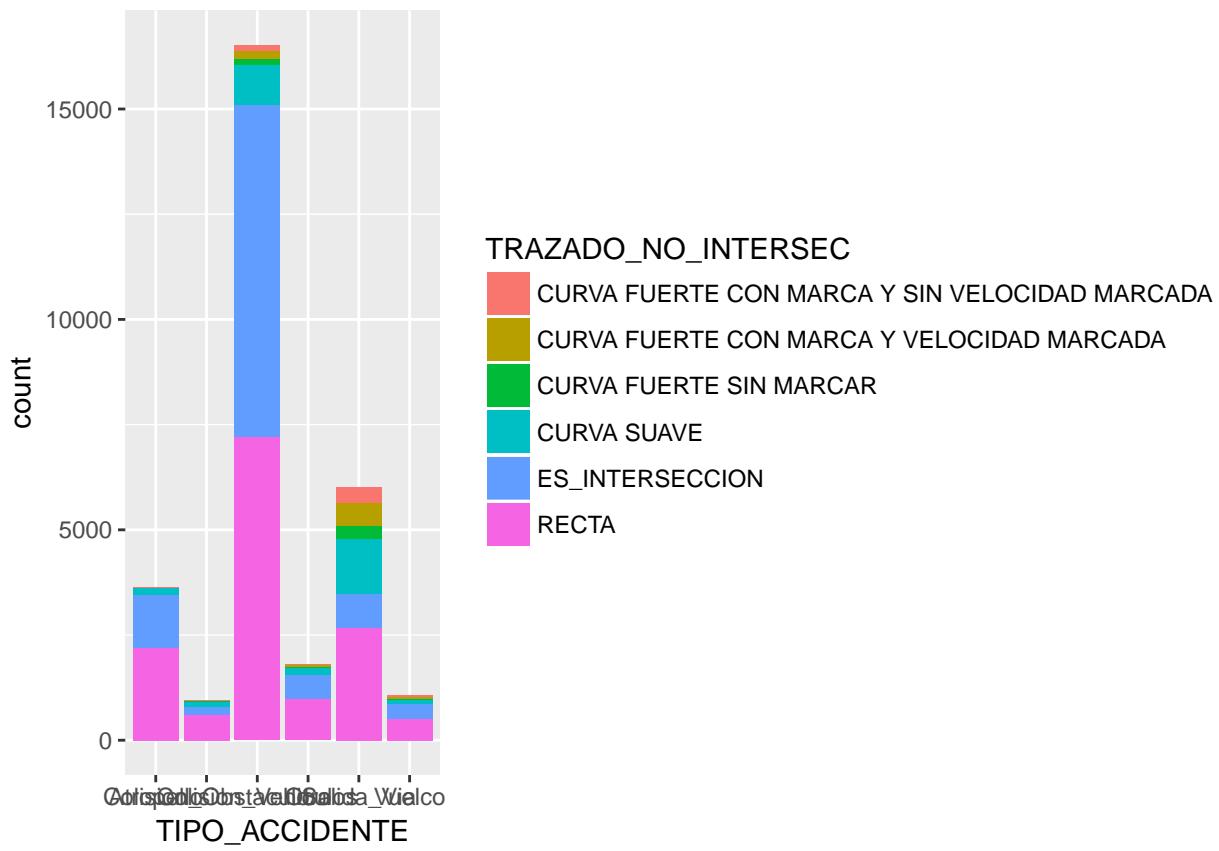
```
vector.sin.otrotipo.y.viaconvencional <- ((accidentes.train.sin.variables.1$TIPO_VIA == "OTRO TIPO") |  
valores.sin.otrotipo.y.viaconvencional <- accidentes.train.sin.variables.1[!vector.sin.otrotipo.y.viacon  
ggplot(data = valores.sin.otrotipo.y.viaconvencional) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=T
```



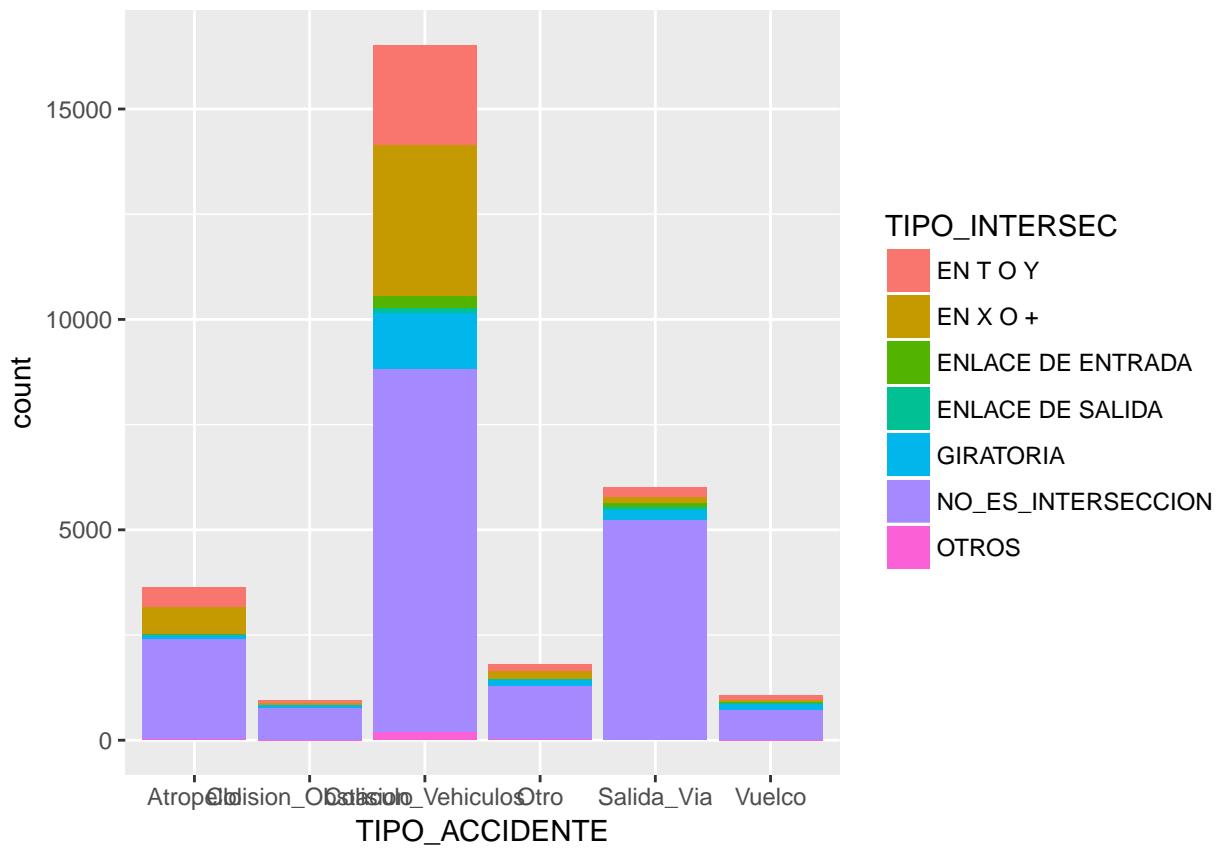
se observa que sea una variable demasiada importante.

```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TRAZADO))
```

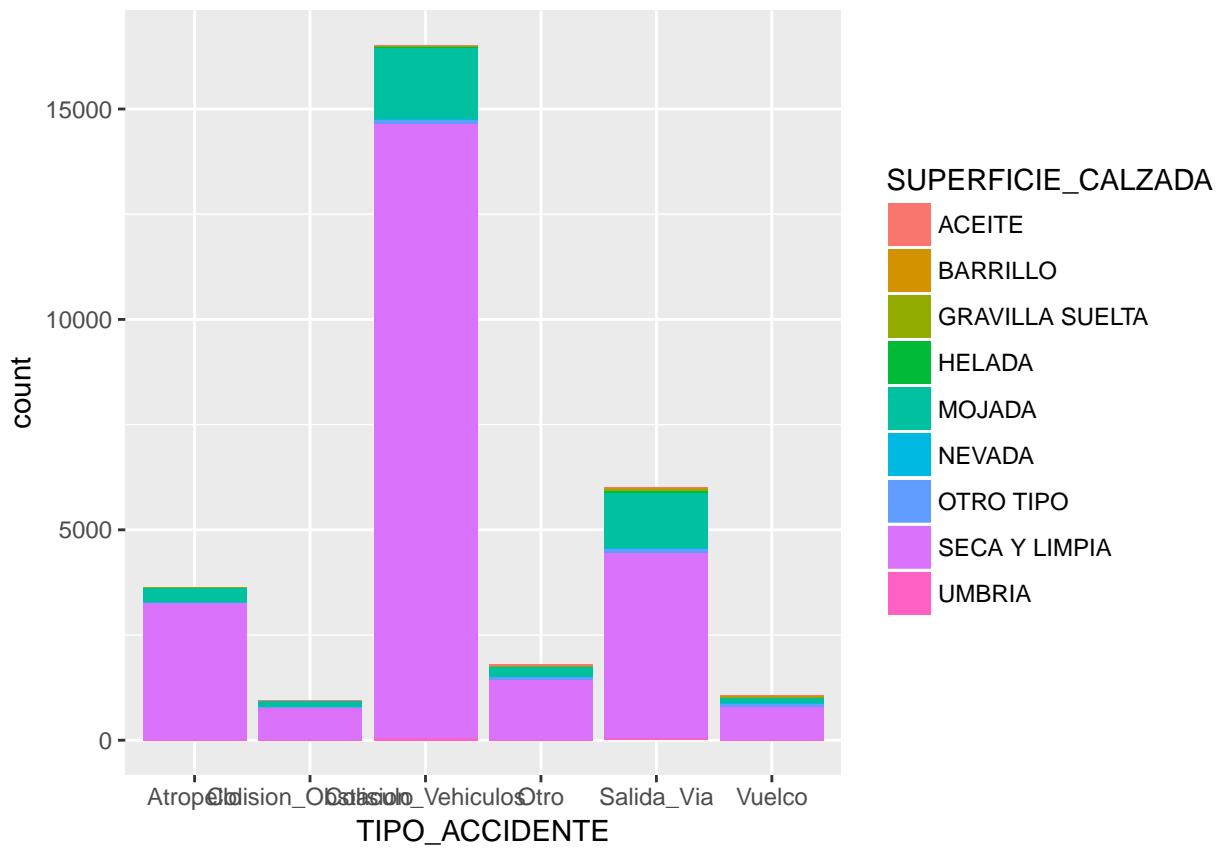
No



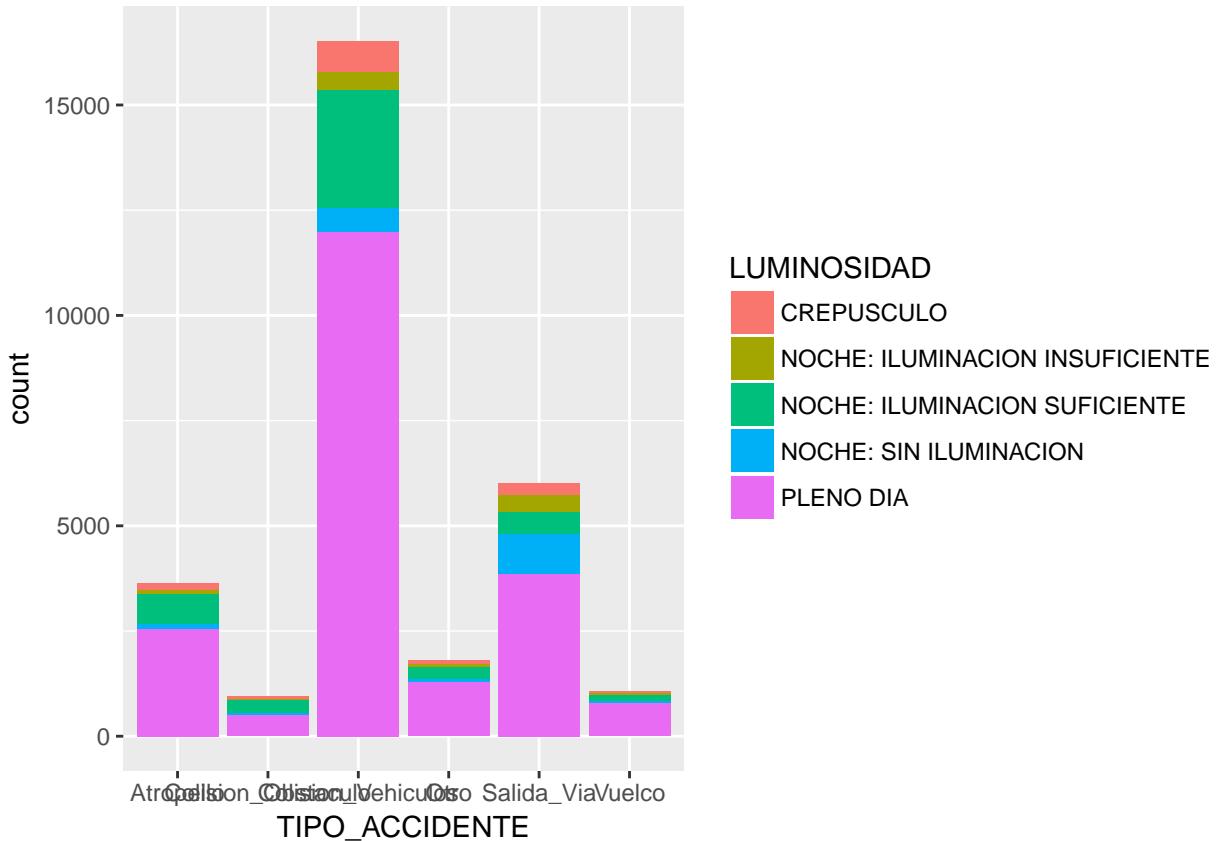
```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=TIPO_IN
```



```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=SUPERFI
```



```
ggplot(data = accidentes.train.sin.variables.1) + geom_bar(mapping = aes(x=TIPO_ACCIDENTE, fill=LUMINOS))
```



Por lo que no podemos sacar demasiada información así que no añadiremos ninguna a las que ya estamos usando de momento.

4 Visión preliminar de los datos

Como anteriormente ya hicimos el summary, no será necesario volver a hacerlo. Lo que si vamos a hacer es un str, para obtener la información de las variables.

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

```
## 'data.frame': 30002 obs. of 22 variables:
## $ ANIO : int 2009 2011 2008 2013 2009 2008 2010 2010 2013 2009 ...
## $ MES : Factor w/ 12 levels "Abril","Agosto",...: 8 5 8 10 1 6 6 7 11 10 ...
## $ HORA : Factor w/ 448 levels "0", "0,016666667", ...: 266 266 136 328 49 411 31 130 ...
## $ DIASEMANA : Factor w/ 7 levels "DOMINGO", "JUEVES", ...: 7 3 6 7 7 6 4 1 7 6 ...
## $ PROVINCIA : Factor w/ 52 levels "Albacete", "Alicante/Alacant", ...: 13 39 49 11 2 23 9 ...
## $ COMUNIDAD_AUTONOMA : Factor w/ 18 levels "Andalucia", "Aragon", ...: 1 13 11 7 11 1 9 11 14 9 ...
## $ ISLA : Factor w/ 10 levels "FORMENTERA", "FUERTEVENTURA", ...: 9 9 9 9 9 9 9 9 9 9 ...
## $ TOT_VICTIMAS : int 1 1 1 3 1 2 3 1 1 1 ...
## $ TOT_MUERTOS : int 0 0 1 0 0 1 0 0 0 0 ...
## $ TOT_HERIDOS_GRAVES : int 0 0 0 0 0 1 0 0 0 0 ...
## $ TOT_HERIDOS_LEVES : int 1 1 0 3 1 0 3 1 1 1 ...
## $ TOT_VEHICULOS_IMPPLICADOS: int 2 2 1 3 1 1 3 2 1 4 ...
## $ ZONA : Factor w/ 4 levels "CARRETERA", "TRAVESIA", ...: 4 1 1 4 1 1 4 4 4 4 ...
## $ ZONA_AGRUPADA : Factor w/ 2 levels "VIAS INTERURBANAS", ...: 2 1 1 2 1 1 2 2 2 2 ...
## $ RED_CARRETERA : Factor w/ 5 levels "OTRAS TITULARIDADES", ...: 4 2 5 4 3 5 4 4 4 4 ...
## $ TIPO_VIA : Factor w/ 9 levels "AUTOPISTA", "AUTOVIA", ...: 4 6 6 4 1 6 4 4 4 4 ...
```

```

## $ TRAZADO_NO_INTERSEC      : Factor w/ 6 levels "CURVA FUERTE CON MARCA Y SIN VELOCIDAD MARCADA",...
## $ TIPO_INTERSEC            : Factor w/ 7 levels "EN T O Y","EN X O +",...: 6 1 6 6 6 1 2 6 6 ...
## $ SUPERFICIE_CALZADA      : Factor w/ 9 levels "ACEITE","BARRILLO",...: 8 8 8 5 8 8 8 8 8 ...
## $ LUMINOSIDAD              : Factor w/ 5 levels "CREPUSCULO","NOCHE: ILUMINACION INSUFICIENTE",...: 5
## $ FACTORES_ATMOSFERICOS    : Factor w/ 9 levels "BUEN TIEMPO",...: 1 1 1 3 1 1 1 1 1 ...
## $ TIPO_ACCIDENTE           : Factor w/ 6 levels "Atropello","Colision_Obstaculo",...: 3 3 5 3 5 5 3 3

```

Si queremos información más detallada:

```
describe(accidentes.train.sin.variables.2[1])
```

```

## accidentes.train.sin.variables.2[1]
##
## 1 Variables      30002 Observations
## -----
## TOT_VICTIMAS
##      n missing distinct      Info      Mean      Gmd      .05      .10
## 30002      0       17     0.609     1.429     0.6909      1       1
##   .25      .50       .75      .90      .95
##   1       1       2       2       3
##
## Value      1       2       3       4       5       6       7       8       9       10
## Frequency 21826  5503  1540   681   248   105   43   25   13   8
## Proportion 0.727  0.183  0.051  0.023  0.008  0.003  0.001  0.001  0.000  0.000
## Value      11      12      13      15      17      18      19
## Frequency  3       1       2       1       1       1       1
## Proportion 0.000  0.000  0.000  0.000  0.000  0.000  0.000
## -----

```

Esto lo podemos hacer con las variables que veamos oportunas. Otra forma de ver más información es:

```
basicStats(accidentes.train.sin.variables.2[1])
```

```

##          TOT_VICTIMAS
## nobs      30002.000000
## NAs       0.000000
## Minimum   1.000000
## Maximum   19.000000
## 1. Quartile 1.000000
## 3. Quartile 2.000000
## Mean      1.429371
## Median    1.000000
## Sum       42884.000000
## SE Mean   0.005258
## LCL Mean  1.419066
## UCL Mean  1.439677
## Variance  0.829334
## Stdev     0.910678
## Skewness   3.817690
## Kurtosis  27.886723

```

5 Imputación de valores perdidos

Vamos a usar uso del paquete mice para imputar los datos.

