Algoritmos de Gran Escala

Andrea García Tapia, Andrea Frenández , Mario Becerra

24 de mayo de 2015

Análisis Exploratorio de Datos

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.

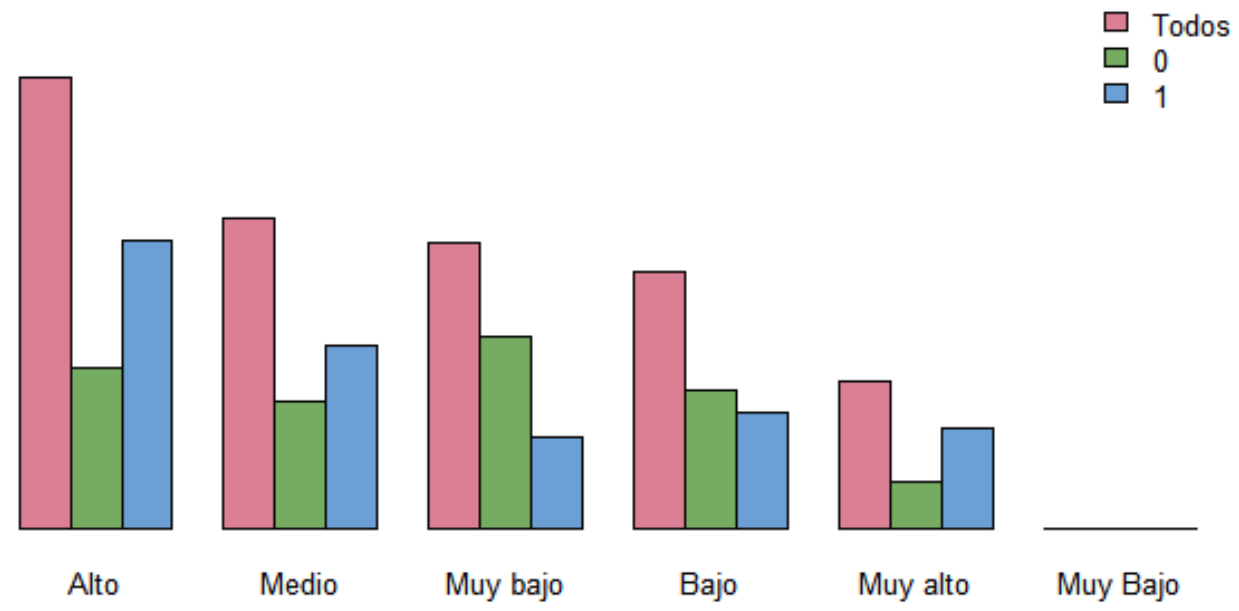
Descripción del Dataset

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

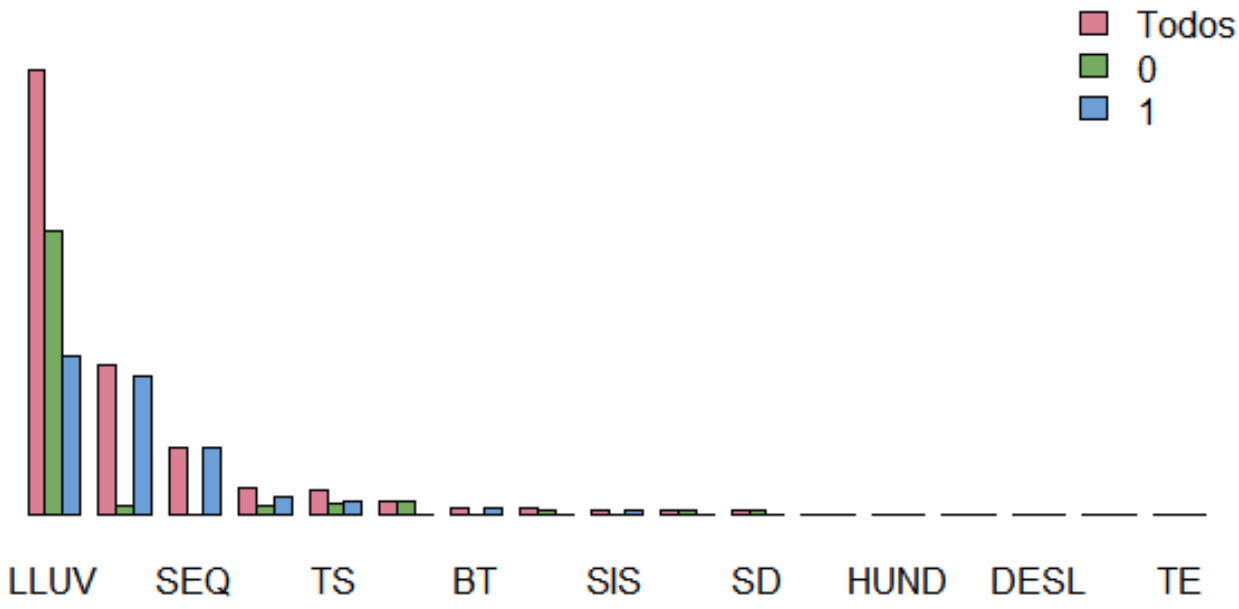
Tipo de declaratoria (dependiente)	Tipo de declaratoria, según el diario oficial de la federación (1 si fue declarado, 0 eoc))
