**PS2 – Big Data. Presentado por el Grupo 5: Carlos Avilán y Francisco Ortiz.**

**Enlace Git-Hub: https://github.com/faor10/Problem-Set-2/**

## 1. Introducción

En este documento presenta el proceso para predecir la pobreza en los hogares de Colombia, esto a partir de una idea inspirada en una reciente competición del Banco Mundial. Esta predicción se obtuvo a partir de dos maneras: como un problema de clasificación y como un problema de predicción del ingreso. Se explica la obtención de los datos con los que se desarrolló el proceso, los modelos utilizados y sus resultados en cada caso. Finalmente, se presentan las conclusiones en donde se dan indicios de la ventaja del uso del modelo de clasificación, del remuestreo y de la utilización de hiperparámetros para mejorar las predicciones.

## 2. Datos

Para la estimación del ingreso por persona se utilizó la variable *Ingtot* de la base de datos *test personas*. Se estimó el ingreso de diferentes maneras con las variables del DANE, por ejemplo, utilizando solo el ingreso laboral y los resultados no cambiaron mucho en este caso, por lo cual se escogió *Ingtot*. Algunas muestras en que la variable *Ingtot* no indicaba ningún valor (*missing value*) fueron eliminadas de la muestra, en su mayoría eran individuos que no reportaban ingresos como los hijos de un hogar. Imputar un valor a estos *missing values* ya sea con un cero o un valor por vecinos más cercanos lo consideramos impertinente ya que podría sesgar los resultados. Por otro lado, para aquellos *missing values* de las variables categóricas, fueron imputados a categorías como "otros" ó "no sabe no informa". La principal razón para no eliminar estas observaciones radica en que estos *missing values* representaban un gran porcentaje de la variable y eliminarlos podría generar un problema de inferencia estadística a diferencia de los valores faltantes de *Ingtot* los cuales representaban un porcentaje mínimo.

Finalmente, estas son las estadísticas descriptivas de las variables utilizadas para el desarrollo del trabajo en la base de datos donde entrenamos nuestros modelos:

| Characteristic | 0, N = 406,888[1] | 1, N = 136,696[1] |
|---|---|---|
| Ingtot | 800,134 (1,435,310) | 153,387 (269,121) |
| Ingtotugarr | 3,030,667 (2,822,409) | 889,033 (583,454) |
| Edad^2 | 1,733 (1,758) | 1,184 (1,569) |
| Edad | 36 (21) | 27 (21) |
| Arriendo | 537,356 (4,198,255) | 279,529 (1,767,826) |
| HorasTrabajadas | 23 (25) | 13 (22) |
| CuartosHogar | 3.72 (1.22) | 3.21 (1.13) |
| CuartosUsad | 2.34 (0.96) | 2.20 (0.90) |
| NPerUG | 3.94 (1.88) | 5.13 (2.39) |
| Nper | 3.96 (1.88) | 5.13 (2.38) |
| Fex_c.x | 66 (88) | 70 (96) |
| Fex_dpto.x | 66 (91) | 69 (101) |

[1] Mean (SD)

| Variable categórica | Característica |
|---|---|
| Oficio | Con 99 oficios, el oficio más frecuente es "Conductor de vehículos de transporte" con el 2.8% del total de oficios. |
| MaxNivEdu | Con 9 niveles educativos, el más frecuente con el 25% de las observaciones el máximo nivel educativo es "primaria incompleta" seguido de "secundaria incompleta" con el 24%. |
| TipoDeViv | Con 6 tipos de vivienda el más frecuente se encuentra que es "propia ó la está pagando" con el 39%, seguido de "arriendo o subarriendo" con el 37%. |

*Tabla 1. Estadísticas descriptivas de las variables utilizadas en la base de datos de entrenamiento.*

## 3.Modelos y resultados

### 3.1 Modelo de clasificación

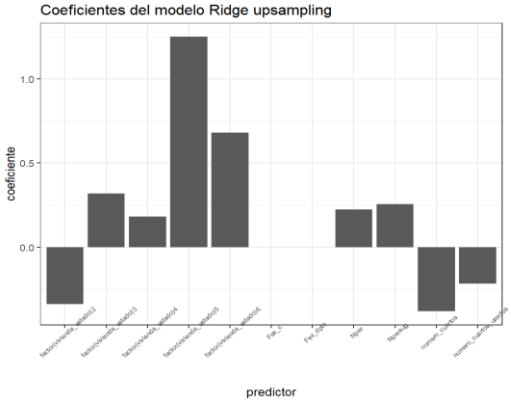
**Esta misma forma funcional se utilizó para los 7 modelos realizados en el procedimiento de clasificación.**

$$Pobre = B_0 + B_1\, Numero\ Cuartos + B_2 Numero\ Cuartos\ usados + B_3 Estado\ Vivienda + B_4 Nper + B_5 Npersug + B_6 Factor\ Expansión + B_7 Factor\ Expansión\ Depto + u$$