5.1 Imputación de variables

Veamos que variables teníamos con valores perdidos.

```
summary(accidentes.train.variables.eliminadas)
```

```
##      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
## SEÑAL DE STOP              : 1750  OTRA_CAUSA          : 491
## SOLO MARCAS VIALES        : 1659  FACTORES ATMOSFERICOS : 374
## SEÑAL 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 ARGEN: 8
## NINGUNA MEDIDA      :21024
## OTRA MEDIDA          : 278
## NA's                 : 8675
##
```

Vemos que dos de estas variables que podrían ser más interesantes son visibilidad restringida y prioridad, por lo que vamos a proceder a imputar sus valores perdidos.

```
accidentes.train.a.imputar <- cbind(accidentes.train.sin.variables.2, accidentes.train.variables.eliminadas)
accidentes.test.a.imputar <- cbind(accidentes.test.sin.variables.2, accidentes.test.variables.eliminadas)
set.seed(1234)
train.imputados.incompletos <- mice::mice(accidentes.train.a.imputar, m=1, method="pmm")
##
```

```

## iter imp variable
## 1 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA

train.imputados <- mice::complete(train.imputados.incompletos)
test.imputados.incompletos <- mice::mice(accidentes.test.a.imputar, m=5, method="pmm")

##
## iter imp variable
## 1 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 1 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 2 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 3 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 4 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 1 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 2 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 3 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 4 PRIORIDAD VISIBILIDAD_RESTRINGIDA
## 5 5 PRIORIDAD VISIBILIDAD_RESTRINGIDA

test.imputados <- mice::complete(test.imputados.incompletos)

```

5.2 Prueba del modelo con imputación de valores perdidos

Hagamos por lo tanto una prueba de como afecta la imputación de valores perdidos.

```

set.seed(1234)
ct3 <- ctree(TIPO_ACCIDENTE ~., train.imputados)
testPred3 <- predict(ct3, newdata = test.imputados)

```

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

```
#ct3
```

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

```

salida.tercer.modelo <- as.matrix(testPred3)
salida.tercer.modelo <- cbind(c(1:(dim(salida.tercer.modelo)[1])), salida.tercer.modelo)

```

```

colnames(salida.tercer.modelo) <- c("Id", "Prediction")
write.table(salida.tercer.modelo,file="predicciones/TerceraPrediccion.txt",sep=",",quote = F,row.names =

```

El resultado de este modelo para la competición de Kaggle, subido el 19/02/2017 a las 17:42, con un total de 14 personas entregadas, se ha quedado en la posición 9 con una puntuación del 0.81753. Bajando muy poco con respecto a la anterior puntuación.

#	Δ5d	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	new	Anabel Gómez	0.83175	12	Sun, 19 Feb 2017 13:06:40 (-2.9d)
2	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
3	new	Jonathan Espinosa	0.82780	12	Sun, 19 Feb 2017 11:41:59
4	↓2	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
5	↓1	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
6	↓3	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
7	new	WhiteShadow	0.82345	6	Sat, 18 Feb 2017 14:23:36 (-17.9h)
8	new	Jorge Jimena	0.82059	4	Sun, 19 Feb 2017 16:12:15 (-0.2h)
9	↓3	PacoPollos	0.81891	3	Sun, 19 Feb 2017 16:41:50 (-47.9h)
Your Best Entry ↑					
Your submission scored 0.81753 , which is not an improvement of your best score. Keep trying!					
10	↓5	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
11	↓4	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
12	↓4	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	↓5	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 3: Tercera puntuación obtenida en Kaggle

6 Detección de anomalías

Veamos como detectar valores anómalos en nuestros datos.

6.1 Uso del paquete outliers

Veamos si tenemos valores perdidos en nuestros datos, solo con valores que no son discretas.

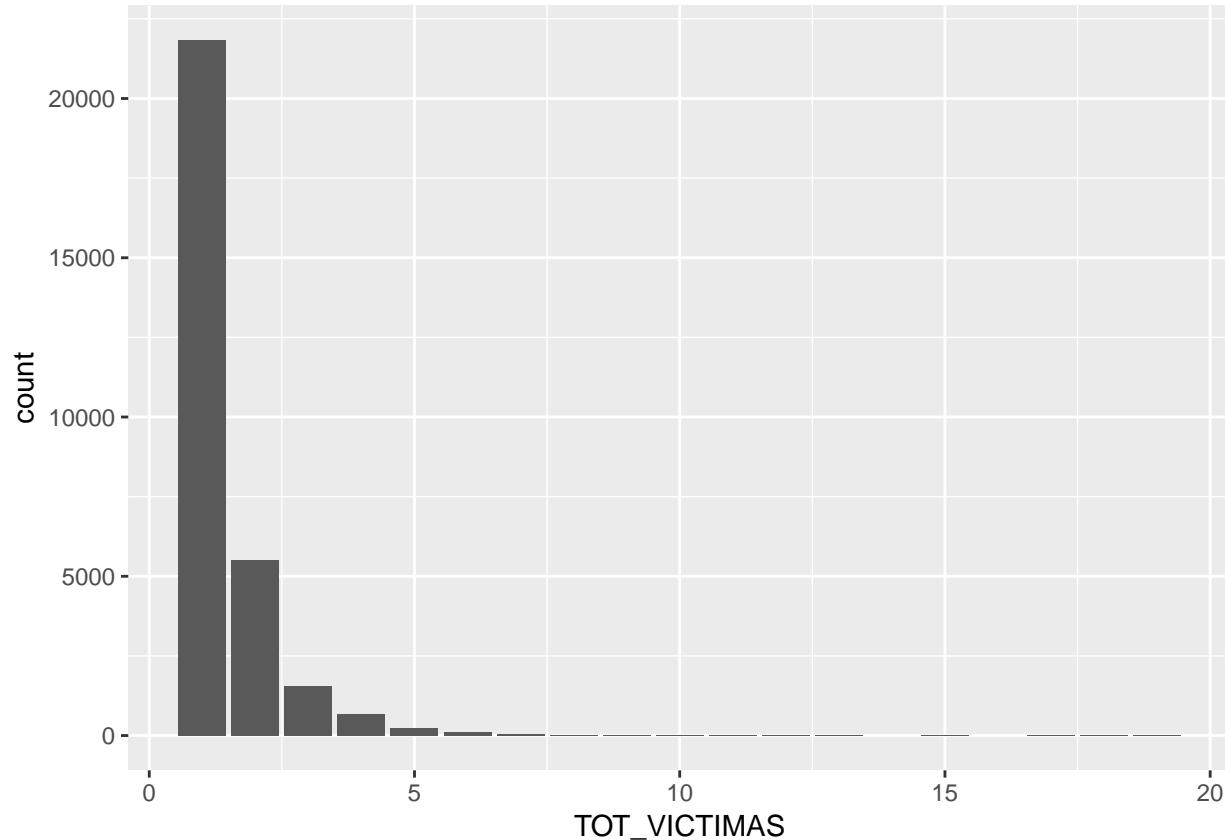
```

valores.anomalous <- outlier(train.imputados[,1:5])
print(valores.anomalous)

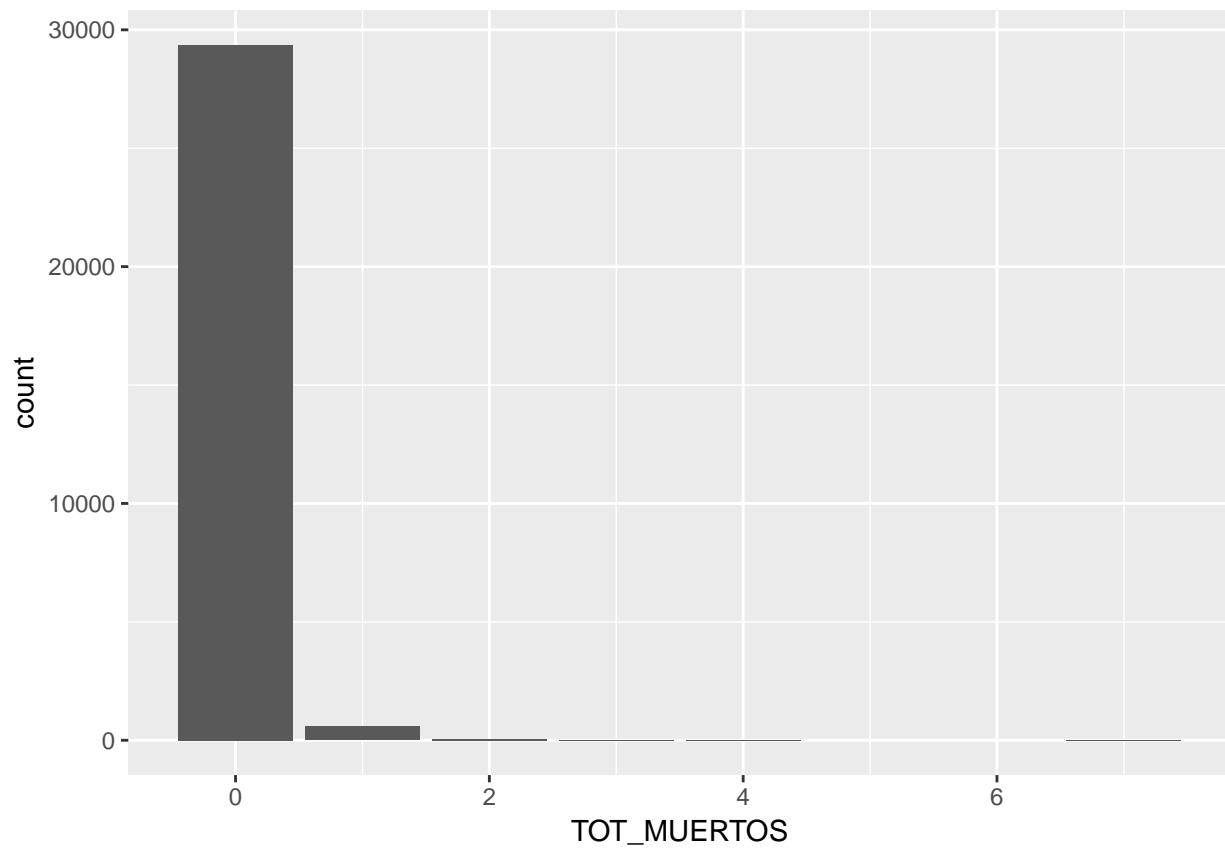
```

##	TOT_VICTIMAS	TOT_MUERTOS	TOT_HERIDOS_GRAVES
##	19	7	9

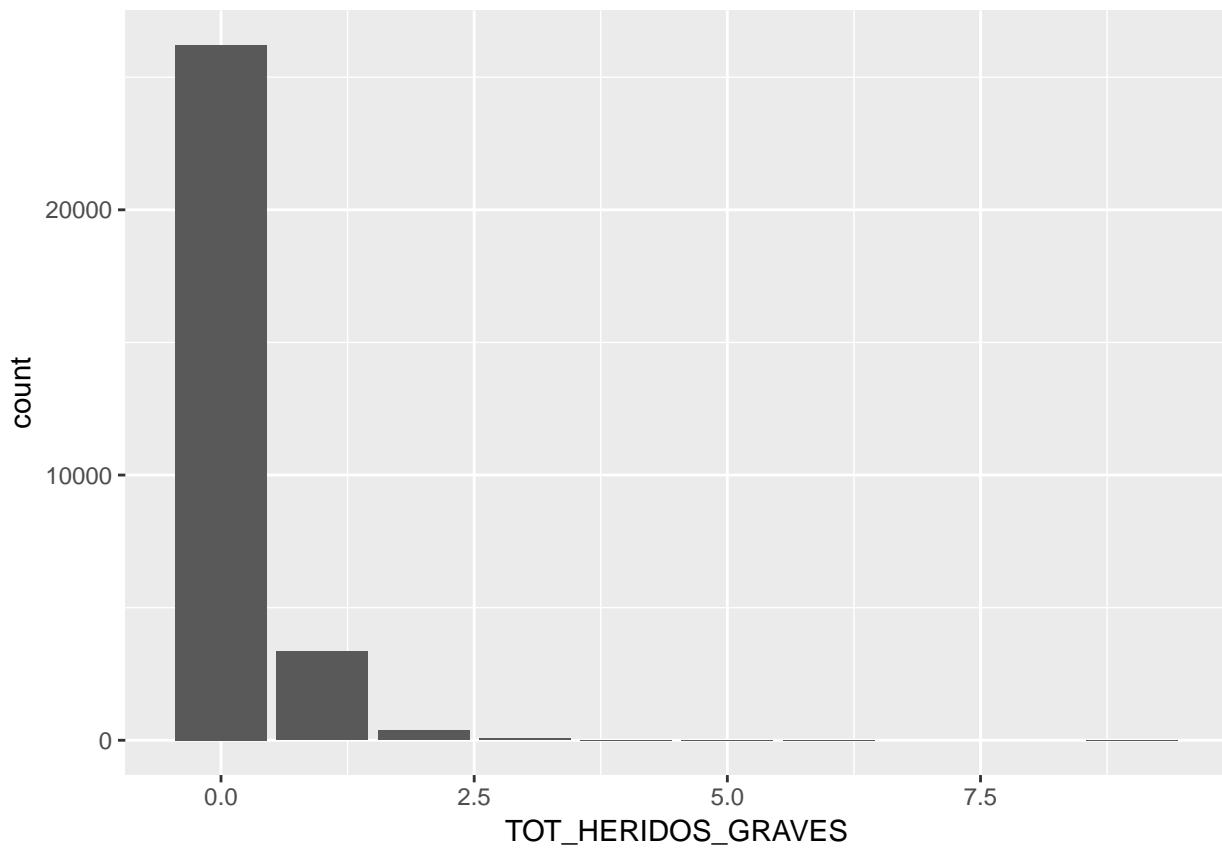
```
##          TOT_HERIDOS_LEVES TOT_VEHICULOS_IMPLICADOS
##                           18                      21
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_VICTIMAS))
```

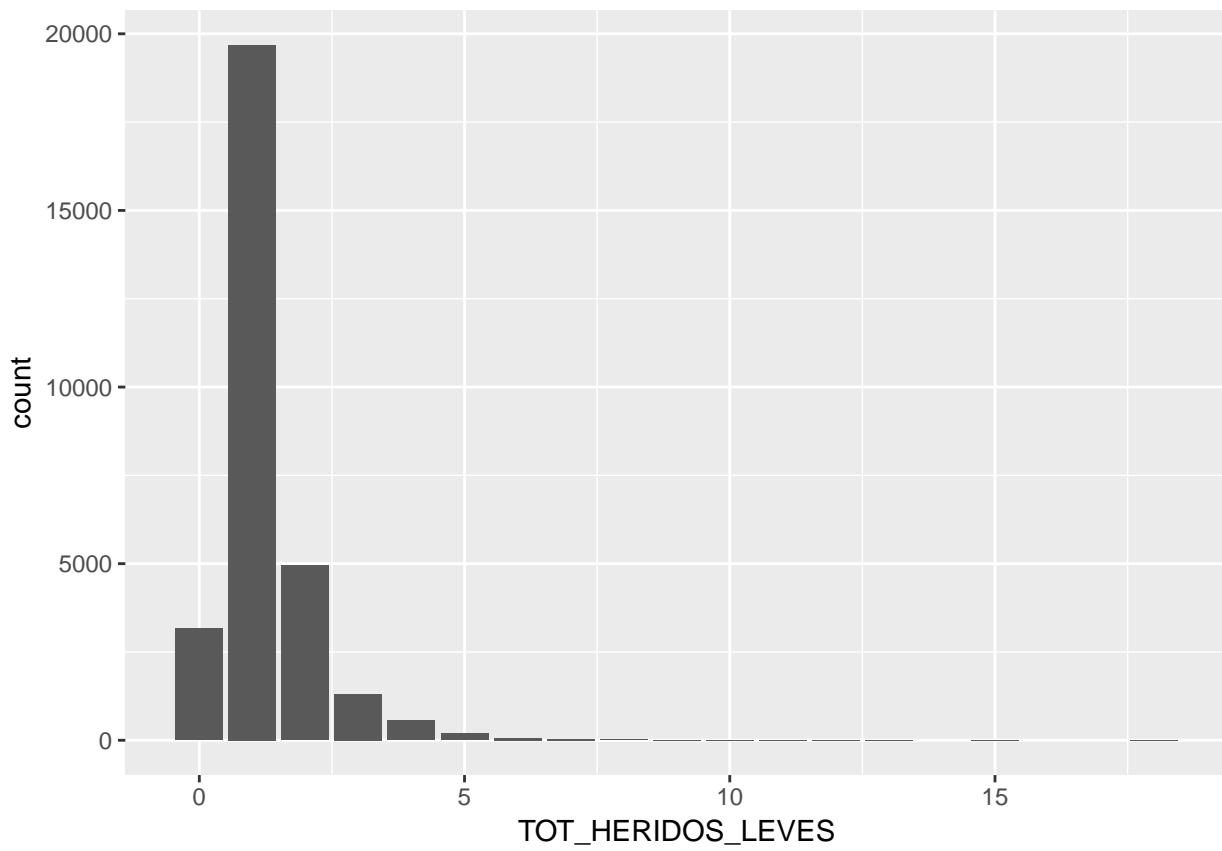


```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_MUERTOS))
```

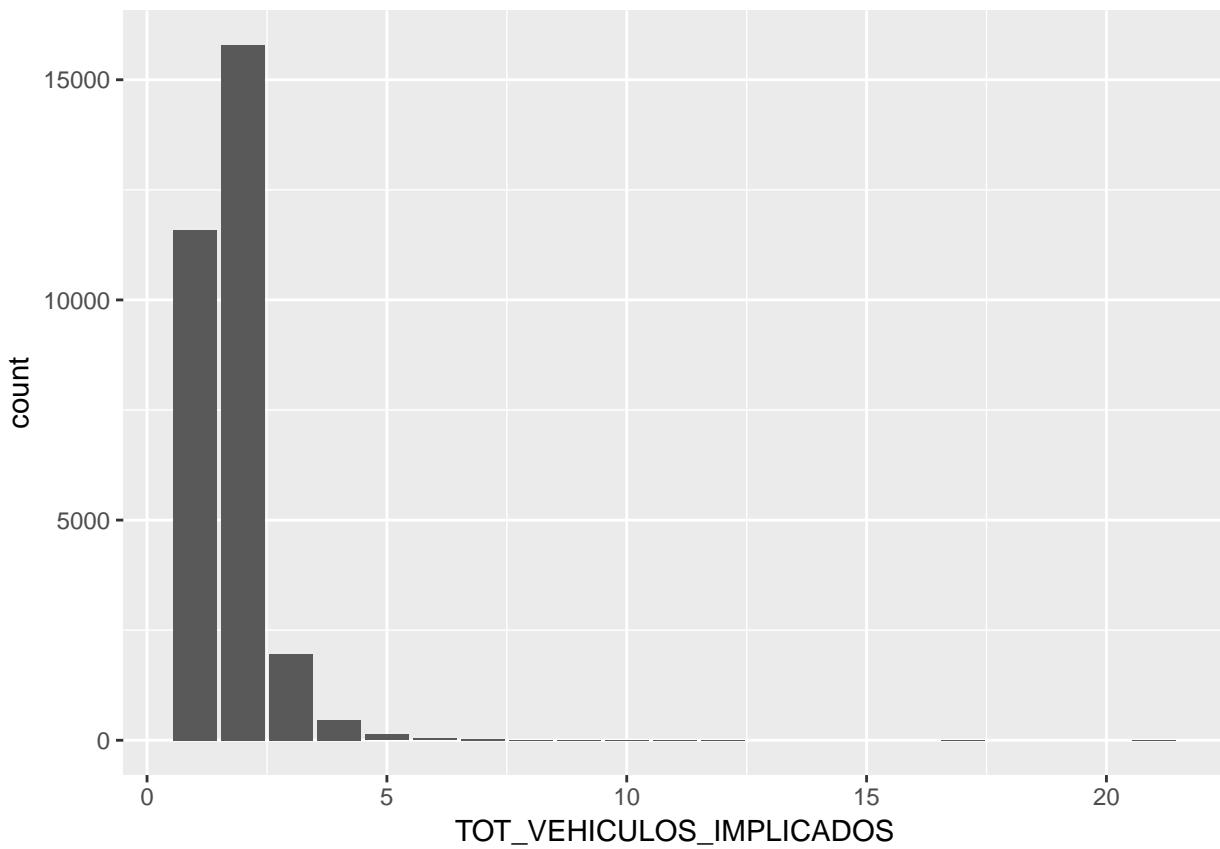


```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_HERIDOS_GRAVES))
```





```
ggplot(data = train.imputados) + geom_bar(mapping = aes(x=TOT_VEHICULOS_IMPLICADOS))
```



Viendo que en cada variable tenemos distintos valores anómalos como sería el valor 19 en TOT_VICTIMAS.