ANAL	Porcentaje de población analfabeta de 15 años o más
SPRI	Porcentaje de población sin primaria completa de 15 años o más
OVSDS	Porcentaje de ocupantes en viviendas sin drenaje 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

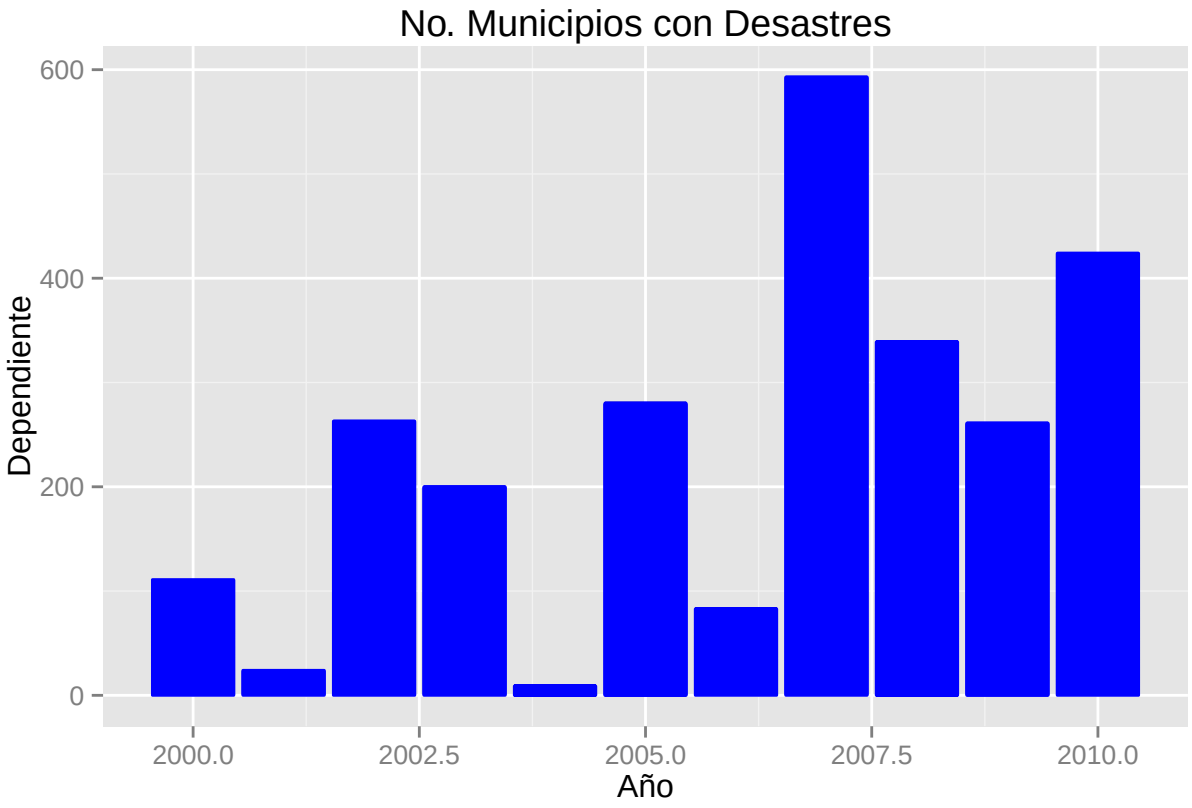
Se dividió el conjunto de datos (4750 observaciones con 25 variables) en datos de entrenamiento (70%) y de prueba (30%).

