Algoritmos de Gran Escala

Andrea García Tapia, Andrea Frenández , Mario Becerra

24 de mayo de 2015

```
library(glmnet)
```

```
## Loading required package: Matrix
## Loaded glmnet 1.9-8
```

```
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
##
## The following object is masked from 'package:Matrix':
##
##     expand
```

```
library(ggplot2)
library(lubridate)
library(stringr)
knitr::opts_chunk$set(warning=FALSE, cache=TRUE, echo=FALSE)
```

Análisis Exploratorio de Datos

```
## Joining by: c("Dependiente", "Tipo.de.fenomeno", "Duracion", "MUNICIPIOS.afectados", "ANAL", "SPRI",
```

México ha tenido un incremento en los costos económicos de desastres asociados a fenómenos hidrometeorológicos, huracanes e inundaciones, entre otros. En 2010 se presentaron las mayores pérdidas económicas en la historia del país por fenómenos hidrometeorológicos y geológicos; en total se perdió el 0.8% del PIB y se estima que, una vez calculado en su totalidad, el daño por las tormentas tropicales Ingrid y Manuel en 2013 supere los valores anteriores.

Una pregunta clave que todavía no se contesta en México es si este incremento en daños y pérdidas se debe a un cambio en la distribución de los desastres o a observaciones atípicas. El Sistema de Protección Civil (SINAPROC) define desastre "al resultado de la ocurrencia de uno o más agentes perturbadores severos y o extremos, concatenados o no, de origen natural o de la actividad humana, que cuando acontecen en un tiempo y en una zona determinada, causan daños y que por su magnitud exceden la capacidad de respuesta de la comunidad afectada"; sin embargo no esta definida qué es la capacidad de respuesta de la comunidad afectada ni existen indicadores. Nuestro sistema es reactivo y las reglas de operación no son muy claras. EL Panel Intergubernamental de Cambio Climático (IPCC) prevee un aumento en la frecuencia e intensidad de los desastres hidrometeoroógicos debido al cambio climático.

Actualmente el SINAPROC funciona de la siguiente manera: cuando ocurre un desastre el Gobierno Estatal solicita una evaluación al Gobierno Federal. Este a su vez solicita al Servicio Meteorológico Nacional (SMN), al Sismológico, Comisión Nacional Forestal (CONAFOR) o al Centro Nacional de Prevención de Desastres (CENAPRED), dependiendo el tipo de desastre, la corroboración del evento. Una vez corroborado el Gobierno Federal decide si lo declara o no . Si lo declara tiene tres opciones: Contingencia Climática, Desastre, Emergencia o una combinación de las últimas dos. Esta declaratoria hace toda la diferencia ya que si no es declarado, el evento solo recibe ayuda de protección civil local. Por el contario si lo declaran desastre (contingencia climatica, desastre o emergencia) se activa el programa de reconstrucción del FONDEN, el programa de apoyos de SAGARPA (CADENA) y diversos programas de apoyo social como el programa de Empleo Temporal de SEDESOL. Es por ello que es tan importante tener reglas claras. Este proyecto busca clarificar las reglas del proceso de declaratoria de desastres naturales y encontrar un modelo que ayude al Gobierno Federal acelerar los procesos de decalratoria, ya que actuar de manera oportuna es vital.

Los datos fueron obtenidos del Centro Nacional de Prevención de Desastres (CENAPRED) para los desastres Hidrometeorológicos de 2000-2010. La base se llama Impacto Socio Económico y es con la que realizan la serie anual de los libros con el mismo nombre. Se unió con la base Marginación de CONEVAL y con una base de Riesgos realizada por el Centro Mario Molina (CMM). La base de Riesgos fue realizada para 5 peligros (huracán, inundación, sequía, incendio forestal, deslave) calculados a partir de las características geofísicas del país y las tasas de retorno de los desastres.

Descripción del Dataset

La base se conforma de 25 variables, entre las cuales hay características geográficas (riesgos), características socioeconómicas de la población y características del evento.