6.2 Paquete mvoutlier

Voy a intentar usar el paquete mvoutlier.

```
require(mvoutlier)

## Loading required package: mvoutlier
## Warning: package 'mvoutlier' was built under R version 3.3.2
## Loading required package: sgeostat
## sROC 0.1-2 loaded
#resultado.busqueda.anomalias <- uni.plot(train.imputados[1:200,1:2])
```

Como se puede ver, se ha obtenido un error el cual no he podido solucionar.

6.3 Eliminación de valores anómalos

En función de lo obtenido con el paquete outlier, voy a intentar realizar algo con este paquete para ver que tal se comporta nuestro dataset.

```
valores.anomalous.train <- outlier(train.imputados[,1:5])
valores.anomalous.test <- outlier(test.imputados[,1:5])
print(valores.anomalous.train)
```

```

##          TOT_VICTIMAS           TOT_MUERTOS           TOT_HERIDOS_GRAVES
##                19                      7                         9
##      TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                18                      21
print(valores.anomalous.test)

##          TOT_VICTIMAS           TOT_MUERTOS           TOT_HERIDOS_GRAVES
##                10                      5                         5
##      TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                10                      11

Veamos, por ejemplo, para la variable TOT_VICTIMAS, cuantas instancias cumplen tener mas de 19 victimas o de 10.

vector.con.victimas.19 <- train.imputados$TOT_VICTIMAS >= 19
sum(vector.con.victimas.19)

## [1] 1

vector.con.victimas.18 <- train.imputados$TOT_VICTIMAS >= 18
sum(vector.con.victimas.18)

## [1] 2

vector.con.victimas.17 <- train.imputados$TOT_VICTIMAS >= 17
sum(vector.con.victimas.17)

## [1] 3

vector.con.victimas.10 <- train.imputados$TOT_VICTIMAS >= 10
sum(vector.con.victimas.10)

## [1] 18

valores.con.victimas.10 <- train.imputados[vector.con.victimas.10,]
valores.con.victimas.10$TIPO_ACCIDENTE

## [1] Salida_Via          Colision_Vehiculos Colision_Vehiculos
## [4] Colision_Vehiculos Colision_Vehiculos Colision_Vehiculos
## [7] Salida_Via          Colision_Vehiculos Salida_Via
## [10] Colision_Vehiculos Colision_Vehiculos Colision_Vehiculos
## [13] Salida_Via          Otro          Colision_Vehiculos
## [16] Colision_Vehiculos Colision_Vehiculos Colision_Vehiculos
## attr(,"contrasts")
##                  2 3 4 5 6
## Atropello          0 0 0 0 0
## Colision_Obstaculo 1 0 0 0 0
## Colision_Vehiculos 0 1 0 0 0
## Otro              0 0 1 0 0
## Salida_Via         0 0 0 1 0
## Vuelco             0 0 0 0 1
## 6 Levels: Atropello Colision_Obstaculo Colision_Vehiculos ... Vuelco

```

Vemos que no son demasiados datos, ya que en total son 18 instancias, por lo que vamos a probar a eliminarlas a ver el comportamiento del paquete outlier de nuevo.

```

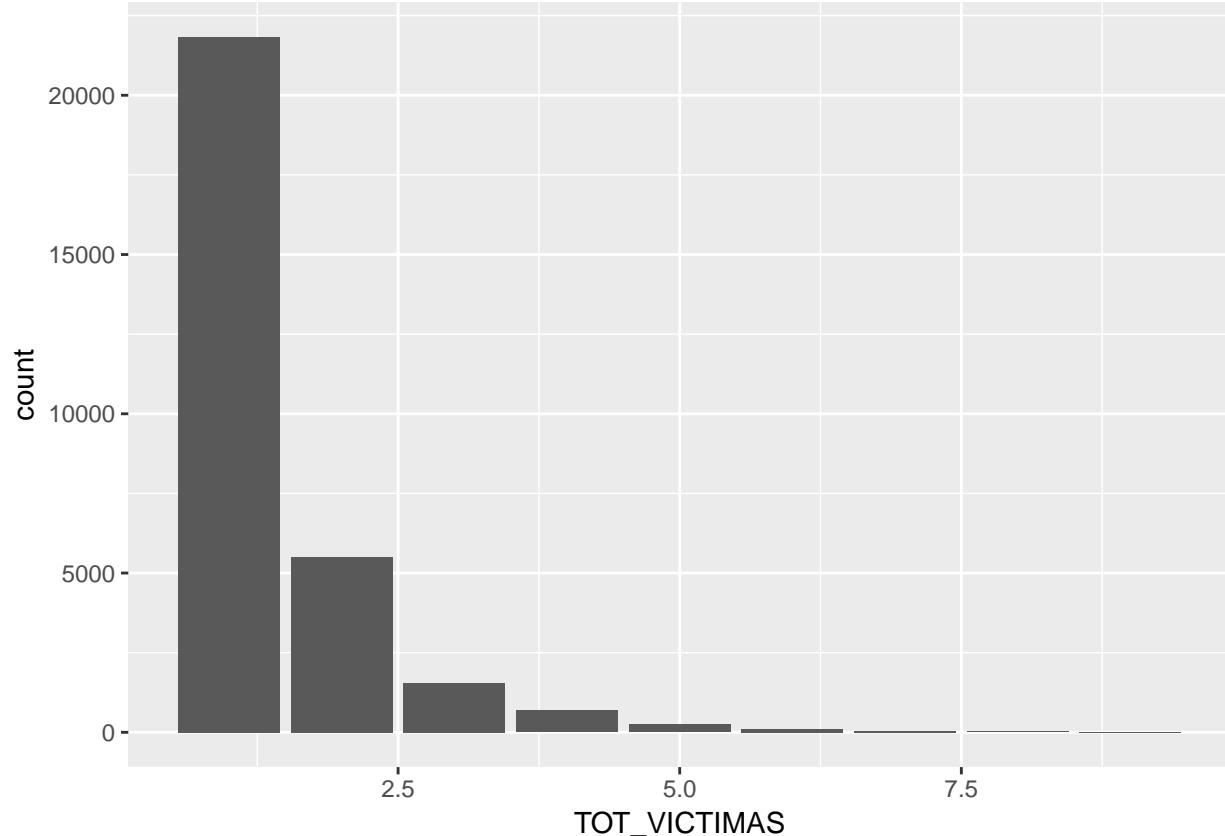
train.sin.outliers <- train.imputados[!vector.con.victimas.10,]
valores.anomalous.sin.victimas.10 <- outlier(train.sin.outliers[,1:5])
print(valores.anomalous.sin.victimas.10)

```

```

##          TOT_VICTIMAS           TOT_MUERTOS           TOT_HERIDOS_GRAVES
##                      9                         7                           6
##          TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                      9                        17
ggplot(data = train.sin.outliers) + geom_bar(mapping = aes(x=TOT_VICTIMAS))

```



Vamos a probar a eliminar algunas instancias, con los criterios de otras variables.

```

vector.con.muertos.7 <- train.sin.outliers$TOT_MUERTOS >=7
sum(vector.con.muertos.7)

## [1] 1

vector.con.muertos.6 <- train.sin.outliers$TOT_MUERTOS >= 6
sum(vector.con.muertos.6)

## [1] 1

vector.con.muertos.5 <- train.sin.outliers$TOT_MUERTOS >= 5
sum(vector.con.muertos.5)

## [1] 1

vector.con.muertos.4 <- train.sin.outliers$TOT_MUERTOS >= 4
sum(vector.con.muertos.4)

## [1] 6

train.sin.outliers <- train.sin.outliers[!vector.con.muertos.4,]
valores.anomalous.sin.muertos.4 <- outlier(train.sin.outliers[,1:5])
print(valores.anomalous.sin.muertos.4)

```

```

##          TOT_VICTIMAS      TOT_MUERTOS      TOT_HERIDOS_GRAVES
##                9                  3                      6
##      TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                9                  17

```

Vamos a realizarlo más rápidamente

```

vector.con.anomalias <- ((train.sin.outliers$TOT_HERIDOS_GRAVES >= 6) | (train.sin.outliers$TOT_HERIDOS_GRAVES >= 5) | (train.sin.outliers$TOT_HERIDOS_GRAVES >= 4))

## [1] 10

vector.con.anomalias <- ((train.sin.outliers$TOT_HERIDOS_GRAVES >= 5) | (train.sin.outliers$TOT_HERIDOS_GRAVES >= 4))

## [1] 31

train.sin.outliers <- train.sin.outliers[!vector.con.anomalias,]

```

Pero, que pasaría si elimináramos en función de las anomalías que nos marca el test:

```

print(valores.anomalous.test)

##          TOT_VICTIMAS      TOT_MUERTOS      TOT_HERIDOS_GRAVES
##                10                 5                      5
##      TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                10                 11

vector.con.anomalias <- ((train.imputados$TOT_HERIDOS_GRAVES > 5) | (train.imputados$TOT_HERIDOS_LEVES > 4))

## [1] 14

```

En total eliminaríamos 14 instancias. Vamos a comprobarlo:

```

train.sin.outliers <- train.imputados[!vector.con.anomalias,]
valores.anomalous.train <- outlier(train.sin.outliers[,1:5])
print(valores.anomalous.train)

##          TOT_VICTIMAS      TOT_MUERTOS      TOT_HERIDOS_GRAVES
##                10                 4                      5
##      TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                10                 11

print(valores.anomalous.test)

##          TOT_VICTIMAS      TOT_MUERTOS      TOT_HERIDOS_GRAVES
##                10                 5                      5
##      TOT_HERIDOS_LEVES  TOT_VEHICULOS_IMPLICADOS
##                10                 11

```

6.4 Prueba del modelo con imputación de valores perdidos

Hagamos por lo tanto una prueba de como afecta la imputación de valores perdidos.

```

set.seed(1234)
ct4 <- ctree(TIPO_ACCIDENTE ~ ., train.sin.outliers)
testPred4 <- predict(ct4, newdata = test.imputados)

```

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

```
#ct4
```

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

```
salida.modelo.4 <- as.matrix(testPred4)
salida.modelo.4 <- cbind(c(1:(dim(salida.modelo.4)[1])), salida.modelo.4)
colnames(salida.modelo.4) <- c("Id", "Prediction")
write.table(salida.modelo.4, file="predicciones/Prediccion4.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 19/02/2017 a las 20:12, con un total de 16 personas entregadas, se ha quedado en la posición 10 con una puntuación del 0.81753. Bajando muy poco con respecto a la anterior puntuación.

#	Δ5d	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	new	Anabel Gómez	0.83175	12	Sun, 19 Feb 2017 13:06:40 (-2.9d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	new	Jonathan Espinosa	0.82780	13	Sun, 19 Feb 2017 18:42:10 (-7h)
5	↓3	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
6	↓2	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
7	↓4	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
8	new	WhiteShadow	0.82345	6	Sat, 18 Feb 2017 14:23:36 (-17.9h)
9	new	Jorge Jimena	0.82059	4	Sun, 19 Feb 2017 16:12:15 (-0.2h)
10	↓4	PacoPollos	0.81891	4	Sun, 19 Feb 2017 19:11:28 (-2.1d)
Your Best Entry ↑					
Your submission scored 0.81753 , which is not an improvement of your best score. Keep trying!					
11	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
12	↓7	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
13	↓6	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
14	↓6	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
15	new	Xisco Fauli	0.48735	2	Wed, 15 Feb 2017 23:16:45
16	↓7	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 4: Cuarta puntuación obtenida en Kaggel

7 Transformación de los datos

Tal y como se vio en el guión de prácticas en el punto 7, vamos a aplicar la transformación para ver que tal nos funciona.

7.1 Transformando los datos

Vamos a aplicar centrado y escalado sobre el conjunto de datos con los valores ya imputados, para las variables que se consideran continuas.

```
valores.preprocesados <- caret::preProcess(train.sin.outliers[,1:5],method=c("center","scale"))
valores.transofrmados <- predict(valores.preprocesados,train.sin.outliers[,1:5])
train.transformado <- cbind(valores.transofrmados,train.sin.outliers[,6:11])
valores.preprocesados.test <- caret::preProcess(test.imputados[,1:5],method=c("center","scale"))
valores.transofrmados.test <- predict(valores.preprocesados.test,test.imputados[,1:5])
test.transformado <- cbind(valores.transofrmados.test,test.imputados[,6:10])
```

7.2 Prueba del modelo con transformación de los datos

Hagamos por lo tanto una prueba de como afecta la transformación de los datos.

```
set.seed(1234)
ct5 <- ctree(TIPO_ACCIDENTE ~., train.transformado)
testPred5 <- predict(ct5, newdata = test.transformado)
```

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

```
#ct5
```

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

```
salida.modelo.5 <- as.matrix(testPred5)
salida.modelo.5 <- cbind(c(1:(dim(salida.modelo.5)[1])), salida.modelo.5)
colnames(salida.modelo.5) <- c("Id", "Prediction")
write.table(salida.modelo.5,file="predicciones/Prediccion5.txt",sep=",",quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 20/02/2017 a las 13:15, con un total de 18 personas entregadas, se ha quedado en la posición 12 con una puntuación del 0.55147. Bajando mucho con respecto a la anterior puntuación, por lo que esta transformación no la tendremos en cuenta.

8 Discretización

Para este conjunto de datos no se realiza discretización ya que no tenemos variables continuas como para poder discretizarlas.

9 Selección de características

Para este apartado comenzaremos con los dataset originales.

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
```

#	$\Delta 6d$	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	new	Anabel Gómez	0.83175	15	Mon, 20 Feb 2017 07:44:42 (-3.7d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	new	Jonathan Espinosa	0.82859	14	Sun, 19 Feb 2017 19:39:14
5	new	Xisco Fauli	0.82810	10	Mon, 20 Feb 2017 10:50:53 (-1.3h)
6	new	ManuelMontero	0.82582	3	Sun, 19 Feb 2017 20:10:00
7	↓5	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
8	↓5	RubenSanchez	0.82533	9	Fri, 17 Feb 2017 16:19:18 (-2.7d)
9	new	RonCR	0.82365	2	Tue, 14 Feb 2017 16:24:28
10	new	WhiteShadow	0.82345	6	Sat, 18 Feb 2017 14:23:36 (-17.9h)
11	new	Jorge Jimena	0.82306	6	Sun, 19 Feb 2017 20:48:53
12	↓7	PacoPollos	0.81891	5	Mon, 20 Feb 2017 12:15:12 (-2.8d)
Your Best Entry ↑					
Your submission scored 0.55147 , which is not an improvement of your best score. Keep trying!					
13	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
14	new	Salva Moreno	0.81891	2	Mon, 20 Feb 2017 12:00:33
15	↓11	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
16	↓10	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
17	↓10	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
18	↓10	LauraDelPinoDíaz	0.12290	1	Mon, 13 Feb 2017 22:51:17

Figure 5: Quinta puntuación obtenida en Kaggel

9.1 Paquete FSelector

9.1.1 Aproximación filter: chi.squared

Determina los pesos de los atributos discretos usando el test de independencia chi-cuadrado (con respecto a la variable clase). Calculamos los pesos de los atributos: la medida devuelta indica el nivel de dependencia de cada atributo frente a la variable clase

```
set.seed(1234)
pesos <- FSelector::chi.squared(TIPO_ACCIDENTE~, train.original)
pesos
```

	attr_importance
##	0.0000000
## ANIO	0.03094819
## MES	0.15532002
## HORA	0.06981337
## DIASEMANA	0.14327070
## PROVINCIA	0.12198994
## COMUNIDAD_AUTONOMA	0.02436129
## ISLA	0.09669636
## TOT_VICTIMAS	0.06428765
## TOT_MUERTOS	0.08816985
## TOT_HERIDOS_GRAVES	0.13237988
## TOT_HERIDOS_LEVES	0.63503097
## ZONA	0.26553819
## ZONA_AGRUPADA	0.45894923
## CARRETERA	0.52879460
## RED_CARRETERA	0.19748117
## TIPO_VIA	0.18875170
## TRAZADO_NO_INTERSEC	0.19181245
## TIPO_INTERSEC	0.14228134
## ACOND_CALZADA	0.09438668
## PRIORIDAD	0.22060851
## SUPERFICIE_CALZADA	0.11239155
## LUMINOSIDAD	0.12226652
## FACTORES_ATMOSFERICOS	0.08055707
## VISIBILIDAD_RESTRINGIDA	0.09419773
## OTRA_CIRCUNSTANCIA	0.06053977
## ACERAS	0.22765102
## DENSIDAD_CIRCULACION	0.12681797
## MEDIDAS_ESPECIALES	0.04702684

Vamos a seleccionar los 7 mejores

```
subset <- FSelector::cutoff.k(pesos, 7)
las.7.mas.importantes.chi.squared <- as.simple.formula(subset, "TIPO_ACCIDENTE")
las.7.mas.importantes.chi.squared
```

```
## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + CARRETERA + ZONA_AGRUPADA +
##      ZONA + ACERAS + PRIORIDAD + RED_CARRETERA
## <environment: 0x7ff9073e8ea0>
```

Por lo que vamos a montar un modelo con estas variables

```
train.filter.chi.squared <- train.original[, c("TOT_VEHICULOS_IMPLICADOS", "CARRETERA", "ZONA_AGRUPADA", "ZON
test.filter.chi.squared <- test.original[, c("TOT_VEHICULOS_IMPLICADOS", "CARRETERA", "ZONA_AGRUPADA", "ZON
```

```
summary(train.filter.chi.squared)
```

```
##   TOT_VEHICULOS_IMPLICADOS    CARRETERA          ZONA_AGRUPADA
##   Min. : 1.000            A-7 : 294    VIAS INTERURBANAS:13335
##   1st Qu.: 1.000            A-2 : 278    VIAS URBANAS      :16667
##   Median : 2.000            AP-7 : 229
##   Mean   : 1.738            N-340 : 229
##   3rd Qu.: 2.000            A-4  : 184
##   Max.   :21.000            (Other):12098
##   NA's   :16690
##
##           ZONA          ACERAS          PRIORIDAD
##   CARRETERA :13278  NO HAY ACERA:21416  NINGUNA (SOLO NORMA) :13495
##   TRAVESIA  : 241   SI HAY ACERA: 5437  SEMAFORO             : 1778
##   VARIANTE  :  57   NA's              : 3149  SEÑAL DE STOP       : 1750
##   ZONA URBANA:16426
##
##           RED_CARRETERA
##   OTRAS TITULARIDADES      :  318
##   TITULARIDAD AUTONOMICA   : 3890
##   TITULARIDAD ESTATAL      : 4021
##   TITULARIDAD MUNICIPAL     :19077
##   TITULARIDAD PROVINCIAL (DIPUTACION, CABILDO O CONSELL): 2696
##
##           TIPO_ACCIDENTE
##   Atropello        : 3642
##   Colision_Obstaculo:  952
##   Colision_Vehiculos:16520
##   Otro             : 1807
##   Salida_Via       : 6013
##   Vuelco           : 1068
##
```

Vemos que la variable CARRETERA tiene un alto número de valores perdidos por lo que la vamos a descartar, a pesar de que la selección de características nos ha dicho que es importante.