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

	0	1
0	254	158
1	354	592

Table 1: Matriz de confusión de regresión logística

	cm$byClass
Sensitivity	0.4177632
Specificity	0.7893333
Pos Pred Value	0.6165049
Neg Pred Value	0.6257928
Prevalence	0.4477172
Detection Rate	0.1870398
Detection Prevalence	0.3033873
Balanced Accuracy	0.6035482

cm$byClass

Máquina de Soporte Vectorial en Paralelo

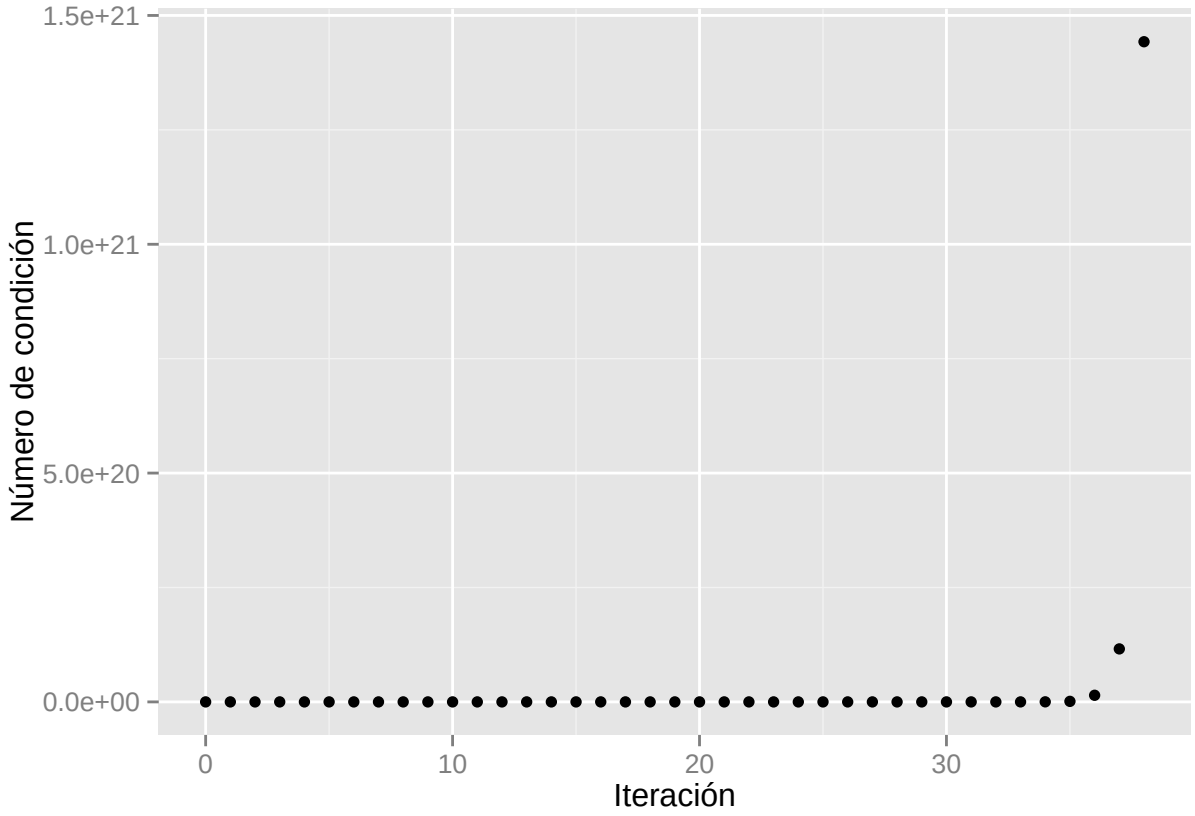
Los resultados de clasificación con la máquina de soporte vectorial fueron mejores que con la

```
## [1] "intercepto"
## [1] 0.09747292419257945
```

