Daniel Casas Bautista

Código: **202120803**

Lucia Fillippo Aguillón

Código: **202213187**

Irina Andrea Vélez López

Código: **201119114**

Miguel Ángel Victoria Simbaqueva

Código: **202224043**

**Problem Set 3**

Big Data & Machine Learning

**Fecha de entrega: 30 de julio de 2023**

**Resumen:** El presente informe presenta la solución al Problem Set 3 de la clase Big Data & Machine Learning, en donde se aplicaron diversos conceptos y herramientas para la limpieza de bases de datos y el desarrollo de un modelo de predicción de la pobreza en territorio colombiano.

En el repositorio GitHub que abajo se referencia contiene el presente informe junto a los scripts de RStudio donde se generaron todos los resultados. Este repositorio se encuentra en: <https://github.com/irivelez/PS3_Predicting_Poverty.git>

1. **Introducción**

El desarrollo de políticas públicas orientado a mejorar el bienestar de la población pobre y vulnerable es fundamental para para fomentar una sociedad equitativa que brinde igualdad de oportunidades a sus habitantes. En este orden de ideas, el diseño de las políticas públicas será óptimo en la medida en que se focalice de mejor manera la población objetivo, para que la distribución de los recursos disponibles llegue a aquellas personas que más lo necesitan.

Teniendo presente lo anterior, el objetivo del presente documento es predecir la situación de pobreza de los hogares colombianos, con el fin de que las políticas públicas orientadas a esta población sean correctamente dirigidas, evitando errores de inclusión y de exclusión. Para lograr esto, se utilizará un modelo de predicción de la pobreza de los hogares, a partir de datos obtenidos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad en el año 2018 por parte del Departamento Administrativo Nacional de Estadística – DANE. En el presente documento se utilizará un modelo predictivo que se basará en el siguiente modelo:

Donde la condición de pobreza existirá cuando el indicador *I* señale queel ingreso *Inc* sea menor a la línea de pobreza *Pl*. El modelo determinará la condición de pobreza por dos vías; en primer lugar, una estrategia de clasificación para predecir hogares pobres (1) y no pobres (0); en segundo lugar, se utilizarán regresiones para determinar el ingreso de los hogares y así determinar si se encuentra por debajo o por encima de la línea de pobreza. Dentro del ejercicio a realizar se tendrán que identificar ROC[[1]](#footnote-2), falsos positivos, falsos negativos y demás elementos para predecir de la manera más acertada los hogares que serían objeto de análisis de las políticas relacionadas con este problema.

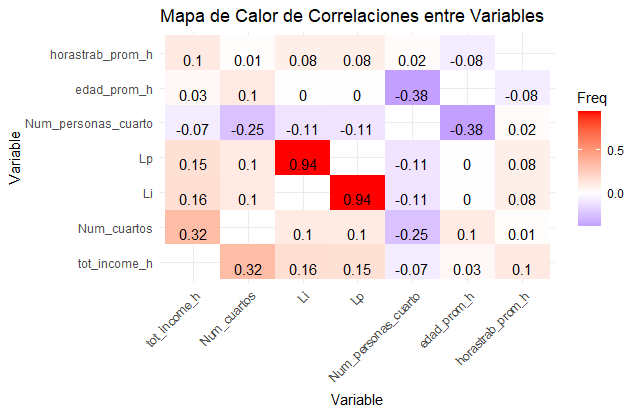
1. **Datos**

Como se mencionó previamente, para el desarrollo del Problem Set se utilizarán datos obtenidos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad en el año 2018 por parte del Departamento Administrativo Nacional de Estadística – DANE, la cual contiene información que permite realizar un análisis de la pobreza en Colombia, al contener datos de ingreso, de mercado laboral, sociodemográfica, entre otros. Por ejemplo, la variable de ingreso del hogar es una variable continua que cuenta con un total de 164.960 observaciones. El valor de ingreso en el primer cuartil es de $800.000 y el valor medio de ingreso es de $2.102.586, el cual es 1.6 veces el salario mínimo[[2]](#footnote-3) que aplicaba en dicho año. La línea de pobreza (Lp) refleja el límite de ingresos por debajo del cual un hogar es considerado pobre, señalando que el valor mínimo es de $167.222, el máximo de $303.8107 y la media de $271.605. El nivel de pobreza, teniendo en cuenta los anteriores datos de ingreso, reflejan un total de 33.024 personas en condición de pobreza monetaria en la muestra, mientras que hay 131.936 que no lo están. Para 2018, de acuerdo con el DANE, la línea nacional de pobreza monetaria fue de $257.433[[3]](#footnote-4).