```
train.filter.chi.squared[ "CARRETERA" ] <- NULL
test.filter.chi.squared[ "CARRETERA" ] <- NULL
```

9.1.2 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct6 <- ctree(TIPO_ACCIDENTE ~., train.filter.chi.squared)
testPred6 <- predict(ct6, newdata = test.filter.chi.squared)
```

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

```
#ct6
```

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

```

salida.modelo.6 <- as.matrix(testPred6)
salida.modelo.6 <- cbind(c(1:(dim(salida.modelo.6)[1])), salida.modelo.6)
colnames(salida.modelo.6) <- c("Id", "Prediction")
write.table(salida.modelo.6,file="predicciones/Prediccion6.txt",sep=",",quote = F,row.names = F)

```

El resultado de este modelo para la competición de Kaggle, subido el 22/02/2017 a las 13:20, con un total de 21 personas entregadas, se ha quedado en la posición 13 con una puntuación del 0.82089. Mejorando a la que ya se tenía anteriormente, por lo que vemos que esta selección de características ha funcionado correctamente.

9.1.3 Aproximación filter: correlation

Busca los pesos de atributos continuos en base a medidas de correlación. Por lo tanto esta aproximación no podremos realizarla al tener la variable clase no numérica.

9.1.4 Aproximación filter: entropy_based

Encontraremos los pesos de los atributos discretos en base a su correlación con el atributo clase.

```

set.seed(1234)
pesos <- FSelector::information.gain(TIPO_ACCIDENTE~, train.original)
subset <- cutoff.k(pesos,7)
los.7.mas.importantes.information.gain <- as.simple.formula(subset, "TIPO_ACCIDENTE")
los.7.mas.importantes.information.gain

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + CARRETERA + ZONA +
##      ZONA_AGRUPADA + TIPO_VIA + TRAZADO_NO_INTERSEC + PRIORIDAD
## <environment: 0x7ff906fab030>

set.seed(1234)
pesos <- FSelector::gain.ratio(TIPO_ACCIDENTE~, train.original)
subset <- cutoff.k(pesos,7)
los.7.mas.importantes.gain.ratio <- as.simple.formula(subset, "TIPO_ACCIDENTE")
los.7.mas.importantes.gain.ratio

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + ZONA_AGRUPADA + ZONA +
##      TIPO_VIA + CARRETERA + RED_CARRETERA + TRAZADO_NO_INTERSEC
## <environment: 0x7ff9090367c8>

set.seed(1234)
pesos <- FSelector::symmetrical.uncertainty(TIPO_ACCIDENTE~, train.original)
subset <- cutoff.k(pesos,7)
los.7.mas.importantes.symmetrical.uncertainty <- as.simple.formula(subset, "TIPO_ACCIDENTE")
los.7.mas.importantes.symmetrical.uncertainty

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + CARRETERA + ZONA_AGRUPADA +
##      ZONA + TIPO_VIA + TRAZADO_NO_INTERSEC + RED_CARRETERA
## <environment: 0x7ff905923350>

Por lo que en función de estas tres salidas tenemos que las variables más importantes serían: TOT_VEHICULOS_IMPLICADOS, CARRETERA, ZONA, ZONA_AGRUPADA, TIPO_VIA, TRAZADO_NO_INTERSEC, PRIORIDAD y RED_CARRETERA. El único que difiere entre algoritmos es PRIORIDAD y RED_CARRETERA. Recordemos los que teníamos con chi_cuadrado:
las.7.mas.importantes.chi.squared

```

#	Δ6d	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	—	RubenSanchez	0.82889	12	Tue, 21 Feb 2017 22:52:29
5	↑1	Jonathan Espinosa	0.82859	14	Sun, 19 Feb 2017 19:39:14
6	↑6	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
7	new	Jorge Jimena	0.82662	13	Wed, 22 Feb 2017 11:58:13 (-33.8h)
8	↓1	WhiteShadow	0.82632	9	Wed, 22 Feb 2017 12:02:54 (-20.5h)
9	new	ManuelMontero	0.82582	11	Wed, 22 Feb 2017 11:56:12 (-2.7d)
10	↓5	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
11	new	Salva Moreno	0.82573	10	Wed, 22 Feb 2017 01:48:56 (-25.9h)
12	↓9	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
13	↓4	PacoPollos	0.82089	6	Wed, 22 Feb 2017 12:20:01
Your Best Entry ↑					
You improved on your best score by 0.00198.					
You just moved up 1 position on the leaderboard.  Tweet this!					
14	new	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
15	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
16	↓8	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
17	new	Mauricio Orellana	0.73246	1	Wed, 22 Feb 2017 02:42:56
18	↓8	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
19	↓8	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10
20	new	StephanieMoraAndrade	0.41514	5	Tue, 21 Feb 2017 18:54:45 (-0.4h)

Figure 6: Sexta puntuación obtenida en Kaggle

```
## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + CARRETERA + ZONA_AGRUPADA +
##      ZONA + ACERAS + PRIORIDAD + RED_CARRETERA
## <environment: 0x7ff9073e8ea0>
```

Es decir, tenemos diferencias en ACERAS. Vamos a probar un modelo con todas las variables dadas por este nuevo método.

```
train.filter.entropy.bases <- train.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","ZONA","TIPO_VI")]
test.filter.entropy.bases <- test.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","ZONA","TIPO_VI")]
```

Como anteriormente, hemos quitado la variable CARRETERA

9.1.5 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct7 <- ctree(TIPO_ACCIDENTE ~., train.filter.entropy.bases)
testPred7 <- predict(ct7, newdata = test.filter.entropy.bases)
```

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

```
#ct7
```

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

```
salida.modelo.7 <- as.matrix(testPred7)
salida.modelo.7 <- cbind(c(1:(dim(salida.modelo.7)[1])), salida.modelo.7)
colnames(salida.modelo.7) <- c("Id", "Prediction")
write.table(salida.modelo.7,file="predicciones/Prediccion7.txt",sep=",",quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 22/02/2017 a las 14:00, con un total de 22 personas entregadas, se ha quedado en la posición 14 con una puntuación del 0.82227. Mejorando a la que ya se tenía anteriormente, por lo que vemos que esta selección de características ha funcionado mejor.

9.1.6 Aproximación filter: oneR

Método simple de cálculo de pesos para atributos discretos mediante el uso de reglas de asociación con un sólo término en el antecedente.

```
pesos <- FSelector::oneR(TIPO_ACCIDENTE~., train.original)
subset <- cutoff.k(pesos,7)
los.7.mas.importantes.oneR <- as.simple.formula(subset,"TIPO_ACCIDENTE")
los.7.mas.importantes.oneR
```

```
## TIPO_ACCIDENTE ~ ANIO + TOT_VEHICULOS_IMPLICADOS + ACERAS + DENSIDAD_CIRCULACION +
##      ZONA_AGRUPADA + TOT_MUERTOS + TOT_VICTIMAS
## <environment: 0x7ff8f2073920>
```

Podemos ver que tenemos distintos atributos más importantes según este método, por lo que vamos aprobar que tal se comportan estos atributos.

```
train.filter.oneR <- train.original[,c("ANIO","TOT_VEHICULOS_IMPLICADOS","ACERAS","DENSIDAD_CIRCULACION")]
test.filter.oneR <- test.original[,c("ANIO","TOT_VEHICULOS_IMPLICADOS","ACERAS","DENSIDAD_CIRCULACION")]
```

#	$\Delta 6d$	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓2	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	↑1	RubenSanchez	0.82889	12	Tue, 21 Feb 2017 22:52:29
5	↓2	Jonathan Espinosa	0.82859	14	Sun, 19 Feb 2017 19:39:14
6	↑6	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
7	new	David Urdin	0.82691	2	Wed, 22 Feb 2017 12:46:17 (-0h)
8	new	Jorge Jimena	0.82662	14	Wed, 22 Feb 2017 12:54:11 (-34.7h)
9	↓2	WhiteShadow	0.82632	10	Wed, 22 Feb 2017 12:48:25 (-21.2h)
10	new	ManuelMontero	0.82582	12	Wed, 22 Feb 2017 12:26:04 (-2.7d)
11	↓5	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
12	new	Salva Moreno	0.82573	10	Wed, 22 Feb 2017 01:48:56 (-25.9h)
13	↓9	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
14	↓5	PacoPollos	0.82227	7	Wed, 22 Feb 2017 13:00:58
Your Best Entry ↑					
You improved on your best score by 0.00138.					
15	new	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
16	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
17	↓9	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
18	new	Mauricio Orellana	0.73246	1	Wed, 22 Feb 2017 02:42:56
19	↓9	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
20	↓9	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10

Figure 7: Puntuación obtenida en Kaggle

9.1.7 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct8 <- ctree(TIPO_ACCIDENTE ~ ., train.filter.oneR)
testPred8 <- predict(ct8, newdata = test.filter.oneR)
```

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

```
#ct8
```

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

```
salida.modelo.8 <- as.matrix(testPred8)
salida.modelo.8 <- cbind(c(1:(dim(salida.modelo.8)[1])), salida.modelo.8)
colnames(salida.modelo.8) <- c("Id", "Prediction")
write.table(salida.modelo.8, file="prediccciones/Prediccion8.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 23/02/2017 a las 12:51, con un total de 22 personas entregadas, se ha quedado en la posición 14 con una puntuación del 0.81891. La cual no mejora a la mejor que ya teníamos.

9.1.8 Aproximación filter: relief

Algoritmo de búsqueda de pesos de atributos continuos y discretos en base a la distancia entre instancias.

```
pesos <- relief(TIPO_ACCIDENTE~., train.original, neighbours.count = 5, sample.size = 20)
pesos
```

	attr_importance
##	
## ANIO	0.075018769
## MES	0.112034688
## HORA	0.103111443
## DIASEMANA	-0.047879632
## PROVINCIA	0.117306135
## COMUNIDAD_AUTONOMA	0.119171925
## ISLA	0.000000000
## TOT_VICTIMAS	-0.003616254
## TOT_MUERTOS	-0.005434975
## TOT_HERIDOS_GRAVES	0.006929172
## TOT_HERIDOS_LEVES	0.001537272
## TOT_VEHICULOS_IMPLICADOS	0.029436848
## ZONA	0.003799304
## ZONA_AGRUPADA	0.001518196
## CARRETERA	0.030862791
## RED_CARRETERA	0.002188091
## TIPO_VIA	0.008518449
## TRAZADO_NO_INTERSEC	0.042740756
## TIPO_INTERSEC	0.034668044
## ACOND_CALZADA	0.041610658
## PRIORIDAD	0.100589138
## SUPERFICIE_CALZADA	0.005795172
## LUMINOSIDAD	0.056955150
## FACTORES_ATMOSFERICOS	0.004863756
## VISIBILIDAD_RESTRINGIDA	0.038455537

#	Δ6d	Team Name	Score ⓘ	Entries	Last Submission UTC (Best - Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	new	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
3	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
4	↑1	RubenSanchez	0.82889	13	Wed, 22 Feb 2017 16:22:32 (-17.5h)
5	↓2	Jonathan Espinosa	0.82859	14	Sun, 19 Feb 2017 19:39:14
6	↑7	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
7	—	WhiteShadow	0.82780	15	Thu, 23 Feb 2017 10:50:04 (-0.4h)
8	new	David Urdin	0.82691	2	Wed, 22 Feb 2017 12:46:17 (-0h)
9	↑1	Jorge Jimena	0.82662	17	Thu, 23 Feb 2017 00:09:49 (-45.9h)
10	new	ManuelMontero	0.82582	12	Wed, 22 Feb 2017 12:26:04 (-2.7d)
11	↓5	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
12	new	Salva Moreno	0.82573	10	Wed, 22 Feb 2017 01:48:56 (-25.9h)
13	↓9	BesayMontesdeocaSantana	0.82543	7	Mon, 13 Feb 2017 20:09:42
14	↓5	Francisco Pérez Hernández	0.82227	8	Thu, 23 Feb 2017 11:51:40 (-22.8h)
Your Best Entry ↑					
Your submission scored 0.81891 , which is not an improvement of your best score. Keep trying!					
15	new	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
16	new	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
17	↓9	fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
18	new	Mauricio Orellana	0.73246	1	Wed, 22 Feb 2017 02:42:56
19	↓8	Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
20	↓8	Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10

Figure 8: Puntuación obtenida en Kaggle

```

## OTRA_CIRCUNSTANCIA      -0.002784773
## ACERAS                  -0.009134334
## DENSIDAD_CIRCULACION   -0.008797839
## MEDIDAS_ESPECIALES      0.009580661

subset <- cutoff.k(pesos,7)
los.7.mas.importantes.relief <- as.simple.formula(subset,"TIPO_ACCIDENTE")
los.7.mas.importantes.relief

## TIPO_ACCIDENTE ~ COMUNIDAD_AUTONOMA + PROVINCIA + MES + HORA +
##      PRIORIDAD + ANIO + LUMINOSIDAD
## <environment: 0x7fc51ce993f8>

```

Vemos que los más importantes, según este método son: COMUNIDAD_AUTONOMA, PROVINCIA, MES, HORA, PRIORIDAD, ANIO y LUMINOSIDAD.

Vamos a probar que tal se comportan estos atributos. (HORA la elimino al tener un gran número de factores y ralentizar los cálculos)

```

train.filter.relief <- train.original[,c("COMUNIDAD_AUTONOMA","PROVINCIA","MES","PRIORIDAD","ANIO","LUMINOSIDAD")]
test.filter.relief <- test.original[,c("COMUNIDAD_AUTONOMA","PROVINCIA","MES","PRIORIDAD","ANIO","LUMINOSIDAD")]

```

9.1.9 Prueba del modelo

Hagamos por lo tanto una prueba.

```

set.seed(1234)
ct9 <- ctree(TIPO_ACCIDENTE ~ ., train.filter.relief)
testPred9 <- predict(ct9, newdata = test.filter.relief)

```

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

```
#ct9
```

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

```

salida.modelo.9 <- as.matrix(testPred9)
salida.modelo.9 <- cbind(c(1:(dim(salida.modelo.9)[1])), salida.modelo.9)
colnames(salida.modelo.9) <- c("Id","Prediction")
write.table(salida.modelo.9,file="predicciones/Prediccion9.txt",sep=",",quote = F,row.names = F)

```

El resultado de este modelo para la competición de Kaggle, subido el 28/02/2017 a las 11:51, con un total de 26 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.59119. La cual empeora mucho a lo que ya se tenía, por lo que no es una buena selección de características.

9.1.10 Aproximación wrapper: cfs

Decir que todos los métodos propuestos de wrapper como: best.first.search, exhaustive.search, greedy.search y hill.climbing.search, no me han funcionado debido al tipo de datos que tenemos. Por contra, el método cfs si ha funcionado correctamente.

```

set.seed(1234)
subset <- FSelector::cfs(TIPO_ACCIDENTE~.,train.original)
el.mejor.segun.cfs <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.cfs

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS
## <environment: 0x7fc51d502508>

```

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑10	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
3	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
4	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
5	↑5	RubenSanchez	0.82899	14	Sat, 25 Feb 2017 09:42:44
6	↓2	Jonathan Espinosa	0.82859	27	Sun, 26 Feb 2017 17:13:00 (-6.9d)
7	↓2	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
8	new	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	↓3	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	↓3	ManuelMontero	0.82582	12	Wed, 22 Feb 2017 12:26:04 (-2.7d)
11	—	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
12	↓4	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
13	new	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
14	↓5	BesayMontesdeocaSantana	0.82543	11	Thu, 23 Feb 2017 14:48:58 (-9.8d)
15	new	Alberto Baena Priego	0.82326	6	Fri, 24 Feb 2017 12:32:32 (-3.1h)
16	↓3	Francisco Pérez Hernández	0.82227	9	Tue, 28 Feb 2017 10:51:45 (-5.9d)
Your Best Entry ↑					
Your submission scored 0.59119 , which is not an improvement of your best score. Keep trying!					
17	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
18	new	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
19	↓5	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
20	new	Alice Gomez	0.81891	2	Thu, 23 Feb 2017 20:26:21 (-0h)

Figure 9: Puntuación obtenida en Kaggle

Siendo TOT_VEHICULOS_IMPLICADOS, la mejor característica, repitamos esto para obtener las 5 mejores, de forma que eliminamos la que mejor se obtiene.

```
set.seed(1234)
train.wrapper.cfs = train.original[,-12]
subset <- FSelector::cfs(TIPO_ACCIDENTE~.,train.wrapper.cfs)
el.mejor.segun.cfs <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.cfs
```

```
## TIPO_ACCIDENTE ~ ZONA_AGRUPADA + CARRETERA + TRAZADO_NO_INTERSEC
## <environment: 0x7fc5008600b0>
```

Ahora obtenemos ZONA_AGRUPADA, CARRETERA y TRAZADO_NO_INTERSEC.

```
set.seed(1234)
train.wrapper.cfs = train.original[,-c(12,14,15,18)]
subset <- FSelector::cfs(TIPO_ACCIDENTE~.,train.wrapper.cfs)
el.mejor.segun.cfs <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.cfs
```

```
## TIPO_ACCIDENTE ~ TOT_HERIDOS_LEVES + ZONA + RED_CARRETERA + TIPO_VIA +
##      TIPO_INTERSEC + PRIORIDAD + SUPERFICIE_CALZADA + ACERAS
## <environment: 0x7fc500540358>
```

Obteniendo: TOT_HERIDOS_LEVES, ZONA, RED_CARRETERA, TIPO_VIA, TIPO_INTERSEC, PRIORIDAD, SUPERFICIE_CALZADA y ACERAS.