Tipo de declaratoria (dependiente)	Tipo de declaratoria, según el diario oficial de la federación (1 si fue declarado, 0 eoc))
ANAL	Porcentaje de población analfabeta de 15 años o más
SPRI	Porcentaje de población sin primaria completa de 15 años o más
OVSDS	Porcentaje de ocupantes en viviendas sin drenaje ni servicio sanitario exclusivo
OVSEE	Porcentaje de ocupantes en viviendas sin energía eléctrica
OVSAE	Porcentaje de ocupantes en viviendas sin agua entubada
VHAC	Porcentaje de viviendas con algún nivel de hacinamiento
OVPT	Porcentaje de ocupantes en viviendas con piso de tierra
PL<5000	Porcentaje de población en localidades con menos de 5 000 habitantes
PO2SM	Porcentaje de población ocupada con ingreso de hasta 2 salarios mínimos
IM	Índice de marginación
GM	Grado de marginación
Sum_POBTOT	Población total
R_Inun	Riesgo de inundación
R_Hur	Riesgo de huracán
R_Des	Riesgo Deslizamiento
R_Seq	Riesgo de sequía
R_IF	Riesgo de incendio forestal
R_Den	Riesgo de dengue
Num Mun	Número de municipios afectados por el desastre en cuestión
Fecha de Inicio	Fecha de inicio del desastre
Fecha de Fin	Fecha de fin del desastre
Año	Año de ocurrencia del desastre
Duración	Duración del desastre en días
Clave del Estado	Clave de la entidad federativa según INEGI
Municipio	Nombre del municipio del registro en cuestión
Tipo de fenomeno	Tipo de fenómeno: lluvia, inundación, deslizamiento tectónico, etc

```
## Source: local data frame [4,750 x 25]
##
##    Dependiente Tipo.de.fenomeno Duracion MUNICIPIOS.afectados ANAL  SPRI
## 1            1              SEQ       10                   58 9.00 41.17
## 2            1               BT        1                   56 3.19 16.75
```

```
## 3            1             LLUV        3                   56 3.19 16.75
## 4            0               SD        1                   56 3.19 16.75
## 5            0             LLUV        1                   56 3.19 16.75
## 6            0             LLUV        1                   56 3.19 16.75
## 7            0               FV        1                   56 2.83 13.37
## 8            0             LLUV        1                   56 2.83 13.37
## 9            0             LLUV        2                   56 2.83 13.37
## 10           0               FV        1                   56 2.83 13.37
## ..         ...              ...      ...                  ...  ...   ...
## Variables not shown: OVSDS (dbl), OVSEE (dbl), OVSAE (dbl), VHAC (dbl),
##   OVPT (dbl), PL.5000 (dbl), PO2SM (dbl), IM (dbl), GM (chr),
##   area_mun..ha. (dbl), Sum_POBTOT (int), R_Inun (int), R_Hur (int), R_Des
##   (int), R_Seq (int), R_IF (int), R_Den (int), a_o (int),
##   Clave.completa.TEXTO. (int)
```