	-1	1
-1	312	194
1	296	556

Table 3: Matriz de confusión de SVM

	cm2$byClass
Sensitivity	0.5131578947368421
Specificity	0.7413333333333333
Pos Pred Value	0.6166007905138340
Neg Pred Value	0.6525821596244131
Prevalence	0.4477172312223859
Detection Rate	0.2297496318114874
Detection Prevalence	0.3726067746686304
Balanced Accuracy	0.6272456140350877



Problemas

Durante la elaboración de este proyecto nos enfrentamos a diversos problemas.

Implementación del Cluster

No entendíamos cómo funcionaba el NFS server:

- Al utilizar dos carpetas con nombres distintos, es decir `mirrorNFS` y `carpetaNodo`, no sabíamos que era necesario cerrar el loop y correr `mpirun` en el master desde `carpetaNodo`.
- Cuando intentamos tener un cluster en nuestras computadoras, con las diferentes versiones y distribuciones de Linux y una computadora MAC, había muchos errores.
- La versión de MPI se actualizó mientras hacíamos las tareas y con un `update` tuvimos que desinstalar y reinstalar todo de nuevo.
- En realidad, todos los problemas por el NFS se solucionaron cuando entendimos que éste era solamente una manera de pegarle al master los esclavos pero que master siempre considera que está corriendo todo en sus versiones.

Compilación de archivos:

- Las librerías de Lapack, ATLAS, BLAS y las librerías de matemáticas no se cargan en `mpicc` automáticamente para la distribución de Linux que usamos (Ubuntu, 14.04, unity). Nos tardamos mucho en entender los mensajes de compilación.
- En general, aprender a utilizar `c` para adaptar los códigos en la tarea 3 fue muy problemático.

Paralelización de SVM

- Nos pasó lo mismo que con `mpi`: la versión y localización de `R` en `master` y en cada uno de los `nodos` debe ser exactamente la misma para todas las computadoras en el cluster.
- La instalación de `rmpi` es muy problemática. Debes de realizar:

```
sudo apt-get install libcr-dev mpich2 mpich2-doc
R
install.packages('Rmpi')
install.packages('snow')
```

- Primero, intentamos con `Rmpi`. Sin embargo, fue muy difícil debuggear e intentar paralelizar el proceso. Lo que nos pasaba es que podíamos prender el cluster en el master pero freezeaba. Cuando leimos diferentes foros de ayuda, resultó que este es uno de los errores más comunes pero puede tener muchas causas. Eventualmente, al poner la bandera de `manual=T` en el comando para iniciar el cluster, si se podía pero teníamos que ir a cada nodo y ejecutar el `script` de `R` generado en cada uno. Esto no escala.
- Lo que nos funcionó mejor fue utilizar `snow` directo y mejor controlar el cluster desde ahí. Un ejemplo funcional para prenderlo de manera interactiva fue:

```r
library(snow)
# Prendes el cluster
cl <- makeSOCKcluster(c("slave05","localhost"))
# clusterApply funciona como apply pero mandas el objeto cluster,
# x *para iterar*, la funcion(suma, para ejemplificar) y parametros adicionales
clusterApply(cl, 1:2, get("+"), 3)
# clusterEvalQ permite enviar comandos a ejecutar en cada nodo para la
# sesion de R abierta en cada uno
clusterEvalQ(cl, library(boot))
# detenemos el cluster
stopCluster(cl)
```