Por lo tanto, ya que tenemos muchas características, vamos a probar el modelo con todas, salvo CARRETERA.

```
train.wrapper.cfs <- train.original[,c("TOT_VEHICULOS_IMPLICADOS", "ZONA_AGRUPADA", "TRAZADO_NO_INTERSEC")]
test.wrapper.cfs <- test.original[,c("TOT_VEHICULOS_IMPLICADOS", "ZONA_AGRUPADA", "TRAZADO_NO_INTERSEC", "TIPO_VIA", "TIPO_INTERSEC", "PRIORIDAD", "SUPERFICIE_CALZADA", "ACERAS")]
```

9.1.11 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct10 <- ctree(TIPO_ACCIDENTE ~., train.wrapper.cfs)
testPred10 <- predict(ct10, newdata = test.wrapper.cfs)
```

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

```
#ct10
```

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

```
salida.modelo.10 <- as.matrix(testPred10)
salida.modelo.10 <- cbind(c(1:(dim(salida.modelo.10)[1])), salida.modelo.10)
colnames(salida.modelo.10) <- c("Id", "Prediction")
write.table(salida.modelo.10, file="predicciones/Prediccion10.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 28/02/2017 a las 12:17, con un total de 26 personas entregadas, se ha quedado en la posición 15 con una puntuación del 0.82395. Por lo que se ha mejorado a la mejor obtenida por mi hasta el momento.

9.1.12 Aproximación wrapper: consistency

Probamos este método:

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑10	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
3	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
4	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
5	↑5	RubenSanchez	0.82899	14	Sat, 25 Feb 2017 09:42:44
6	↓2	Jonathan Espinosa	0.82859	27	Sun, 26 Feb 2017 17:13:00 (-6.9d)
7	↓2	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
8	new	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	↓3	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	↓3	ManuelMontero	0.82582	12	Wed, 22 Feb 2017 12:26:04 (-2.7d)
11	—	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
12	↓4	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
13	new	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
14	↓5	BesayMontesdeocaSantana	0.82543	11	Thu, 23 Feb 2017 14:48:58 (-9.8d)
15	↓2	Francisco Pérez Hernández	0.82395	10	Tue, 28 Feb 2017 11:17:56
Your Best Entry ↑					
You improved on your best score by 0.00168.					
You just moved up 1 position on the leaderboard.				 Tweet this!	
16	new	Alberto Baena Priego	0.82326	6	Fri, 24 Feb 2017 12:32:32 (-3.1h)
17	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
18	new	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
19	↓5	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52
20	new	Alice Gomez	0.81891	2	Thu, 23 Feb 2017 20:26:21 (-0h)

Figure 10: 10 puntuación obtenida en Kaggle

```

set.seed(1234)
subset <- consistency(TIPO_ACCIDENTE~, train.original)
el.mejor.segun.consistency <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.consistency

## TIPO_ACCIDENTE ~ MES + HORA + DIASEMANA + PROVINCIA + TOT_VICTIMAS +
##      TOT_HERIDOS_LEVES + TOT_VEHICULOS_IMPLICADOS + CARRETERA +
##      TIPO_VIA + TRAZADO_NO_INTERSEC + TIPO_INTERSEC + ACOND_CALZADA +
##      PRIORIDAD + SUPERFICIE_CALZADA + LUMINOSIDAD + FACTORES_ATMOSFERICOS +
##      OTRA_CIRCUNSTANCIA + ACERAS + DENSIDAD_CIRCULACION + MEDIDAS_ESPECIALES
## <environment: 0x7fc51c137380>

```

Obteniendo 20 características como importantes, no apareciendo: ANIO, COMUNIDAD_AUTONOMA, ISLA, TOT_MUERTOS, TOT_HERIDOS_GRAVES, ZONA, ZONA_AGRUPADA, RED_CARRETERA, VISIBILIDAD_RESTRINGIDA. Por lo tanto, ya que tenemos muchas características, vamos a ejecutar nuestro modelo. Quitando además CARRETERA al ralentizar los cálculos.

```

train.wrapper.consistency <- train.original[,c(2,3,4,5,8,11,12,17,18,19,20,21,22,23,24,26,27,28,29,30)]
test.wrapper.consistency <- test.original[,c(2,3,4,5,8,11,12,17,18,19,20,21,22,23,24,26,27,28,29)]

```

9.1.13 Prueba del modelo

Hagamos por lo tanto una prueba. Quitando además CARRETERA para realizar los cálculos más rápidamente. Además, para que nuestro árbol funcione, he eliminado la variable HORA, junto con MEDIDAS_ESPECIAL, al contener esta un factor que en test no aparece.

```

train.wrapper.consistency <- train.wrapper.consistency[,-2]
test.wrapper.consistency <- test.wrapper.consistency[,-2]
train.wrapper.consistency <- train.wrapper.consistency[,-18]
test.wrapper.consistency <- test.wrapper.consistency[,-18]
set.seed(1234)
ct11 <- ctree(TIPO_ACCIDENTE ~., train.wrapper.consistency)
testPred11 <- predict(ct11, newdata = test.wrapper.consistency)

```

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

```
#ct11
```

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

```

salida.modelo.11 <- as.matrix(testPred11)
salida.modelo.11 <- cbind(c(1:(dim(salida.modelo.11)[1])), salida.modelo.11)
colnames(salida.modelo.11) <- c("Id", "Prediction")
write.table(salida.modelo.11,file="predicciones/Prediccion11.txt",sep=",",quote = F, row.names = F)

```

El resultado de este modelo para la competición de Kaggel, subido el 01/03/2017 a las 16:56, con un total de 26 personas entregadas, se ha quedado en la posición 15 con una puntuación del 0.82237. Empeorando un poco a la mejor actual.

9.1.14 Aproximación embedded: random森林.importance

Método de cálculo de pesos de importancia de atributos calculados sobre un modelo construido usando el algoritmo RandomForest. Al no funcionar este método con variables que tengan más de 53 categorías hemos tenido que eliminar algunas variables como: CARRETERA y HORA. El tercer argumento con valor 1 significa la reducción en la media de fiabilidad predictiva.

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑7	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
3	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
4	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
5	↓1	RubenSanchez	0.82899	15	Wed, 01 Mar 2017 14:08:28 (-4.2d)
6	↓1	Jonathan Espinosa	0.82859	27	Sun, 26 Feb 2017 17:13:00 (-6.9d)
7	↓1	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
8	↓1	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	↓1	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	—	ManuelMontero	0.82582	12	Wed, 22 Feb 2017 12:26:04 (-2.7d)
11	—	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
12	—	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
13	↑5	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
14	↓1	BesayMontesdeocaSantana	0.82543	11	Thu, 23 Feb 2017 14:48:58 (-9.8d)
15	↓1	Francisco Pérez Hernández	0.82395	12	Wed, 01 Mar 2017 15:56:15 (-28.6h)
Your Best Entry ↑					
Your submission scored 0.82237 , which is not an improvement of your best score. Keep trying!					
16	new	Alberto Baena Priego	0.82326	7	Wed, 01 Mar 2017 09:42:13 (-5d)
17	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
18	↑2	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30
19	↓4	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
20	↓4	alaineiturria	0.81891	2	Wed, 01 Mar 2017 15:51:29 (-9.9d)

Figure 11: 11 puntuación obtenida en Kaggel

```

train.embedded.random.forest.importance <- train.original[,c(1,2,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19)
set.seed(1234)
pesos1 <- FSelector::random.fores t.importantce(TIPO_ACCIDENTE~,train.embedded.random.fores t.importance,
subset1 <- cutoff.k(pesos1,7)
los.7.mas.importantes.random.fores t.importance1 <- as.simple.formula(subset1, "TIPO_ACCIDENTE")
los.7.mas.importantes.random.fores t.importance1

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + PRIORIDAD + SUPERFICIE_CALZADA +
##          TIPO_INTERSEC + ZONA + ZONA_AGRUPADA + TIPO_VIA
## <environment: 0x7fc51bc90e30>

```

El tercer argumento con valor 2 significa la reducción en la media de impureza de nodos.

```

set.seed(1234)
pesos2 <- FSelector::random.fores t.importantce(TIPO_ACCIDENTE~,train.embedded.random.fores t.importance,
subset2 <- cutoff.k(pesos2,7)
los.7.mas.importantes.random.fores t.importance2 <- as.simple.formula(subset2, "TIPO_ACCIDENTE")
los.7.mas.importantes.random.fores t.importance2

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS + PRIORIDAD + MES +
##          PROVINCIA + DIASEMANA + TIPO_INTERSEC + COMUNIDAD_AUTONOMA
## <environment: 0x7fc51ba51fb8>

```

Saliendo con ambos métodos las mismas variables como las más importantes.

```

train.embedded.random.fores t.importance <- train.original[,c("TOT_VEHICULOS_IMPLICADOS","PRIORIDAD","SU")
test.embedded.random.fores t.importance <- test.original[,c("TOT_VEHICULOS_IMPLICADOS","PRIORIDAD","SU")

```

9.1.15 Prueba del modelo

Hagamos por lo tanto una prueba. Quitando además CARRETERA para realizar los cálculos más rápidamente.

```

set.seed(1234)
ct12 <- ctree(TIPO_ACCIDENTE ~., train.embedded.random.fores t.importance)
testPred12 <- predict(ct12, newdata = test.embedded.random.fores t.importance)

```

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

```
#ct12
```

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

```

salida.modelo.12 <- as.matrix(testPred12)
salida.modelo.12 <- cbind(c(1:(dim(salida.modelo.12)[1])), salida.modelo.12)
colnames(salida.modelo.12) <- c("Id","Prediction")
write.table(salida.modelo.12,file="predicciones/Prediccion12.txt",sep=",",quote = F,row.names = F)

```

El resultado de este modelo para la competición de Kaggel, subido el 01/03/2017 a las 16:32, con un total de 26 personas entregadas, se ha quedado en la posición 15 con una puntuación del 0.82326. Por lo que ha empeorado muy poco a la mejor obtenida por mi.

9.2 Paquete caret

Para este apartado comenzaremos con los dataset originales.

```

rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")

```

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best - Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑7	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
3	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
4	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
5	↓1	RubenSanchez	0.82899	15	Wed, 01 Mar 2017 14:08:28 (-4.2d)
6	↓1	Jonathan Espinosa	0.82859	27	Sun, 26 Feb 2017 17:13:00 (-6.9d)
7	↓1	Xisco Fauli	0.82839	12	Tue, 21 Feb 2017 19:28:50 (-18.8h)
8	↓1	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	↓1	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	—	ManuelMontero	0.82582	12	Wed, 22 Feb 2017 12:26:04 (-2.7d)
11	—	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
12	—	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
13	↑5	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
14	↓1	BesayMontesdeocaSantana	0.82543	11	Thu, 23 Feb 2017 14:48:58 (-9.8d)
15	↓1	Francisco Pérez Hernández	0.82395	11	Wed, 01 Mar 2017 15:32:42 (-28.2h)
Your Best Entry ↑					
Your submission scored 0.82326 , which is not an improvement of your best score. Keep trying!					
16	new	Alberto Baena Priego	0.82326	7	Wed, 01 Mar 2017 09:42:13 (-5d)
17	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
18	↑2	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30
19	↓4	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)
20	↓4	alaineiturria	0.81891	1	Sun, 19 Feb 2017 17:56:52

Figure 12: 12 puntuación obtenida en Kaggle

```

test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]

```

9.2.1 Esquema de valoración con aprendizaje de random forest

Probemos este paquete

```

train.caret <- train.sin.na[,-3]
set.seed(1234)
control <- caret::rfeControl(functions = rfFuncs, method = "cv", number = 10)
results <- caret::rfe(train.caret[,1:20], train.caret[,21], sizes=c(1:20), rfeControl=control)
results

##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold)
##
## Resampling performance over subset size:
##
##   Variables Accuracy Kappa AccuracySD KappaSD Selected
##      1     0.7356 0.5476  0.001958 0.003753
##      2     0.7356 0.5488  0.002060 0.005799
##      3     0.7428 0.5633  0.008822 0.018186
##      4     0.7682 0.6103  0.003999 0.006967
##      5     0.7746 0.6198  0.003718 0.006957
##      6     0.8058 0.6746  0.024927 0.043455
##      7     0.8257 0.7092  0.004365 0.007473
##      8     0.8251 0.7080  0.004406 0.007606
##      9     0.8250 0.7084  0.004879 0.008312
##     10     0.8255 0.7091  0.004767 0.008148
##     11     0.8256 0.7094  0.005054 0.008533
##     12     0.8263 0.7104  0.004471 0.007590
##     13     0.8254 0.7090  0.004504 0.007600
##     14     0.8257 0.7093  0.004064 0.006927
##     15     0.8252 0.7086  0.004978 0.008451
##     16     0.8262 0.7106  0.004493 0.007700
##     17     0.8254 0.7093  0.004945 0.008465
##     18     0.8268 0.7114  0.004381 0.007590      *
##     19     0.8266 0.7110  0.004279 0.007404
##     20     0.8262 0.7098  0.004392 0.007524

##
## The top 5 variables (out of 18):
##   TOT_VEHICULOS_IMPLICADOS, SUPERFICIE_CALZADA, COMUNIDAD_AUTONOMA, TRAZADO_NO_INTERSEC, TIPO_INTER
predictors(results)

## [1] "TOT_VEHICULOS_IMPLICADOS" "SUPERFICIE_CALZADA"
## [3] "COMUNIDAD_AUTONOMA"      "TRAZADO_NO_INTERSEC"
## [5] "TIPO_INTERSEC"          "ZONA"
## [7] "FACTORES_ATMOSFERICOS"   "ZONA_AGRUPADA"
## [9] "TIPO_VIA"                "RED_CARRETERA"
## [11] "TOT_HERIDOS_LEVES"       "LUMINOSIDAD"

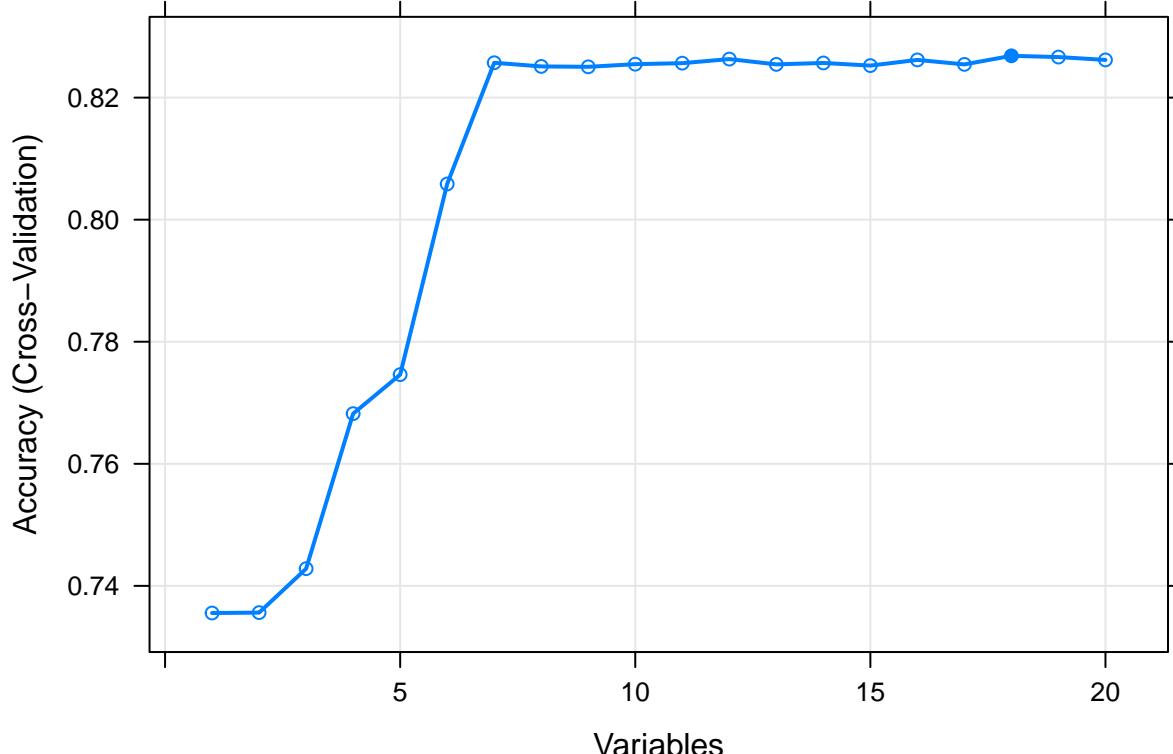
```

```

## [13] "TOT_VICTIMAS"           "DIASEMANA"
## [15] "TOT_HERIDOS_GRAVES"     "ANIO"
## [17] "TOT_MUERTOS"            "MES"

plot(results, type=c("g", "o"), lw=2)

```



Por lo tanto nos quedamos con las variables: TOT_VEHICULOS_IMPLICADOS, SUPERFICI_CALZADA, COMUNIDAD_AUTONOMA, TRAZADO_NO_INTERSEC, TIPO_INTERSEC, FACTORES_ATMOSFERICOS y ZONA.

```

train.caret <- train.caret[,c(11,18,5,16,17,20,12,21)]
test.caret <- test.sin.na[,c(12,19,6,17,18,21,13)]

```

9.2.2 Prueba del modelo

Hagamos por lo tanto una prueba.

```

set.seed(1234)
ct13 <- ctree(TIPO_ACCIDENTE ~., train.caret)
testPred13 <- predict(ct13, newdata = test.caret)

```

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

```
#ct13
```

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

```

salida.modelo.13 <- as.matrix(testPred13)
salida.modelo.13 <- cbind(c(1:(dim(salida.modelo.13)[1])), salida.modelo.13)
colnames(salida.modelo.13) <- c("Id", "Prediction")
write.table(salida.modelo.13,file="predicciones/Prediccion13.txt",sep=",",quote = F,row.names = F)

```

El resultado de este modelo para la competición de Kaggle, subido el 03/03/2017 a las 12:44, con un total

de 28 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82286. Por lo que ha empeorado muy poco a la mejor obtenida por mi.