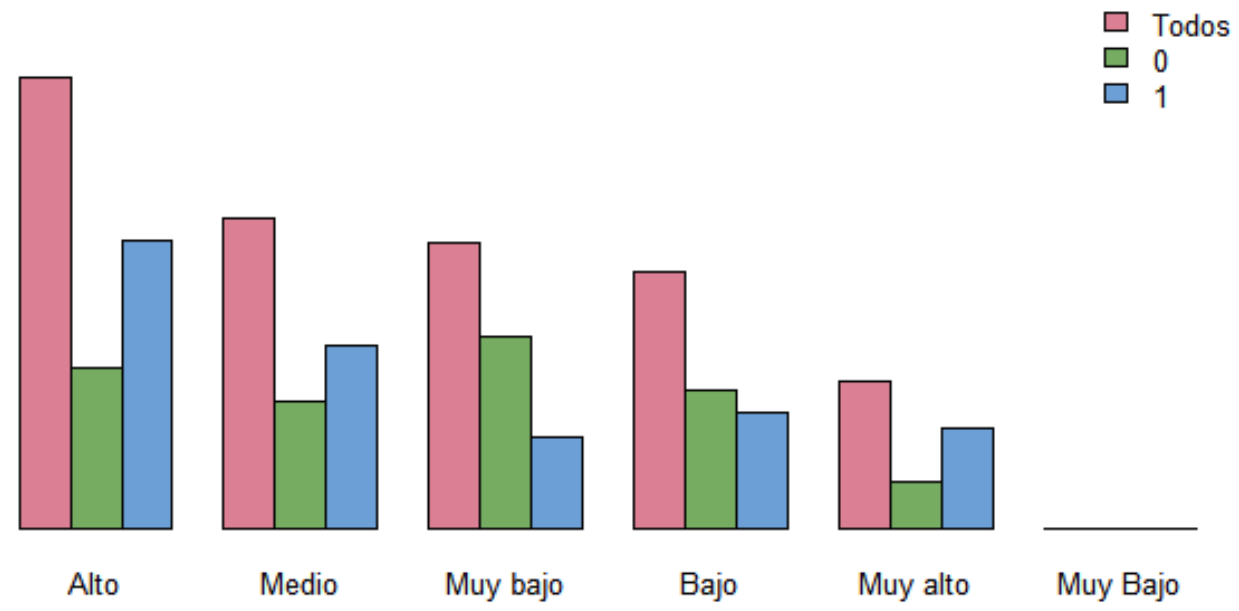
```
##           Dependiente      Tipo.de.fenomeno              Duracion
##             "integer"           "character"             "integer"
##  MUNICIPIOS.afectados                  ANAL                 SPRI
##             "integer"             "numeric"             "numeric"
##                 OVSDS                 OVSEE                 OVSAE
##             "numeric"             "numeric"             "numeric"
##                  VHAC                  OVPT               PL.5000
##             "numeric"             "numeric"             "numeric"
##                 PO2SM                    IM                    GM
##             "numeric"             "numeric"           "character"
##         area_mun..ha.            Sum_POBTOT                R_Inun
##             "numeric"             "integer"             "integer"
##                 R_Hur                 R_Des                 R_Seq
##             "integer"             "integer"             "integer"
##                  R_IF                 R_Den                   a_o
##             "integer"             "integer"             "integer"
## Clave.completa.TEXTO.
##             "integer"
```

