

Memoria Competición Kaggel Preprocesamiento

Francisco Pérez Hernández

17/2/2017

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.

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 SUFICIENTE : 4793 OTRO : 715
## NOCHE: SIN ILUMINACION : 1815 LLUVIA FUERTE: 499
## PLENO DIA :20997 VIENTO FUERTE: 156
## NIEBLA LIGERA: 83
## (Other) : 173
## VISIBILIDAD_RESTRINGIDA OTRA_CIRCUNSTANCIA
## SIN RESTRICCION :16982 NINGUNA :24967
## CONFIGURACION DEL TERRENO: 989 OTRA : 942
## OTRA_CAUSA : 491 OBRAS : 263
## FACTORES ATMOSFERICOS : 374 FUERTE DESCENSO : 227
## EDIFICIOS : 229 CAMBIO DE RASANTE: 100
## (Other) : 252 (Other) : 264
## NA's :10685 NA's : 3239
## ACERAS DENSIDAD_CIRCULACION MEDIDAS_ESPECIALES
## NO HAY ACERA:21416 CONGESTIONADA: 308 CARRIL REVERSIBLE : 17
## SI HAY ACERA: 5437 DENSA : 1479 HABILITACION ARCEN: 8
## NA's : 3149 FLUIDA :17505 NINGUNA MEDIDA :21024
## NA's :10710 OTRA MEDIDA : 278
## NA's : 8675
##
## TIPO_ACCIDENTE
## Atropello : 3642
## Colision_Obstaculo: 952
## Colision_Vehiculos:16520
## Otro : 1807
## Salida_Via : 6013
## Vuelco : 1068
##

```

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

```

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

```

1.2 Primera prueba con un modelo

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

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

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

```
ct

##
## 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 kaggle

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

```
dada <- as.matrix(testPred)
write.table(dada,file="predicciones/PrimeraPrediccion.txt",sep="," ,quote = F)
```

Por lo que ya tenemos un fichero con la salida del conjunto de test. Lo único que tendremos que modificar es la primera línea del archivo para añadir "Id, Prediction". El resultado de este primer modelo para la competición de Kaggle, 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		Luis Suárez	0.82948	2	Fri, 10 Feb 2017 19:54:58
2		fgraggel	0.81120	3	Thu, 09 Feb 2017 17:40:29 (-8d)
3	new	PacoPollos	0.73246	1	Sat, 11 Feb 2017 18:51:32
4		Héctor Garbisu	0.55147	1	Thu, 02 Feb 2017 20:42:24
5		Francisco Javier Campón Peinado	0.55147	1	Fri, 03 Feb 2017 20:15:10

Figure 1: Primera puntuación obtenida en Kaggle