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑5	Xisco Fauli	0.83027	21	Fri, 03 Mar 2017 01:46:51 (-3.9h)
3	↓1	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	14	Thu, 02 Mar 2017 13:31:23 (-10.7d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	14	Thu, 02 Mar 2017 20:08:41 (-17d)
16	—	Francisco Pérez Hernández	0.82395	13	Fri, 03 Mar 2017 11:44:32 (-3d)
Your Best Entry ↑					
Your submission scored 0.82286 , which is not an improvement of your best score. Keep trying!					
17	↓2	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
18	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
19	↑3	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30
20	↓3	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)

Figure 13: 13 puntuación obtenida en Kaggle

9.3 Paquete Boruta

Para este apartado comenzaremos con los dataset originales.

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]
```

9.3.1 Obtención de estadísticas sobre los atributos

```
set.seed(1234)
Bor.son <- Boruta(TIPO_ACCIDENTE~, data=train.sin.na,doTrace = 2)

## 1. run of importance source...
## 2. run of importance source...
## 3. run of importance source...
## 4. run of importance source...
## 5. run of importance source...
## 6. run of importance source...
## 7. run of importance source...
## 8. run of importance source...
## 9. run of importance source...
## 10. run of importance source...
## 11. run of importance source...
## 12. run of importance source...
## After 12 iterations, +5.2 mins:

## confirmed 19 attributes: ANIO, COMUNIDAD_AUTONOMA, DIASEMANA, FACTORES_ATMOSFERICOS, HORA and 14 mo
## rejected 1 attribute: ISLA;
## still have 1 attribute left.

## 13. run of importance source...
## 14. run of importance source...
## 15. run of importance source...
## 16. run of importance source...
## 17. run of importance source...
## 18. run of importance source...
## 19. run of importance source...
## 20. run of importance source...
## 21. run of importance source...
## 22. run of importance source...
## 23. run of importance source...
## 24. run of importance source...
## 25. run of importance source...
## 26. run of importance source...
## 27. run of importance source...
## 28. run of importance source...
## 29. run of importance source...
```

```

## 30. run of importance source...
## 31. run of importance source...
## 32. run of importance source...
## 33. run of importance source...
## 34. run of importance source...
## 35. run of importance source...
## 36. run of importance source...
## 37. run of importance source...
## 38. run of importance source...
## 39. run of importance source...
## 40. run of importance source...
## 41. run of importance source...
## 42. run of importance source...
## After 42 iterations, +17 mins:
## confirmed 1 attribute: MES;
## no more attributes left.
Bor.son

```

```

## Boruta performed 42 iterations in 17.29001 mins.
## 20 attributes confirmed important: ANIO, COMUNIDAD_AUTONOMA,
## DIASEMANA, FACTORES_ATMOSFERICOS, HORA and 15 more;
## 1 attributes confirmed unimportant: ISLA;

```

```

stats <- attStats(Bor.son)
stats

```

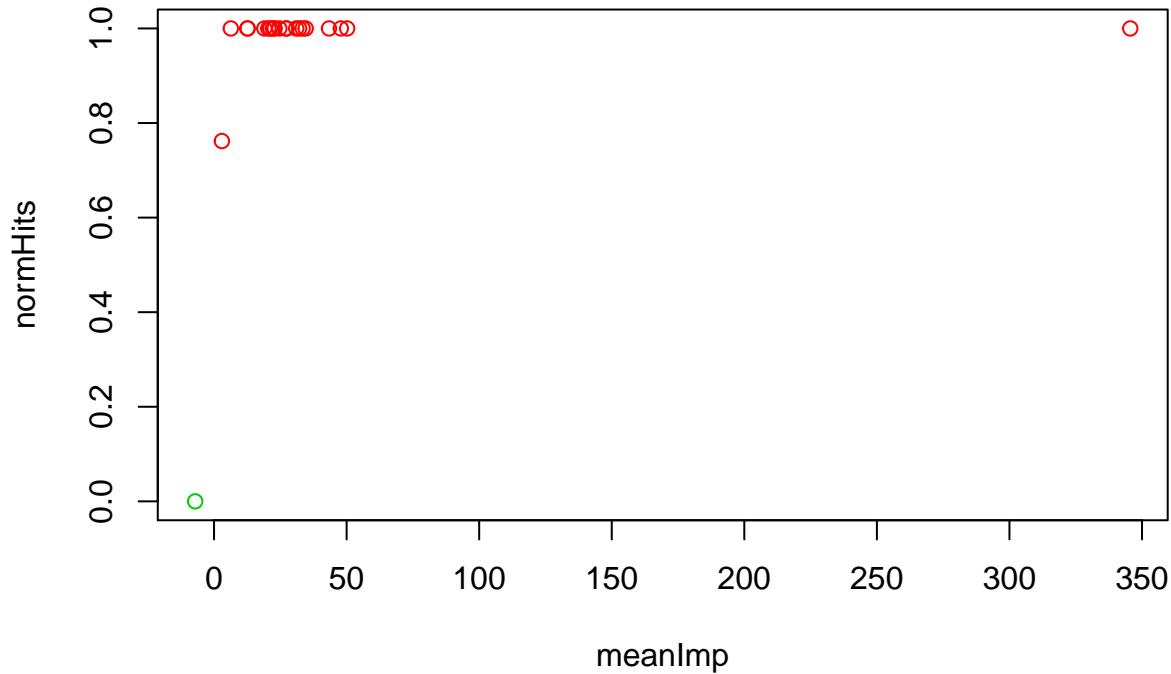
	meanImp	medianImp	minImp	maxImp
## ANIO	6.337452	6.453078	3.4110212	9.403673
## MES	2.952967	2.911986	0.8062434	5.076589
## HORA	21.189500	21.016267	17.9746464	24.026017
## DIASEMANA	12.540347	12.491457	9.8553525	16.466104
## PROVINCIA	24.624474	24.655544	21.6214213	27.746619
## COMUNIDAD_AUTONOMA	32.095891	32.037744	29.1281293	34.360393
## ISLA	-7.118103	-7.268114	-9.6533542	-4.918010
## TOT_VICTIMAS	22.850232	22.900747	19.4130266	25.070471
## TOT_MUERTOS	12.697489	12.785330	10.9135491	14.431201
## TOT_HERIDOS_GRAVES	18.989096	18.918382	17.5523079	21.376157
## TOT_HERIDOS_LEVES	27.056570	27.001044	23.9341939	31.273737
## TOT_VEHICULOS_IMPLICADOS	345.574591	347.174147	315.2261709	368.695487
## ZONA	34.563310	34.646987	31.9239781	36.505290
## ZONA_AGRUPADA	31.005910	30.995802	29.1147874	32.665360
## RED_CARRETERA	22.174816	22.175171	20.6267151	24.823239
## TIPO_VIA	27.390177	27.509398	24.8809769	29.615937
## TRAZADO_NO_INTERSEC	47.817823	48.306608	44.4681576	51.167342
## TIPO_INTERSEC	50.201198	50.118075	43.8547046	55.095652
## SUPERFICIE_CALZADA	43.352704	43.806093	38.8227329	47.769905
## LUMINOSIDAD	20.487392	20.533000	17.8024958	22.848690

```

## FACTORES_ATMOSFERICOS      33.419599 33.481231 29.8107475 36.641064
##                                         normHits decision
## ANIO                           1.0000000 Confirmed
## MES                            0.7619048 Confirmed
## HORA                           1.0000000 Confirmed
## DIASEMANA                      1.0000000 Confirmed
## PROVINCIA                      1.0000000 Confirmed
## COMUNIDAD_AUTONOMA             1.0000000 Confirmed
## ISLA                            0.0000000 Rejected
## TOT_VICTIMAS                   1.0000000 Confirmed
## TOT_MUERTOS                     1.0000000 Confirmed
## TOT_HERIDOS_GRAVES             1.0000000 Confirmed
## TOT_HERIDOS_LEVES              1.0000000 Confirmed
## TOT_VEHICULOS_IMPLICADOS       1.0000000 Confirmed
## ZONA                            1.0000000 Confirmed
## ZONA_AGRUPADA                  1.0000000 Confirmed
## RED_CARRETERA                  1.0000000 Confirmed
## TIPO_VIA                        1.0000000 Confirmed
## TRAZADO_NO_INTERSEC             1.0000000 Confirmed
## TIPO_INTERSEC                   1.0000000 Confirmed
## SUPERFICIE_CALZADA             1.0000000 Confirmed
## LUMINOSIDAD                      1.0000000 Confirmed
## FACTORES_ATMOSFERICOS          1.0000000 Confirmed

plot(normHits~meanImp, col=stats$decision, data=stats)

```



```

stats[stats$minImp > 20,]

##                                     meanImp medianImp    minImp    maxImp normHits
## PROVINCIA                      24.62447  24.65554  21.62142  27.74662      1
## COMUNIDAD_AUTONOMA             32.09589  32.03774  29.12813  34.36039      1
## TOT_HERIDOS_LEVES              27.05657  27.00104  23.93419  31.27374      1
## TOT_VEHICULOS_IMPLICADOS      345.57459 347.17415 315.22617 368.69549      1

```

```

## ZONA          34.56331 34.64699 31.92398 36.50529    1
## ZONA_AGRUPADA 31.00591 30.99580 29.11479 32.66536    1
## RED_CARRETERA 22.17482 22.17517 20.62672 24.82324    1
## TIPO_VIA      27.39018 27.50940 24.88098 29.61594    1
## TRAZADO_NO_INTERSEC 47.81782 48.30661 44.46816 51.16734    1
## TIPO_INTERSEC   50.20120 50.11808 43.85470 55.09565    1
## SUPERFICIE_CALZADA 43.35270 43.80609 38.82273 47.76990    1
## FACTORES_ATMOSFERICOS 33.41960 33.48123 29.81075 36.64106    1
##               decision
## PROVINCIA      Confirmed
## COMUNIDAD_AUTONOMA Confirmed
## TOT_HERIDOS_LEVES Confirmed
## TOT_VEHICULOS_IMPLICADOS Confirmed
## ZONA          Confirmed
## ZONA_AGRUPADA Confirmed
## RED_CARRETERA Confirmed
## TIPO_VIA      Confirmed
## TRAZADO_NO_INTERSEC Confirmed
## TIPO_INTERSEC Confirmed
## SUPERFICIE_CALZADA Confirmed
## FACTORES_ATMOSFERICOS Confirmed

```

Hemos visto como Boruta nos indica que el atributo que podemos quitar es ISLA, pero además he podado por minImp < 20, quedándome con los atributos: PROVINCIA, COMUNIDAD_AUTONOMA, TOT_HERIDOS_LEVES, TOT_VEHICULOS_IMPLICADOS, ZONA, ZONA_AGRUPADA, RED_CARRETERA, TIPO_VIA, TRAZADO_NO_INTERSEC, TIPO_INTERSEC, SUPERFICIE_CALZADA y FACTORES_ATMOSFERICOS. Vamos a probar un modelo con estas variables.

```

train.boruta <- train.sin.na[,c(5,6,11,12,13,14,15,16,17,18,19,21,22)]
test.boruta <- test.sin.na[,c(5,6,11,12,13,14,15,16,17,18,19,21)]
```

9.3.2 Prueba del modelo

Hagamos por lo tanto una prueba.

```

set.seed(1234)
ct14 <- ctree(TIPO_ACCIDENTE ~., train.boruta)
testPred14 <- predict(ct14, newdata = test.boruta)
```

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

```
#ct14
```

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

```

salida.modelo.14 <- as.matrix(testPred14)
salida.modelo.14 <- cbind(c(1:(dim(salida.modelo.14)[1])), salida.modelo.14)
colnames(salida.modelo.14) <- c("Id","Prediction")
write.table(salida.modelo.14,file="predicciones/Prediccion14.txt",sep=",",quote = F,row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 03/03/2017 a las 12:47, con un total de 28 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82326. Por lo que ha empeorado muy poco a la mejor obtenida por mi.

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑5	Xisco Fauli	0.83027	21	Fri, 03 Mar 2017 01:46:51 (-3.9h)
3	↓1	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	14	Thu, 02 Mar 2017 13:31:23 (-10.7d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	14	Thu, 02 Mar 2017 20:08:41 (-17d)
16	—	Francisco Pérez Hernández	0.82395	14	Fri, 03 Mar 2017 11:47:04 (-3d)
Your Best Entry ↑					
Your submission scored 0.82326 , which is not an improvement of your best score. Keep trying!					
17	↓2	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
18	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
19	↑3	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30
20	↓3	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)

Figure 14: 14 puntuación obtenida en Kaggle

9.3.3 Combinación con random forest

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]

train.boruta.random森林 <- train.sin.na[,-3]
set.seed(1234)
Bor.ran <- Boruta(TIPO_ACCIDENTE~, data=train.boruta.random森林, doTrace=2)

## 1. run of importance source...
## 2. run of importance source...
## 3. run of importance source...
## 4. run of importance source...
## 5. run of importance source...
## 6. run of importance source...
## 7. run of importance source...
## 8. run of importance source...
## 9. run of importance source...
## 10. run of importance source...
## 11. run of importance source...
## After 11 iterations, +4.4 mins:
## confirmed 18 attributes: ANIO, COMUNIDAD_AUTONOMA, DIASEMANA, FACTORES_ATMOSFERICOS, LUMINOSIDAD and
## rejected 1 attribute: ISLA;
## still have 1 attribute left.
## 12. run of importance source...
## 13. run of importance source...
## 14. run of importance source...
## 15. run of importance source...
## 16. run of importance source...
## 17. run of importance source...
## 18. run of importance source...
## 19. run of importance source...
## 20. run of importance source...
## 21. run of importance source...
## 22. run of importance source...
## 23. run of importance source...
## 24. run of importance source...
```

```
## 25. run of importance source...
## 26. run of importance source...
## 27. run of importance source...
## 28. run of importance source...
## 29. run of importance source...
## 30. run of importance source...
## 31. run of importance source...
## 32. run of importance source...
## 33. run of importance source...
## 34. run of importance source...
## 35. run of importance source...
## 36. run of importance source...
## 37. run of importance source...
## 38. run of importance source...
## 39. run of importance source...
## 40. run of importance source...
## 41. run of importance source...
## 42. run of importance source...
## 43. run of importance source...
## 44. run of importance source...
## 45. run of importance source...
## 46. run of importance source...
## 47. run of importance source...
## 48. run of importance source...
## 49. run of importance source...
## 50. run of importance source...
## 51. run of importance source...
## 52. run of importance source...
## 53. run of importance source...
## 54. run of importance source...
## 55. run of importance source...
## 56. run of importance source...
## 57. run of importance source...
## 58. run of importance source...
## 59. run of importance source...
## 60. run of importance source...
```

```

## 61. run of importance source...
## 62. run of importance source...
## 63. run of importance source...
## 64. run of importance source...
## 65. run of importance source...
## 66. run of importance source...
## 67. run of importance source...
## 68. run of importance source...
## 69. run of importance source...
## 70. run of importance source...
## 71. run of importance source...
## 72. run of importance source...
## 73. run of importance source...
## 74. run of importance source...
## 75. run of importance source...
## 76. run of importance source...
## 77. run of importance source...
## 78. run of importance source...
## 79. run of importance source...
## 80. run of importance source...
## 81. run of importance source...
## 82. run of importance source...
## 83. run of importance source...
## After 83 iterations, +33 mins:
## confirmed 1 attribute: MES;
## no more attributes left.

model1 <- randomForest(TIPO_ACCIDENTE~, data=train.boruta.random森林)
model1

##
## Call:
## randomForest(formula = TIPO_ACCIDENTE ~ ., data = train.boruta.random森林)
##           Type of random forest: classification
##                   Number of trees: 500
## No. of variables tried at each split: 4
##
##           OOB estimate of error rate: 17.36%
## Confusion matrix:
##             Atropello Colision_Obstaculo Colision_Vehiculos Otro
## Atropello          3060                  8                134    15
## Colision_Obstaculo      335                 36                366    18

```

```

## Colision_Vehiculos      1          0          16514    2
## Otro                      634        14          745    88
## Salida_Via                468         3          461    18
## Vuelco                     354        12          190    13
##                               Salida_Via Vuelco  class.error
## Atropello                  420        5 0.1598023064
## Colision_Obstaculo        195        2 0.9621848739
## Colision_Vehiculos         3          0 0.0003631961
## Otro                      312       14 0.9513004981
## Salida_Via                 5045       18 0.1609845335
## Vuelco                     449       50 0.9531835206

model2 <- randomForest(train.boruta.random.forest[,getSelectedAttributes(Bor.ran)],train.boruta.random.fo...  

model2

##  

## Call:  

##   randomForest(x = train.boruta.random.fo...[ , getSelectedAttributes(Bor.ran)] ,           y = train.borut...  

##               Type of random forest: classification  

##                           Number of trees: 500  

## No. of variables tried at each split: 4  

##  

##           OOB estimate of error rate: 17.38%  

## Confusion matrix:  

##                               Atropello Colision_Obstaculo Colision_Vehiculos Otro
## Atropello                  3055          4          134    15
## Colision_Obstaculo        330          38          366    22
## Colision_Vehiculos         0          0          16514    3
## Otro                      635          14          746    87
## Salida_Via                 470          1          460    15
## Vuelco                     349          12          190    15
##                               Salida_Via Vuelco  class.error
## Atropello                  427        7 0.1611751785
## Colision_Obstaculo        193        3 0.9600840336
## Colision_Vehiculos         3          0 0.0003631961
## Otro                      316        9 0.9518539015
## Salida_Via                 5049       18 0.1603193082
## Vuelco                     456       46 0.9569288390

Bor.ran

## Boruta performed 83 iterations in 32.53357 mins.  

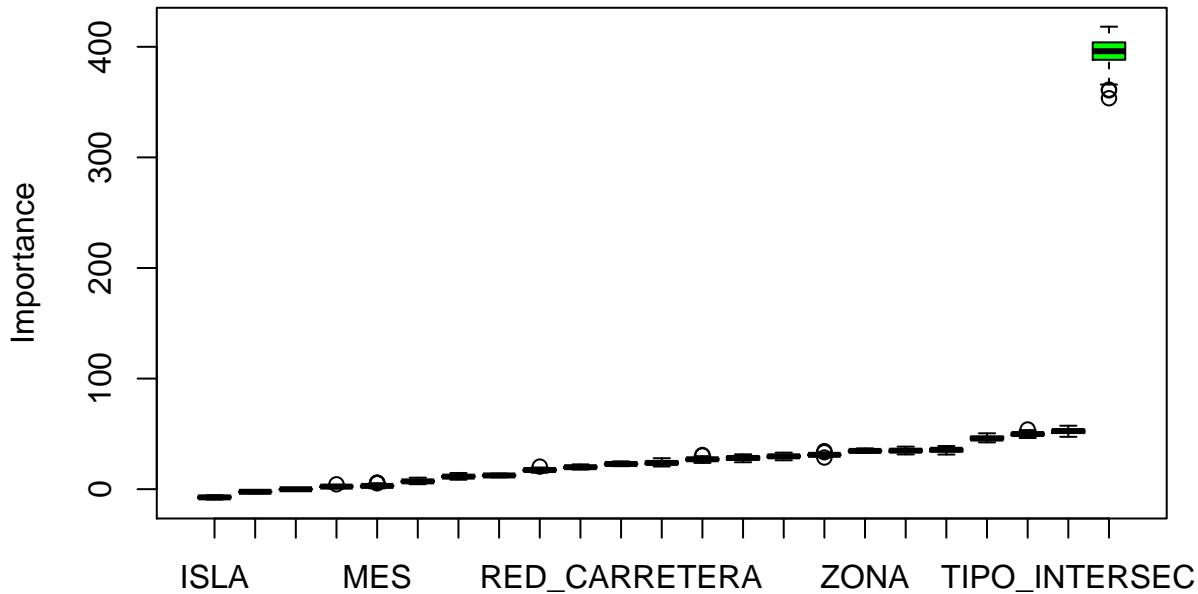
## 19 attributes confirmed important: ANIO, COMUNIDAD_AUTONOMA,  

## DIASEMANA, FACTORES_ATMOSFERICOS, LUMINOSIDAD and 14 more;  

## 1 attributes confirmed unimportant: ISLA;  

plot(Bor.ran)

```



Attributes

Vemos como el atributo que nos indica que se puede eliminar es ISLA. Probemos el modelo sin esta variable.