Se dividió el conjunto de datos (4750 observaciones con 25 variables) en datos de entrenamiento (70%) y de prueba (30%).

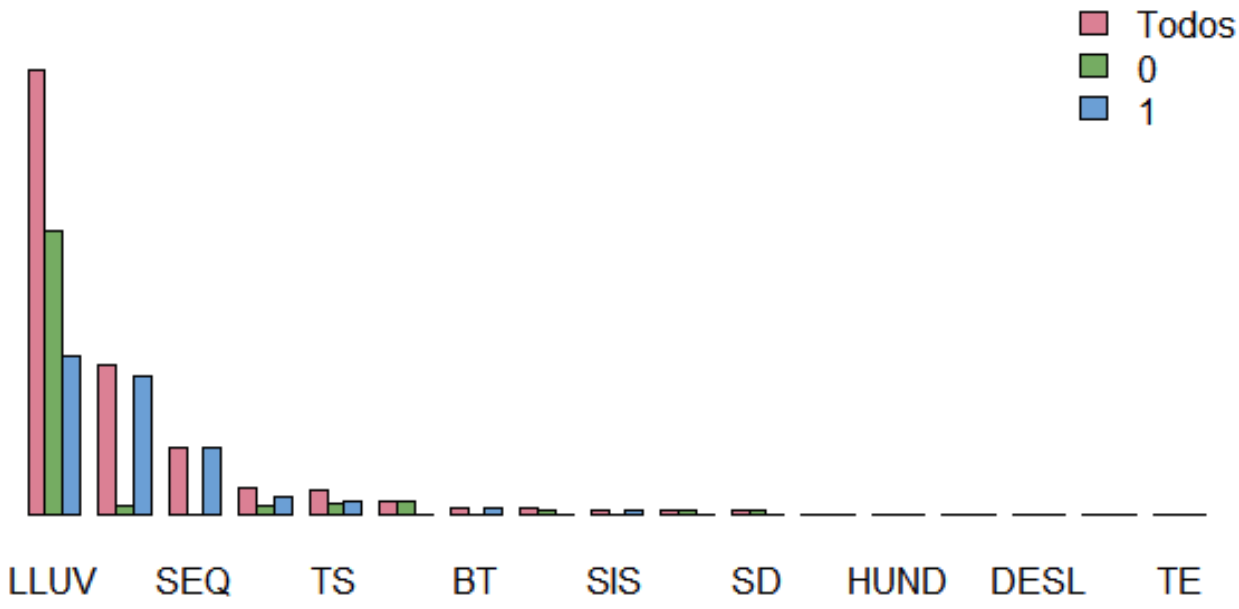
```
##   Dependiente     Tipo.de.fenomeno      Duracion
##  Min.   :0.0000   Length:4750        Min.   :  1.000
##  1st Qu.:0.0000   Class :character   1st Qu.:  1.000
##  Median :1.0000   Mode  :character   Median :  1.000
##  Mean   :0.5446                      Mean   :  5.824
##  3rd Qu.:1.0000                      3rd Qu.:  3.000
##  Max.   :1.0000                      Max.   :122.000
##
##  MUNICIPIOS.afectados      ANAL            SPRI           OVSDS
##  Min.   :  1.00       Min.   : 1.07   Min.   : 5.23   Min.   : 0.00
##  1st Qu.: 15.00       1st Qu.: 6.88   1st Qu.:25.75   1st Qu.: 2.22
##  Median : 39.00       Median :13.00   Median :37.75   Median : 5.79
##  Mean   : 69.39       Mean   :14.75   Mean   :36.77   Mean   :10.89
##  3rd Qu.: 86.00       3rd Qu.:20.11   3rd Qu.:47.06   3rd Qu.:14.27
##  Max.   :564.00       Max.   :75.81   Max.   :87.69   Max.   :93.72
##  NA's   :3            NA's   :3       NA's   :3       NA's   :3
```