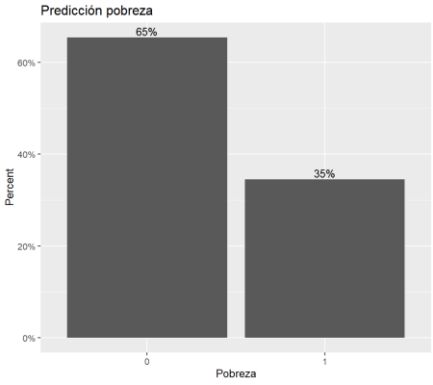
| Modelos | | Lambda mínimo | AUC | Tasa Falsos Negativos | Sensitivity |
|---|---|---|---|---|---|
| Modelo 1 | Estimación con Lasso – Alpha=1 | 0,0006911766 | 0,5714028 | 16,9% | 16,8% |
| Modelo 2 | Estimación con Ridge – Alpha=0 | 0,00957201 | 0,550705 | 17,9% | 11,73% |
| Modelo 3 | Estimación Logit – Lambda=0 | 0 | 0,5747561 | 16,7% | 17,81% |
| Modelo 4 | Estimación con Lasso-Upsampling | 0,0006887055 | 0,6892392 | 7,22% | 64,54% |
| Modelo 5 | Estimación con Ridge-Upsampling | 0,0138377 | 0,6891904 | 7,21% | 64,62% |
| Modelo 6 | Estimación con Lasso-Downsampling | 0,0009165909 | 0,689228 | 7,27% | 64,31% |
| Modelo 7 | Estimación con Ridge-Downsampling | 0,01393137 | 0,6888365 | 7,25% | 64,40% |

*Tabla 2. Modelos utilizados en el problema de clasificación.*

Como se observa en la tabla 2, el mejor modelo de predicción de la pobreza a partir del procedimiento de clasificación fue el Modelo 5 – Estimación Ridge con Upsampling de la muestra. Este es el mejor modelo ya que, en primer lugar, es el modelo con menor cantidad de falsos negativos (7,21%) lo cual se puede interpretar como un buen modelo predictivo de la pobreza al ser este porcentaje bajo. Así mismo, es el modelo con mayor Sensibilidad (64,62%), es decir, la proporción de hogares pobres que efectivamente fueron clasificados como pobres. Por último, es el modelo que también tiene mayor AUC (0,6891904) lo que se puede interpretar como el modelo con mayor poder predictivo.




*Gráfica 1. Coeficientes del modelo Ridge upsampling*




*Gráfica 2. Predicción pobreza*

La gráfica 2 nos muestra un histograma de las predicciones del modelo, en donde se predijo que el 35% de los hogares son Pobres. La gráfica 1 nos muestra la importancia de cada variable independiente en la predicción de la pobreza, determinándose que la variable explicativa que más impacto tiene en la predicción de la pobreza de los hogares es el estado de la vivienda, específicamente, los hogares que tienen vivienda a partir de un usufructo o una posesión sin título tienen una mayor probabilidad de ser pobres frente a los hogares que si tienen un hogar propio. Además, la gráfica nos muestra que el número de las personas en el hogar si es determinante para predecir la pobreza en los hogares, específicamente, entre mayor sea el número de personas en la unidad de gasto del hogar la probabilidad de que sea pobre aumenta significativamente.

### 3.2. Modelo de regresión estimación del ingreso

Se utilizaron 7 modelos para encontrar el más apropiado, en el Modelo 1 se utilizó la información por personas para poder predecir el ingreso de cada una de ellas (*Ingtot*), posteriormente en el Modelo 2 y 3 se utilizaron los hiperparámetros Ridge y Lasso para la estimación de estos. En el Modelo 4 se utilizó la información de hogares para la predicción de la variable *Ingtotugarr,* en los modelos 5 y 6 se utilizaron los hiperparámetros Ridge y Lasso para la estimación de estos.