```
train.boruta.random森林 <- train.boruta.random森林[,-6]
test.boruta.random森林 <- test.sin.na[,-3]
test.boruta.randomForest <- test.boruta.randomForest[,-6]
```

9.3.4 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct15 <- ctree(TIPO_ACCIDENTE ~., train.boruta.randomForest)
testPred15 <- predict(ct15, newdata = test.boruta.randomForest)
```

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

```
#ct15
```

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

```
salida.modelo.15 <- as.matrix(testPred15)
salida.modelo.15 <- cbind(c(1:(dim(salida.modelo.15)[1])), salida.modelo.15)
colnames(salida.modelo.15) <- c("Id", "Prediction")
write.table(salida.modelo.15, file="predicciones/Prediccion15.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 03/03/2017 a las 12:51, con un total de 28 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82286. Por lo que ha empeorado muy poco a la mejor obtenida por mi.

10 Detección de ruido

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
```

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	24	Tue, 21 Feb 2017 17:46:15 (-5.1d)
2	↑5	Xisco Fauli	0.83027	21	Fri, 03 Mar 2017 01:46:51 (-3.9h)
3	↓1	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	14	Thu, 02 Mar 2017 13:31:23 (-10.7d) 1w improvement: none
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	14	Thu, 02 Mar 2017 20:08:41 (-17d)
16	—	Francisco Pérez Hernández	0.82395	15	Fri, 03 Mar 2017 11:51:18 (-3d)
Your Best Entry ↑					
Your submission scored 0.82286 , which is not an improvement of your best score. Keep trying!					
17	↓2	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
18	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
19	↑3	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30
20	↓3	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)

Figure 15: 15 puntuación obtenida en Kaggle

```

test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]

```

Vamos a probar este método.

```

#install.packages("NoiseFiltersR")
#library(NoiseFiltersR)
out <- IPF(TIPO_ACCIDENTE~, data=train.original,s=2)

## Iteration 1: 305 noisy instances removed
## Iteration 2: 10 noisy instances removed
## Iteration 3: 0 noisy instances removed
#summary(out, explicit=TRUE)
identical(out$cleanData, train.original[setdiff(1:nrow(train.original),out$remIdx),])

## [1] TRUE
train.sin.ruido1 <- train.original[setdiff(1:nrow(train.original),out$remIdx),]

set.seed(1234)
out <- IPF(TIPO_ACCIDENTE~, data=train.sin.ruido1,s=2)

## Iteration 1: 313 noisy instances removed
## Iteration 2: 0 noisy instances removed
## Iteration 3: 0 noisy instances removed
#summary(out, explicit=TRUE)
train.sin.ruido2 <- train.sin.ruido1[setdiff(1:nrow(train.sin.ruido1),out$remIdx),]

set.seed(1234)
out <- IPF(TIPO_ACCIDENTE~, data=train.sin.ruido2,s=2)

## Iteration 1: 300 noisy instances removed
## Iteration 2: 8 noisy instances removed
## Iteration 3: 0 noisy instances removed
#summary(out, explicit=TRUE)
train.sin.ruido3 <- train.sin.ruido2[setdiff(1:nrow(train.sin.ruido2),out$remIdx),]

set.seed(1234)
out <- IPF(TIPO_ACCIDENTE~, data=train.sin.ruido3,s=2)

## Iteration 1: 297 noisy instances removed
## Iteration 2: 4 noisy instances removed
## Iteration 3: 2 noisy instances removed
#summary(out, explicit=TRUE)
train.sin.ruido4 <- train.sin.ruido3[setdiff(1:nrow(train.sin.ruido3),out$remIdx),]

set.seed(1234)
out <- IPF(TIPO_ACCIDENTE~, data=train.sin.ruido4,s=2)

```

```

## Iteration 1: 293 noisy instances removed
## Iteration 2: 5 noisy instances removed
## Iteration 3: 0 noisy instances removed
#summary(out, explicit=TRUE)
train.sin.ruido5 <- train.sin.ruido4[setdiff(1:nrow(train.sin.ruido4),out$remIdx),]

numero.de.eliminadas <- nrow(train.original)-nrow(train.sin.ruido5)
numero.de.eliminadas

## [1] 1537

```

Por lo que se han eliminado en total 1537 instancias al ser ruido, pero podrían ser más o menos, depende del número de pasadas que se decidan hacer.

Hagamos ahora un selección de características, con el mejor método hasta ahora, cfs, para crear un modelo y realizar una prueba.

```

set.seed(1234)
subset <- FSelector::cfs(TIPO_ACCIDENTE~,train.sin.ruido5)
el.mejor.segun.cfs <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.cfs

## TIPO_ACCIDENTE ~ TOT_VEHICULOS_IMPLICADOS
## <environment: 0x7fcbe596498>

```

Siendo TOT_VEHICULOS_IMPLICADOS, la mejor característica, repitamos esto para obtener las 5 mejores, de forma que eliminamos la que mejor se obtiene.

```

set.seed(1234)
train.wrapper.cfs = train.sin.ruido5[,-12]
subset <- FSelector::cfs(TIPO_ACCIDENTE~,train.wrapper.cfs)
el.mejor.segun.cfs <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.cfs

```

```

## TIPO_ACCIDENTE ~ ZONA_AGRUPADA + CARRETERA + TRAZADO_NO_INTERSEC
## <environment: 0x7fcbe5b134e8>

```

Ahora obtenemos ZONA_AGRUPADA, CARRETERA y TRAZADO_NO_INTERSEC.

```

set.seed(1234)
train.wrapper.cfs = train.sin.ruido5[,-c(12,14,15,18)]
subset <- FSelector::cfs(TIPO_ACCIDENTE~,train.wrapper.cfs)
el.mejor.segun.cfs <- as.simple.formula(subset, "TIPO_ACCIDENTE")
el.mejor.segun.cfs

```

```

## TIPO_ACCIDENTE ~ TOT_HERIDOS_LEVES + ZONA + RED_CARRETERA + TIPO_VIA +
##      TIPO_INTERSEC + PRIORIDAD + SUPERFICIE_CALZADA + ACERAS
## <environment: 0x7fcbe214ed48>

```

Obteniendo: TOT_HERIDOS_LEVES, ZONA, RED_CARRETERA, TIPO_VIA, TIPO_INTERSEC, PRIORIDAD, SUPERFICIE_CALZADA y ACERAS.

Por lo tanto, ya que tenemos muchas características, vamos a probar el modelo con todas, salvo CARRETERA.

```

train.deteccion.de.ruido <- train.sin.ruido5[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO_INTERSEC")]
test.deteccion.de.ruido <- test.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO_INTERSEC")]

```

Vamos a probar este modelo.

10.0.1 Prueba del modelo

Hagamos por lo tanto una prueba.

```
set.seed(1234)
ct16 <- ctree(TIPO_ACCIDENTE ~., train.deteccion.de.ruido)
testPred16 <- predict(ct16, newdata = test.deteccion.de.ruido)
```

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

```
#ct16
```

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

```
salida.modelo.16 <- as.matrix(testPred16)
salida.modelo.16 <- cbind(c(1:(dim(salida.modelo.16)[1])), salida.modelo.16)
colnames(salida.modelo.16) <- c("Id", "Prediction")
write.table(salida.modelo.16, file="predicciones/Prediccion16.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 03/03/2017 a las 13:00, con un total de 28 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82365. Por lo que ha empeorado muy poco a la mejor obtenida por mi, cosa que no era de esperar ya que es el mismo modelo que el mejor obtenido, pero eliminando instancias de ruido.

11 Modelo SVM Radial

Limpiamos el espacio de trabajo

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]
train.mejor.resultado.sin.na <- train.original[,c("TOT_VEHICULOS_IMPLICADOS", "ZONA_AGRUPADA", "TRAZADO_NO")]
test.mejor.resultado.sin.na <- test.original[,c("TOT_VEHICULOS_IMPLICADOS", "ZONA_AGRUPADA", "TRAZADO_NO")]
```

Vamos a aplicar el modelo SVM Radial a nuestros datos.

```
set.seed(1234)
modelo.svm.radial <- train(TIPO_ACCIDENTE ~., data = train.mejor.resultado.sin.na, methods = "svmRadial")
predicciones.svm.radial <- predict(modelo.svm.radial, test.mejor.resultado.sin.na)
```

Veamos la puntuación en kaggel

```
salida.modelo.17 <- as.matrix(predicciones.svm.radial)
salida.modelo.17 <- cbind(c(1:(dim(salida.modelo.17)[1])), salida.modelo.17)
colnames(salida.modelo.17) <- c("Id", "Prediction")
write.table(salida.modelo.17, file="predicciones/Prediccion17.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggle, subido el 04/03/2017 a las 16:28, con un total de 29 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.81762.

Si lo que queremos es preprocesar los datos con un centrado y escalado

```
set.seed(1234)
modelo.svm.radial2 <- train(TIPO_ACCIDENTE ~., data = train.mejor.resultado.sin.na, methods = "svmRadial")
predicciones.svm.radial2 <- predict(modelo.svm.radial2, test.mejor.resultado.sin.na)
```

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83175	25	Fri, 03 Mar 2017 11:56:25 (-14.8d)
2	↑5	Xisco Fauli	0.83027	21	Fri, 03 Mar 2017 01:46:51 (-3.9h)
3	↓1	WhiteShadow	0.82988	24	Sat, 25 Feb 2017 18:58:44 (-1.1h)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	14	Thu, 02 Mar 2017 13:31:23 (-10.7d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	14	Thu, 02 Mar 2017 20:08:41 (-17d)
16	—	Francisco Pérez Hernández	0.82395	16	Fri, 03 Mar 2017 12:00:47 (-3d)
Your Best Entry ↑					
Your submission scored 0.82365 , which is not an improvement of your best score. Keep trying!					
17	↓2	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
18	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
19	↑3	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30
20	↓3	CarlosBailon	0.82079	3	Tue, 21 Feb 2017 16:41:49 (-0.3h)

Figure 16: 16 puntuación obtenida en Kaggle

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	14	Thu, 02 Mar 2017 13:31:23 (-10.7d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	17	Sat, 04 Mar 2017 15:28:02 (-4.2d)
Your Best Entry ↑					
Your submission scored 0.81762 , which is not an improvement of your best score. Keep trying!					
17	↓2	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
18	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
19	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
20	↑3	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30

Figure 17: 17 puntuación obtenida en Kaggle

Veamos la puntuación en kaggel

```
salida.modelo.18 <- as.matrix(predicciones.svm.radial2)
salida.modelo.18 <- cbind(c(1:(dim(salida.modelo.18)[1])), salida.modelo.18)
colnames(salida.modelo.18) <- c("Id", "Prediction")
write.table(salida.modelo.18,file="predicciones/Prediccion18.txt",sep=",",quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 04/03/2017 a las 17:43, con un total de 29 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.81822.

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	14	Thu, 02 Mar 2017 13:31:23 (-10.7d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	18	Sat, 04 Mar 2017 16:43:13 (-4.2d)
Your Best Entry ↑					
Your submission scored 0.81822 , which is not an improvement of your best score. Keep trying!					
17	↓2	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
18	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
19	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40
20	↑3	Francisco Javier Campón Peinado	0.82108	3	Tue, 28 Feb 2017 12:37:30

Figure 18: 18 puntuación obtenida en Kaggel

Este método tiene un parámetro de coste que regula el coste asociado a los errores de predicción: las diferencias entre el valor predicho y el real. Es posible evaluar diferentes valores de coste directamente:

```

set.seed(1234)
modelo.svm.radial3 <- train(TIPO_ACCIDENTE ~., data = train.mejor.resultado.sin.na, methods = "svmRadial")
predicciones.svm.radial3 <- predict(modelo.svm.radial3, test.mejor.resultado.sin.na)

```

Veamos la puntuación en kaggel

```

salida.modelo.19 <- as.matrix(predicciones.svm.radial3)
salida.modelo.19 <- cbind(c(1:(dim(salida.modelo.19)[1])), salida.modelo.19)
colnames(salida.modelo.19) <- c("Id", "Prediction")
write.table(salida.modelo.19, file="predicciones/Prediccion19.txt", sep=",", quote = F, row.names = F)

```

El resultado de este modelo para la competición de Kaggel, subido el 05/03/2017 a las 12:15, con un total de 31 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82256.

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	15	Sat, 04 Mar 2017 19:18:17 (-13d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	19	Sun, 05 Mar 2017 11:15:16 (-5d)
Your Best Entry ↑					
Your submission scored 0.82256 , which is not an improvement of your best score. Keep trying!					
17	—	CarlosBailon	0.82395	6	Sun, 05 Mar 2017 10:56:27 (-0.3h)
18	↓3	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
19	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
20	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40

Figure 19: 19 puntuación obtenida en Kaggel

También se puede modificar esta llamada para que se utilicen diferentes particionados. Realicemos 2 repeticiones de validación cruzada con k = 10.

```
set.seed(1234)
modelo.svm.radial4 <- train(TIPO_ACCIDENTE ~., data = train.mejor.resultado.sin.na, methods = "svmRadial")
predicciones.svm.radial4 <- predict(modelo.svm.radial4, test.mejor.resultado.sin.na)
```

Veamos la puntuación en kaggel

```
salida.modelo.20 <- as.matrix(predicciones.svm.radial4)
salida.modelo.20 <- cbind(c(1:(dim(salida.modelo.20)[1])), salida.modelo.20)
colnames(salida.modelo.20) <- c("Id", "Prediction")
write.table(salida.modelo.20, file="predicciones/Prediccion20.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 05/03/2017 a las 12:17, con un total de 31 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82237.

12 SVM (Support Vector Machine)

Limpiamos el espacio de trabajo

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]
train.mejor.resultado.sin.na <- train.original[,c("TOT_VEHICULOS_IMPLICADOS", "ZONA_AGRUPADA", "TRAZADO_NO")]
test.mejor.resultado.sin.na <- test.original[,c("TOT_VEHICULOS_IMPLICADOS", "ZONA_AGRUPADA", "TRAZADO_NO")]
```

Vamos a probar esta técnica:

```
set.seed(1234)
modelo.svm <- e1071::svm(TIPO_ACCIDENTE~., data=train.mejor.resultado.sin.na, method="C-classification")
prediccion.svm <- predict(modelo.svm, test.mejor.resultado.sin.na)
```

Veamos la puntuación en kaggel

```
salida.modelo.21 <- as.matrix(prediccion.svm)
salida.modelo.21 <- cbind(c(1:(dim(salida.modelo.21)[1])), salida.modelo.21)
colnames(salida.modelo.21) <- c("Id", "Prediction")
write.table(salida.modelo.21, file="predicciones/Prediccion21.txt", sep=",", quote = F, row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 05/03/2017 a las 12:26, con un total de 31 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.82296.

13 Métodos ensamble de construcción de conjuntos de modelos

13.1 Bagging

Limpiamos el espacio de trabajo

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
```

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	15	Sat, 04 Mar 2017 19:18:17 (-13d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	20	Sun, 05 Mar 2017 11:17:26 (-5d)
Your Best Entry ↑					
Your submission scored 0.82237 , which is not an improvement of your best score. Keep trying!					
17	—	CarlosBailon	0.82395	6	Sun, 05 Mar 2017 10:56:27 (-0.3h)
18	↓3	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
19	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
20	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40

Figure 20: 20 puntuación obtenida en Kaggle

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	15	Sat, 04 Mar 2017 19:18:17 (-13d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	21	Sun, 05 Mar 2017 11:26:28 (-5d)
Your Best Entry ↑					
Your submission scored 0.82296 , which is not an improvement of your best score. Keep trying!					
17	—	CarlosBailon	0.82395	6	Sun, 05 Mar 2017 10:56:27 (-0.3h)
18	↓3	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
19	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
20	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40

Figure 21: Puntuación obtenida en Kaggle

```

test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]
train.mejor.resultado.sin.na <- train.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO")]
test.mejor.resultado.sin.na <- test.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO")]

set.seed(1234)
modelo.bagging.1 <- adabag::bagging(TIPO_ACCIDENTE~,data=train.mejor.resultado.sin.na, control=rpart::control)
predicciones.bagging.1 <- adabag::predict.bagging(modelo.bagging.1,newdata=test.mejor.resultado.sin.na)

set.seed(1234)
modelo.bagging.2 <- adabag::bagging(TIPO_ACCIDENTE~,data=train.mejor.resultado.sin.na,mfinal=20 ,control)
predicciones.bagging.2 <- adabag::predict.bagging(modelo.bagging.2,newdata=test.mejor.resultado.sin.na)

```

Veamos la puntuación en kaggel

```

salida.modelo.22 <- as.matrix(predicciones.bagging.1$class)
salida.modelo.22 <- cbind(c(1:(dim(salida.modelo.22)[1])), salida.modelo.22)
colnames(salida.modelo.22) <- c("Id","Prediction")
write.table(salida.modelo.22,file="predicciones/Prediccion22.txt",sep=",",quote = F,row.names = F)

```

El resultado de este modelo para la competición de Kaggel, subido el 05/03/2017 a las 12:31, con un total de 31 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.81891.