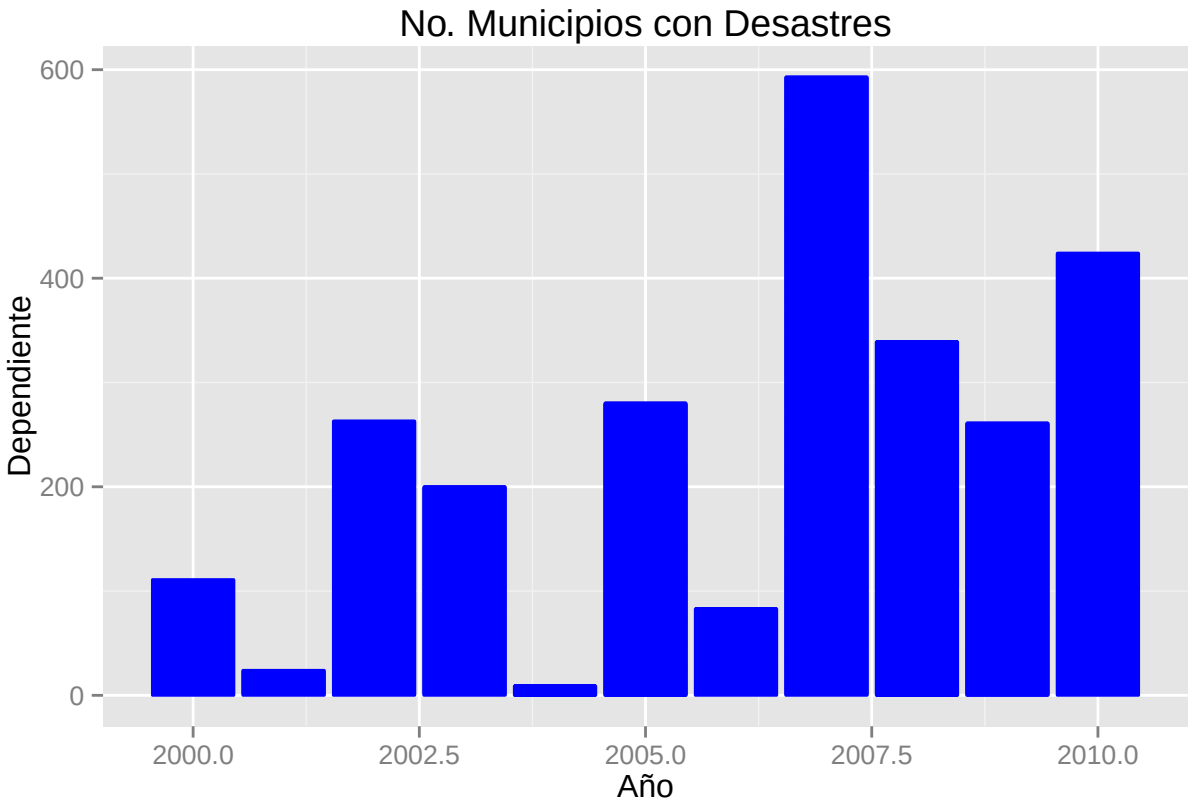
```
##      OVSEE            OVSAE            VHAC            OVPT
##  Min.   : 0.000   Min.   : 0.00   Min.   :13.86   Min.   : 0.000
##  1st Qu.: 1.360   1st Qu.: 3.47   1st Qu.:39.66   1st Qu.: 5.675
##  Median : 3.150   Median : 9.81   Median :48.49   Median :13.440
##  Mean   : 5.629   Mean   :17.51   Mean   :49.35   Mean   :19.568
##  3rd Qu.: 6.730   3rd Qu.:27.15   3rd Qu.:58.76   3rd Qu.:27.470
##  Max.   :85.670   Max.   :96.46   Max.   :89.60   Max.   :97.840
##  NA's   :3        NA's   :3       NA's   :3       NA's   :3
##     PL.5000           PO2SM             IM              GM
##  Min.   :  0.00   Min.   :11.70   Min.   :-2.4500   Length:4750
##  1st Qu.: 26.35   1st Qu.:48.81   1st Qu.:-1.0400   Class :character
##  Median : 61.12   Median :64.81   Median :-0.2700   Mode  :character
##  Mean   : 59.63   Mean   :62.39   Mean   :-0.2526
##  3rd Qu.:100.00   3rd Qu.:77.28   3rd Qu.: 0.4400
##  Max.   :100.00   Max.   :98.99   Max.   : 4.5000
##  NA's   :3        NA's   :3       NA's   :3
##  area_mun..ha.       Sum_POBTOT          R_Inun          R_Hur
##  Min.   :    221   Min.   :    310   Min.   :0.000   Min.   :1.000
##  1st Qu.:  16249   1st Qu.:  11808   1st Qu.:1.000   1st Qu.:1.000
##  Median :  42163   Median :  29699   Median :3.000   Median :3.000
##  Mean   : 126937   Mean   : 115093   Mean   :2.633   Mean   :2.917
##  3rd Qu.: 118952   3rd Qu.:  84706   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5327186   Max.   :1815786   Max.   :5.000   Max.   :5.000
##  NA's   :1         NA's   :1         NA's   :1       NA's   :1
##      R_Des            R_Seq            R_IF           R_Den
##  Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:0.000
##  Median :0.000   Median :3.00   Median :2.000   Median :0.000
##  Mean   :1.203   Mean   :2.99   Mean   :2.463   Mean   :1.143
##  3rd Qu.:2.000   3rd Qu.:4.00   3rd Qu.:3.000   3rd Qu.:0.000
##  Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000
##  NA's   :1       NA's   :1      NA's   :1       NA's   :1
##       a_o       Clave.completa.TEXTO.
##  Min.   :2000   Min.   : 1001
##  1st Qu.:2005   1st Qu.:12057
##  Median :2007   Median :20438
##  Mean   :2006   Mean   :19902
##  3rd Qu.:2008   3rd Qu.:30009
##  Max.   :2010   Max.   :32058
##
```

La distribución por Grado de Marginación nos muestra que los grados altos tienen mas declaratorias.



En cuanto al tipo de fenómeno la mayor parte de las declaratorias se concentran en lluvias y sequías.





Modelos de clasificación

Regresión Logística Regularizada

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 254 158
##          1 354 592
##
##                Accuracy : 0.623
##                  95% CI : (0.5966, 0.6488)
##     No Information Rate : 0.5523
##     P-Value [Acc > NIR] : 7.793e-08
##
##                   Kappa : 0.2136
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4178
##             Specificity : 0.7893
##          Pos Pred Value : 0.6165
##          Neg Pred Value : 0.6258
##              Prevalence : 0.4477
##          Detection Rate : 0.1870
##    Detection Prevalence : 0.3034
```

```
##       Balanced Accuracy : 0.6035
##
##        'Positive' Class : 0
##
```

Máquina de Soporte Vectorial en Paralelo

```
## [1] "intercepto"
## [1] 0.08000000000017195
## [1] "Tasa de clasificación incorrecta en entrenamiento"
## [1] 0.22
## [1] "Tasa de clasificación incorrecta en prueba"
## [1] 0.4050073637702504
```