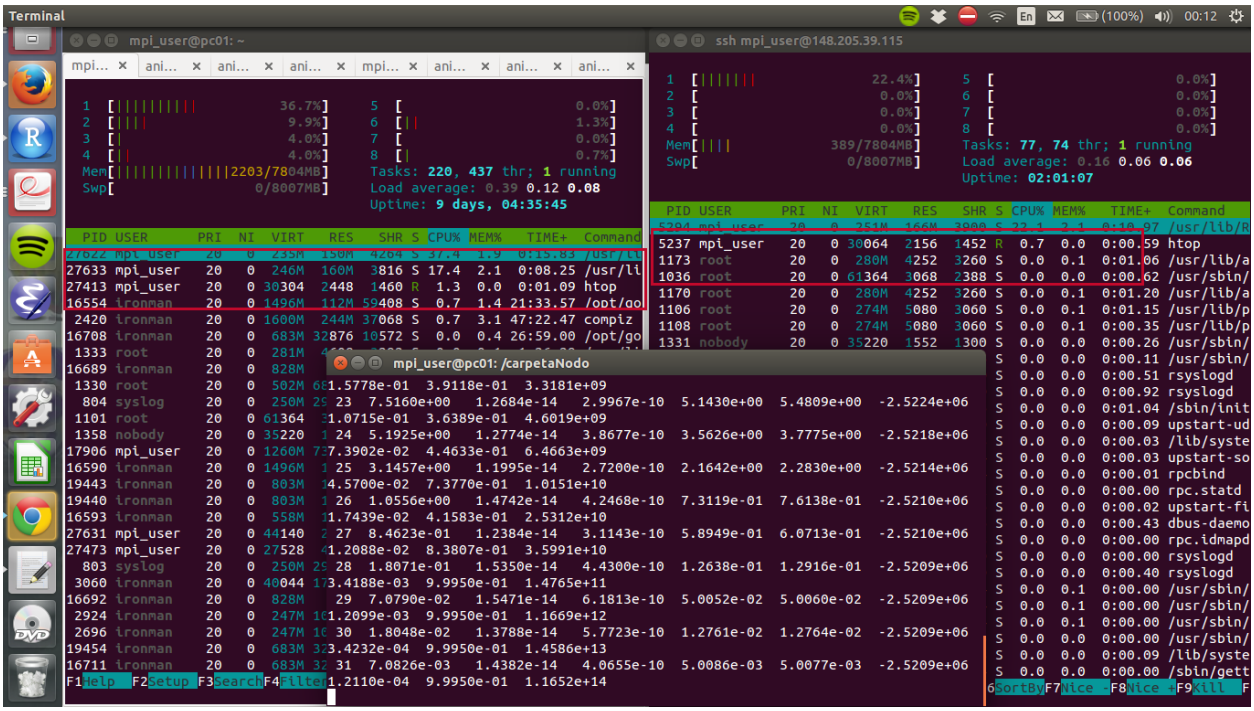
- Esta forma nos permitió dejar de utilizar `mpirun` y poder manipular el cluster desde `R`. Por esto, al final el llamado a svm en paralelo es vía `Rscript`, desde el master y en `carpetaNodo`:

```
Rscript Script.R
```