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best - Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	15	Sat, 04 Mar 2017 19:18:17 (-13d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	22	Sun, 05 Mar 2017 11:31:27 (-5d)
Your Best Entry ↑					
Your submission scored 0.81891 , which is not an improvement of your best score. Keep trying!					
17	—	CarlosBailon	0.82395	6	Sun, 05 Mar 2017 10:56:27 (-0.3h)
18	↓3	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
19	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
20	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40

Veamos la puntuación en kaggel

```
salida.modelo.23 <- as.matrix(predicciones.bagging.2$class)
salida.modelo.23 <- cbind(c(1:(dim(salida.modelo.23)[1])), salida.modelo.23)
colnames(salida.modelo.23) <- c("Id","Prediction")
write.table(salida.modelo.23,file="predicciones/Prediccion23.txt",sep=",",quote = F,row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 05/03/2017 a las 12:37, con un total de 31 personas entregadas, se ha quedado en la posición 16 con una puntuación del 0.81891.

13.2 Random forest

Limpiamos el espacio de trabajo

#	Δ1w	Team Name	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	Anabel Gómez	0.83185	33	Sat, 04 Mar 2017 13:16:19
2	↑5	Xisco Fauli	0.83027	22	Fri, 03 Mar 2017 23:35:17 (-25.7h)
3	↓1	WhiteShadow	0.82988	32	Sat, 04 Mar 2017 15:11:29 (-6.9d)
4	↓1	JacintoCC	0.82958	2	Sun, 19 Feb 2017 17:39:23 (-0h)
5	↓1	Luis Suárez	0.82948	3	Mon, 13 Feb 2017 08:11:24 (-2.5d)
6	↓1	RubenSanchez	0.82899	18	Fri, 03 Mar 2017 08:27:15 (-5.9d)
7	↓1	Jonathan Espinosa	0.82889	32	Thu, 02 Mar 2017 09:53:07 (-4h)
8	—	David Urdin	0.82820	3	Thu, 23 Feb 2017 17:49:50
9	—	Jorge Jimena	0.82741	34	Tue, 28 Feb 2017 00:56:54 (-3d)
10	new	Gzmg2102	0.82612	10	Fri, 03 Mar 2017 11:22:03 (-0.4h)
11	↓1	ManuelMontero	0.82582	15	Sat, 04 Mar 2017 19:18:17 (-13d)
12	↓1	RonCR	0.82582	10	Wed, 22 Feb 2017 11:35:56 (-11.8h)
13	↓1	Salva Moreno	0.82573	12	Thu, 23 Feb 2017 13:54:33 (-2.6d)
14	↓1	Mauricio Orellana	0.82563	12	Sat, 25 Feb 2017 00:08:49 (-22h)
15	↓1	BesayMontesdeocaSantana	0.82543	16	Fri, 03 Mar 2017 16:40:50 (-17.9d)
16	—	Francisco Pérez Hernández	0.82395	23	Sun, 05 Mar 2017 11:37:02 (-5d)
Your Best Entry ↑					
Your submission scored 0.81891 , which is not an improvement of your best score. Keep trying!					
17	—	CarlosBailon	0.82395	6	Sun, 05 Mar 2017 10:56:27 (-0.3h)
18	↓3	Alberto Baena Priego	0.82385	8	Wed, 01 Mar 2017 20:18:00
19	new	lorenadq	0.82148	2	Fri, 03 Mar 2017 15:41:27 (-0.4h)
20	new	CristinaZuhe	0.82118	2	Mon, 27 Feb 2017 23:13:40

Figure 22: 23 puntuación obtenida en Kaggle

```

rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]
train.mejor.resultado.sin.na <- train.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO_INTERSEC")]
test.mejor.resultado.sin.na <- test.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO_INTERSEC")]

set.seed(1234)
modelo.random森林.10 <- randomForest::randomForest(TIPO_ACCIDENTE~., data=train.mejor.resultado.sin.na, ntree = 10)
print(modelo.random森林.10)

##
## Call:
##   randomForest(formula = TIPO_ACCIDENTE ~ ., data = train.mejor.resultado.sin.na,      ntree = 10)
##   Type of random forest: classification
##   Number of trees: 10
##   No. of variables tried at each split: 3
##
##       OOB estimate of error rate: 18.19%
## Confusion matrix:
##             Atropello Colision_Obstaculo Colision_Vehiculos Otro
##   Atropello          3028              6            134    18
##   Colision_Obstaculo     330             40            353    14
##   Colision_Vehiculos        8             18           16267    26
##   Otro                  646             24            731    55
##   Salida_Via           504             21            443    55
##   Vuelco                351              7            179    36
##
##             Salida_Via Vuelco class.error
##   Atropello          418              7 0.161451122
##   Colision_Obstaculo     200              7 0.957627119
##   Colision_Vehiculos        40              9 0.006170577
##   Otro                  305              27 0.969239374
##   Salida_Via           4886             38 0.178409282
##   Vuelco                450              31 0.970588235

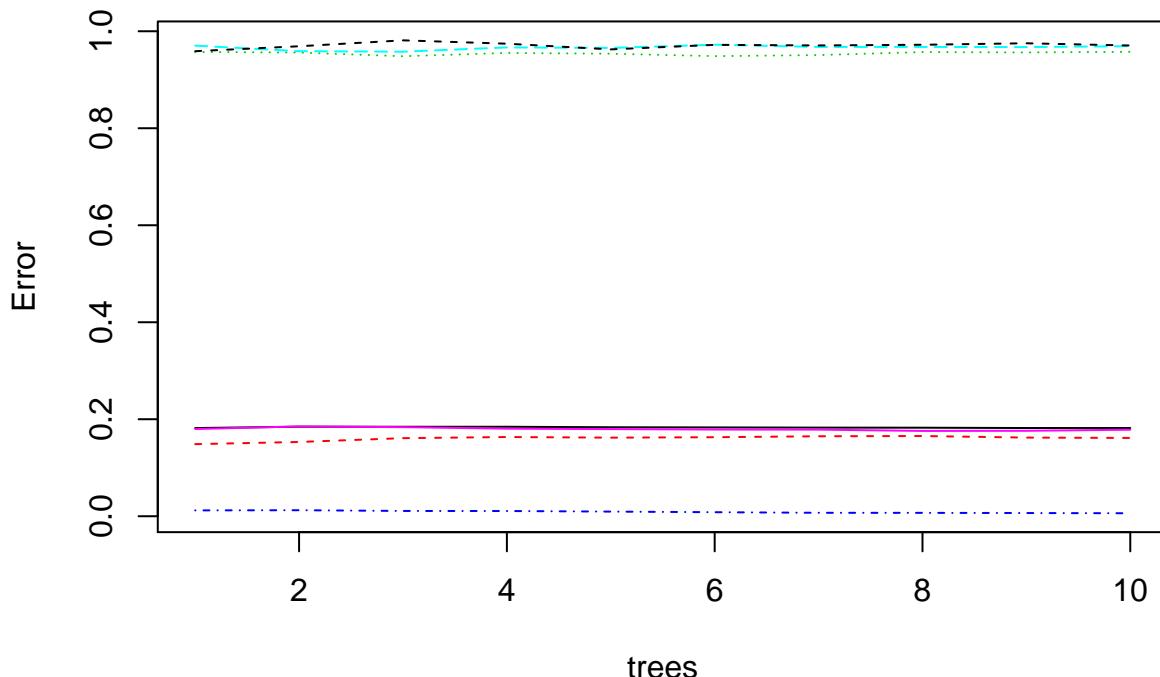
randomForest::importance(modelo.random森林.10)

##
##               MeanDecreaseGini
##   TOT_VEHICULOS_IMPLICADOS    7949.0655
##   ZONA_AGRUPADA                 921.2303
##   TRAZADO_NO_INTERSEC            325.9847
##   TOT_HERIDOS_LEVES              229.3772
##   ZONA                           744.0811
##   RED_CARRETERA                  127.7767
##   TIPO_VIA                         359.3147
##   TIPO_INTERSEC                   270.0264
##   SUPERFICIE_CALZADA                245.3562

plot(modelo.random森林.10)

```

modelo.random森林.10



```
predicciones.rf.10 <- predict(modelo.random.forest.10,newdata=test.mejor.resultado.sin.na)
```

Veamos la puntuación en kaggel

```
salida.modelo.24 <- as.matrix(predicciones.rf.10)
salida.modelo.24 <- cbind(c(1:(dim(salida.modelo.24)[1])), salida.modelo.24)
colnames(salida.modelo.24) <- c("Id", "Prediction")
write.table(salida.modelo.24,file="predicciones/Prediccion24.txt",sep=",",quote = F,row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 06/03/2017 a las :, con un total de 3X personas entregadas, se ha quedado en la posición XX con una puntuación del 0..

```
set.seed(1234)
modelo.random.forest.100 <- randomForest::randomForest(TIPO_ACCIDENTE~.,data=train.mejor.resultado.sin.na)
print(modelo.random.forest.100)
```

```
##
## Call:
##   randomForest(formula = TIPO_ACCIDENTE ~ ., data = train.mejor.resultado.sin.na,
##                 ntree = 100)
##   Type of random forest: classification
##   Number of trees: 100
##   No. of variables tried at each split: 3
##
##   OOB estimate of  error rate: 17.64%
##   Confusion matrix:
##   Atropello Colision_Obstaculo Colision_Vehiculos Otro
##   Atropello          3066            3         132   11
##   Colision_Obstaculo      324           29         361   27
##   Colision_Vehiculos        1            1       16502    4
##   Otro                  638           16         751   68
##   Salida_Via          478            6         459   28
```

19.2. Random forest

Se trata también de una técnica **ensemble** que genera un co*dección que se usan posteriormente para generar predicciones. El de generación de modelos se presenta a continuación:*

- seleccionar el número de modelos a crear, m
- para cada modelo a construir
 - generar una muestra usando **bootstrap** a partir de los n datos
 - generar un modelo (usualmente un árbol) a partir de esta muestra y este nuevo nodo a introducir:
 - seleccionar al azar k variables predictoras de las p disponibles, donde k es menor que el número de variables disponibles al principio

Figure 23: 24 puntuación obtenida en Kaggle

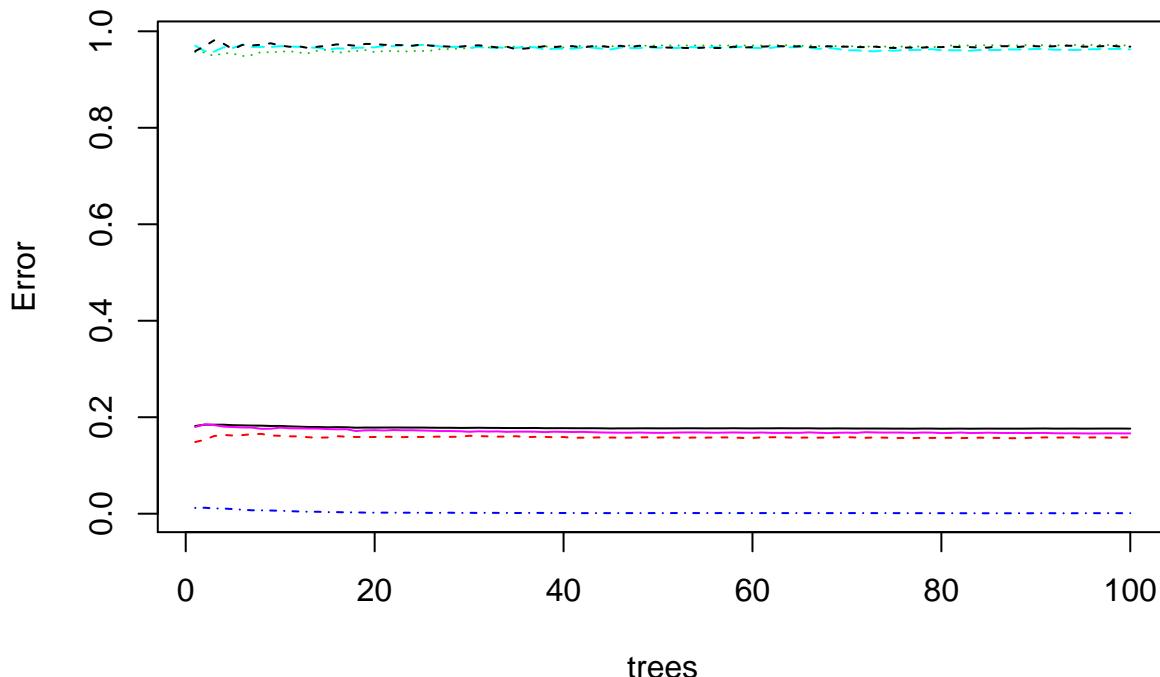
```
## Vuelco           346          8          188    40
## Salida_Via Vuelco class.error
## Atropello        427          3 0.158154860
## Colision_Obstaculo 209          2 0.969537815
## Colision_Vehiculos 12          0 0.001089588
## Otro              321          13 0.962368567
## Salida_Via       5012         30 0.166472643
## Vuelco            452          34 0.968164794

randomForest::importance(modelo.random森林.100)

##                                     MeanDecreaseGini
## TOT_VEHICULOS_IMPLICADOS      8066.3529
## ZONA_AGRUPADA                  526.6196
## TRAZADO_NO_INTERSEC            401.7404
## TOT_HERIDOS_LEVES              228.8489
## ZONA                           634.5463
## RED_CARRETERA                 202.4311
## TIPO_VIA                       494.7834
## TIPO_INTERSEC                  285.8930
## SUPERFICIE_CALZADA             267.8467

plot(modelo.random森林.100)
```

modelo.random.forest.100



```
predicciones.rf.100 <- predict(modelo.random.forest.100,newdata=test.mejor.resultado.sin.na)
```

Veamos la puntuación en kaggel

```
salida.modelo.25 <- as.matrix(predicciones.rf.100)
salida.modelo.25 <- cbind(c(1:(dim(salida.modelo.25)[1])), salida.modelo.25)
colnames(salida.modelo.25) <- c("Id", "Prediction")
write.table(salida.modelo.25,file="predicciones/Prediccion25.txt",sep=",",quote = F,row.names = F)
```

El resultado de este modelo para la competición de Kaggel, subido el 06/03/2017 a las :, con un total de 3X personas entregadas, se ha quedado en la posición XX con una puntuación del 0..

13.3 Boosting

Limpiamos el espacio de trabajo

```
rm(list=ls())
train.original <- read.csv("accidentes-kaggle.csv")
test.original <- read.csv("accidentes-kaggle-test.csv")
train.sin.na <- train.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24,30)]
test.sin.na <- test.original[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,22,23,24)]
train.mejor.resultado.sin.na <- train.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO")]
test.mejor.resultado.sin.na <- test.original[,c("TOT_VEHICULOS_IMPLICADOS","ZONA_AGRUPADA","TRAZADO_NO")]
```

Vamos a probar este método:

```
set.seed(1234)
modelo.boosting <- adabag::boosting(TIPO_ACCIDENTE~,data=train.mejor.resultado.sin.na, mfinal = 10, con
barplot(modelo.boosting$imp[order(modelo.boosting$imp, decreasing = TRUE)],
ylim = c(0, 100), main = "Variables Relative Importance",
col = "lightblue")
```

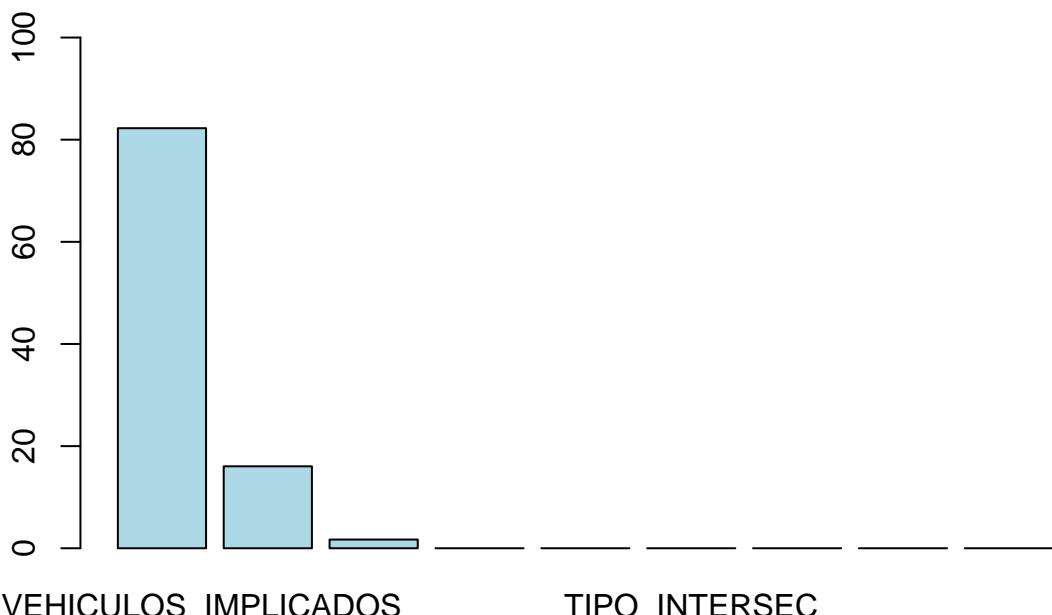
19.2. Random forest

Se trata también de una técnica **ensemble** que genera un coí decisión que se usan posteriormente para generar predicciones. El de generación de modelos se presenta a continuación:

- seleccionar el número de modelos a crear, m
- para cada modelo a construir
 - generar una muestra usando **bootstrap** a partir de los c
 - generar un modelo (usualmente un árbol) a partir de esté nuevo nodo a introducir:
 - seleccionar al azar k variables predictoras de las di menor que el número de variables disponibles al prir

Figure 24: 25 puntuación obtenida en Kaggle

Variables Relative Importance



```
prediccion.boosting <- predict.boosting(modelo.boosting, newdata = test.mejor.resultado.sin.na)
modelo.boosting$importance[modelo.boosting$importance>0]
```

```
## TOT_VEHICULOS_IMPLICADOS          ZONA          ZONA_AGRUPADA
##                 82.246757        1.709175       16.044068
```

Veamos la puntuación en kaggel

```

salida.modelo.26 <- as.matrix(prediccion.boosting$class)
salida.modelo.26 <- cbind(c(1:(dim(salida.modelo.26)[1])), salida.modelo.26)
colnames(salida.modelo.26) <- c("Id", "Prediction")
write.table(salida.modelo.26,file="predicciones/Prediccion26.txt",sep=",",quote = F,row.names = F)

```

El resultado de este modelo para la competición de Kaggel, subido el 06/03/2017 a las :, con un total de 3X personas entregadas, se ha quedado en la posición XX con una puntuación del 0..

19.2. Random forest

Se trata también de una técnica **ensemble** que genera un coi decisión que se usan posteriormente para generar predicciones. El de generación de modelos se presenta a continuación:

- seleccionar el número de modelos a crear, m
- para cada modelo a construir
 - generar una muestra usando **bootstrap** a partir de los n datos
 - generar un modelo (usualmente un árbol) a partir de esta muestra y este nuevo nodo a introducir:
 - seleccionar al azar k variables predictoras de las p disponibles, donde k es menor que el número de variables disponibles al principio

Figure 25: 26 puntuación obtenida en Kaggel

Notas: Tengo que subir la 24 y 25.