Finamente, en el Modelo 7 se hizo un merge de la información de personas con las de hogar y por medio del hiperparámetro Lasso se realizó la estimación de este para poder predecir de la mejor manera el ingreso total de cada individuo. La razón por la que se utilizó Lasso con un lamba de 239 radicó en que se encontró una

mejora en cuanto al MSE comparándolo con Ridge o el modelo sin ningún hiperparámetro. Las variables explicativas para el ingreso fueron la edad, el máximo nivel educativo, el oficio, las horas trabajadas y una variable arriendo la cual consistió en sumar el valor del arriendo que las personas pagaban o una estimación de cuanto pagarían de arriendo en el caso que fueran propietarios. Consideramos que estas variables eran apropiadas para poder identificar el ingreso de una persona para posteriormente entrenarlo con la base de datos de *train personas*.

| Modelo | | MSE | Lambda óptimo |
|---|---|---|---|
| Modelo 1 | $Ingtot = Edad + Edad^2 + MaxNivEdu + Oficio + HorasTrabajadas$ | 1.058E+12 | No aplica |
| Modelo 2 | Mismo modelo 1 estimado con Ridge | 1.060E+12 | 49197 |
| Modelo 3 | Mismo modelo 1 estimado con Lasso | 1.058E+12 | 239 |
| Modelo 4 | $Ingtotugarr = CuartosHogar + TipoDeViv + Arriendo + NperUG$ | 5.888E+12 | No aplica |
| Modelo 5 | Mismo modelo 4 estimado con Ridge | 5.891E+12 | 91457 |
| Modelo 6 | Mismo modelo 5 estimado con Lasso | 5.890E+12 | 3443 |
| Modelo 7 (estimación con Lasso) | $Ingtot = Edad + Edad^2 + MaxNivEdu + Oficio + HorasTrabajadas + Arriendo$ | 1.058E+12 | 239 |

*Tabla 3. Modelos utilizados para la estimación del ingreso*

Con el modelo 7 escogido predecimos el ingreso para cada individuo en la base de datos *test personas,* seguidamente estos ingresos fueron sumados por hogar para finalmente compararlos con la línea de pobreza *Lp* de la base de datos *test hogar* y así definir si un hogar era pobre o no, convirtiendo esta comparación en un valor binario de 1 y 0. Para encontrar los Falsos negativos y Positivos dividimos la base de datos con que entrenamos nuestro modelo en un *train* y *test.* Se encontraron 42112 true negatives, 15118 true positives, 8569 false negative y 39118 false negatives. Las tasas de falsos encontradas fueron las siguientes:

$$False\ Positive\ Rate = \frac{FP}{TP + FP} = \frac{39118}{15158 + 39118} = 0.720 \qquad False\ Negative\ Rate = \frac{FN}{TN + FN} = \frac{8569}{42112 + 8569} = 0.169$$

Dado que estamos prediciendo pobreza, nuestro interés es que no sé predigan falsos negativos (hogares pobres que se predicen como no pobres). En este caso se obtiene una tasa de falsos negativos de 16,9%.

## 4. Conclusiones

Este ejercicio predictivo de la pobreza de los hogares colombianos nos muestra que la forma para predecir una variable puede ser a partir de diversos procedimientos, siendo de suma importancia desarrollar diversos modelos para compararlos entre sí y encontrar el que mejor prediga.

En la predicción de la pobreza a partir del problema de clasificación se encuentra que el remuestreo es una forma eficiente para que nuestros modelos hayan aumentado su poder predictivo y disminuido sustancialmente los falsos positivos a comparación de los modelos sin remuestreo. En este caso en específico, nuestra muestra presentaba un desequilibrio significativo en la variable “Pobre” por lo que esto podía estar impactando negativamente la efectividad de los modelos.

Nuestro análisis nos llevó a determinar que el mejor modelo de predicción para la pobreza de los hogares a partir de nuestra forma funcional es un ridge con upsampling, estando los demás modelos con remuestreo muy cerca del poder predictivo de este modelo. Algo a destacar de este modelo es que los falsos negativos encontrados fueron de 7.21%, una cifra más baja a comparación del 16.9% que se obtuvo a partir del mejor modelo utilizando la predicción del ingreso. Esto nos da indicios de las ventajas de utilizar modelos de clasificación para este tipo de predicciones. Finamente, en el caso del modelo de predicción se encuentra una ventaja al utilizar hiperparámetros, en específico de Lasso, donde se encuentra el menor MSE ya que la penalización de variables por medio del lama incide en una mejor estimación.