Los resultados se muestran a continuación

```
mpi_user@pc01:/carpetaNodo$ Rscript Script.R
Loading required package: methods

Attaching package: 'snow'

The following objects are masked from 'package:parallel':

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, clusterSplit, makeCluster, parApply,
    parCapply, parLapply, parRapply, parSapply, splitIndices,
    stopCluster

   i    |F|       |rcx1|        |rb|         |f3|        |f4|         obj         mu        alfa       cond
  0  1.6757e+08   1.6657e+08   1.6200e+05  1.2972e+07  1.2972e+07  6.5174e+11  2.5000e+05  5.0363e-01  1.2529e+04
  1  8.3127e+07   8.2679e+07   8.0412e+04  6.4676e+06  5.7037e+06  1.5791e+11  1.2790e+05  9.9950e-01  2.2601e+04
  2  1.6742e+06   4.1339e+04   4.0206e+01  1.3882e+06  9.3484e+05  -1.0864e+09  2.6646e+04  9.9950e-01  1.1719e+05
  3  8.2849e+05   2.0670e+01   2.0103e-02  5.7984e+05  5.9177e+05  -5.2292e+08  1.2803e+04  9.9950e-01  2.3188e+05
  4  1.1417e+05   1.0335e-02   1.0052e-05  7.3467e+04  8.7386e+04  -1.1020e+08  2.6697e+03  9.9950e-01  9.3870e+05
  5  8.2984e+04   5.1674e-06   4.9570e-09  5.4266e+04  6.2783e+04  -5.3692e+07  1.2810e+03  9.9950e-01  1.3558e+06
  6  1.5027e+04   2.5837e-09   1.1440e-09  3.7551e+03  1.4550e+04  -1.2552e+07  2.6748e+02  9.9950e-01  9.5372e+06
  7  8.4187e+03   1.2915e-12   2.6919e-10  5.2515e+03  6.5800e+03  -7.1840e+06  1.2817e+02  6.7881e-01  1.4355e+07
  8  3.9515e+03   4.1608e-13   2.8566e-10  2.3949e+03  3.1430e+03  -4.5609e+06  5.9323e+01  4.6055e-01  4.2724e+07
  9  2.6953e+03   2.3045e-13   3.3505e-10  1.6789e+03  2.1085e+03  -3.7776e+06  3.7904e+01  3.9714e-01  6.4480e+07
 10  1.8038e+03   1.3540e-13   3.2823e-10  1.1351e+03  1.4018e+03  -3.3267e+06  2.5207e+01  5.2916e-01  9.1610e+07
 11  9.6019e+02   6.8423e-14   2.3144e-10  6.1092e+02  7.4078e+02  -2.9392e+06  1.3874e+01  3.7782e-01  1.6166e+08
 12  6.8902e+02   3.7676e-14   3.8213e-10  4.4431e+02  5.2663e+02  -2.7990e+06  9.5846e+00  3.1886e-01  2.1187e+08
 13  5.0418e+02   2.5546e-14   2.0452e-10  3.2707e+02  3.8370e+02  -2.7161e+06  6.9708e+00  3.6886e-01  2.4499e+08
 14  3.3971e+02   1.5136e-14   3.0972e-10  2.2151e+02  2.5756e+02  -2.6483e+06  4.7531e+00  3.8066e-01  3.5845e+08
 15  2.2849e+02   1.2859e-14   3.0221e-10  1.4995e+02  1.7240e+02  -2.6028e+06  3.2091e+00  4.6643e-01  5.0549e+08
 16  1.3463e+02   1.1662e-14   2.7065e-10  8.9028e+01  1.0099e+02  -2.5671e+06  1.9340e+00  4.2329e-01  8.1414e+08
 17  8.7987e+01   1.2519e-14   2.7486e-10  5.8701e+01  6.5544e+01  -2.5490e+06  1.2512e+00  3.9123e-01  1.3409e+09
 18  5.9100e+01   1.1558e-14   3.7835e-10  3.9672e+01  4.3806e+01  -2.5386e+06  8.3736e-01  4.2389e-01  1.6897e+09
 19  3.7404e+01   1.2265e-14   3.1981e-10  2.5238e+01  2.7606e+01  -2.5314e+06  5.3545e-01  3.3654e-01  1.5613e+09
 20  2.7257e+01   1.2254e-14   2.9123e-10  1.8467e+01  2.0047e+01  -2.5280e+06  3.8343e-01  3.0288e-01  1.8324e+09
 21  2.0230e+01   1.1701e-14   3.2809e-10  1.3739e+01  1.4849e+01  -2.5259e+06  2.8351e-01  5.1283e-01  2.1831e+09
 22  1.0672e+01   1.2333e-14   2.4801e-10  7.2650e+00  7.8172e+00  -2.5234e+06  1.5778e-01  3.9118e-01  3.3181e+09
 23  7.5160e+00   1.2684e-14   2.9967e-10  5.1430e+00  5.4809e+00  -2.5224e+06  1.0715e-01  3.6389e-01  4.6019e+09
 24  5.1925e+00   1.2774e-14   3.8677e-10  3.5626e+00  3.7775e+00  -2.5218e+06  7.3902e-02  4.4633e-01  6.4663e+09
 25  3.1457e+00   1.1995e-14   2.7200e-10  2.1642e+00  2.2830e+00  -2.5214e+06  4.5700e-02  7.3770e-01  1.0151e+10
 26  1.0556e+00   1.4742e-14   4.2468e-10  7.3119e-01  7.6138e-01  -2.5210e+06  1.7439e-02  4.1583e-01  2.5312e+10
 27  8.4623e-01   1.2384e-14   3.1143e-10  5.8949e-01  6.0713e-01  -2.5210e+06  1.2088e-02  8.3807e-01  3.5991e+10
 28  1.8071e-01   1.5350e-14   4.4300e-10  1.2638e-01  1.2916e-01  -2.5209e+06  3.4188e-03  9.9950e-01  1.4765e+11
 29  7.0790e-02   1.5471e-14   6.1813e-10  5.0052e-02  5.0060e-02  -2.5209e+06  1.2099e-03  9.9950e-01  1.1669e+12
 30  1.8048e-02   1.3788e-14   5.7723e-10  1.2761e-02  1.2764e-02  -2.5209e+06  3.4232e-04  9.9950e-01  1.4586e+13
 31  7.0826e-03   1.4382e-14   4.0655e-10  5.0086e-03  5.0077e-03  -2.5209e+06  1.2110e-04  9.9950e-01  1.1652e+14
 32  1.8073e-03   1.5348e-14   5.1396e-10  1.2780e-03  1.2779e-03  -2.5209e+06  3.4275e-05  9.9950e-01  1.4539e+15
 33  7.0881e-04   1.5212e-14   3.5451e-10  5.0121e-04  5.0120e-04  -2.5209e+06  1.2121e-05  9.9950e-01  1.1624e+16
 34  1.8099e-04   1.4574e-14   3.8833e-10  1.2798e-04  1.2798e-04  -2.5209e+06  3.4319e-06  9.9950e-01  1.4499e+17
 35  7.0936e-05   1.4271e-14   4.5952e-10  5.0159e-05  5.0159e-05  -2.5209e+06  1.2132e-06  9.9950e-01  1.1602e+18
 36  1.8126e-05   1.3890e-14   4.8142e-10  1.2817e-05  1.2817e-05  -2.5209e+06  3.4362e-07  9.9950e-01  1.4462e+19
 37  7.0991e-06   1.5915e-14   3.6238e-10  5.0198e-06  5.0198e-06  -2.5209e+06  1.2143e-07  9.9950e-01  1.1581e+20
 38  1.8152e-06   1.4765e-14   6.0510e-10  1.2835e-06  1.2835e-06  -2.5209e+06  3.4406e-08  9.9950e-01  1.4425e+21
[1] "intercepto"
[1] 0.09747292419259111
[1] "En entrenamiento"
[1] 0.3429602888086642
[1] "En prueba"
[1] 0.3608247422680412
mpi_user@pc01:/carpetaNodo$
```

Para revisar que el trabajo se estaba realizando en master y en nodos, prendimos un `htop` en el master y el esclavo:


Terminal
mpi_user@pc01: ~
ssh mpi_user@148.205.39.115
mpi_user@pc01: /carpetaNodo