1. **Estadísticas descriptivas**

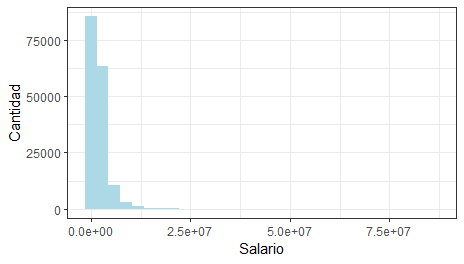
* General

En una revisión inicial de las correlaciones entre las variables independientes, se puede identificar que existe una relación positiva entre el número de cuartos del hogar y el ingreso total, ya que a mayor ingreso existe la posibilidad de tener hogares más amplios; además, otra relación importante a destacar es la relación negativa que hay entre la edad promedio del hogar y el número de cuartos, lo que quiere decir que a mayor edad promedio, la cantidad de cuartos en el hogar se reduce, por lo que los hogares con mayor edad suelen vivir en viviendas más reducidas, esto puede estar también influenciado por varios factores, como el ingreso y los gastos del hogar.



* Ingreso

Los ingresos por hogar están concentrados en valores bajos, esto se puede evidenciar en la gráfica de distribución de ingresos, lo cual concuerda con la distribución de ingresos en Colombia, en donde la mayoría de la población es de ingreso bajo y medio. En promedio, los ingresos por hogar están en 2.000.000 de pesos, el mínimo ingreso en un hogar en de 0, siendo estos los casos de pobreza, y el nivel máximo de ingresos por hogar es de 85.000.000 de pesos, los cuales son los valores que sesgan la distribución de ingresos.



* Pobreza

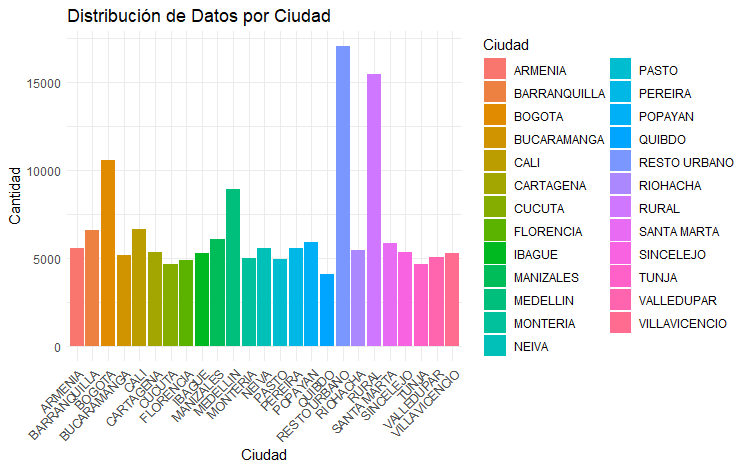
Se puede identificar en el siguiente gráfico de torta, que en los datos obtenidos, el 20% de los hogares está por debajo de la línea de pobreza, mientras que el resto se encuentra sobre este valor. Cabe resaltar que esta distribución proviene de la información obtenida de la base de datos submission template, la cual es una aproximación a los datos reales.

Gráfico, Gráfico circular

Descripción generada automáticamente

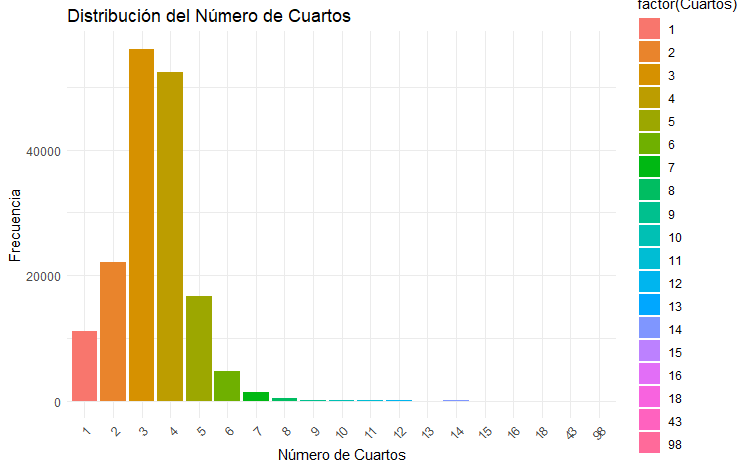
* Dominio

A continuación, se presenta la distribución de los datos de hogares por ciudad. Las columnas más altas corresponden al acumulado, es decir, resto urbano y rural. Se observa también que las grandes ciudades son las que presentan una mayor cantidad de datos en la muestra, estas son Bogotá y Medellín; el resto de las ciudades tiene una cantidad similar de observaciones (rondan las 5.000), esta distribución similar en los datos contribuye la reducción de sesgo en las estimaciones



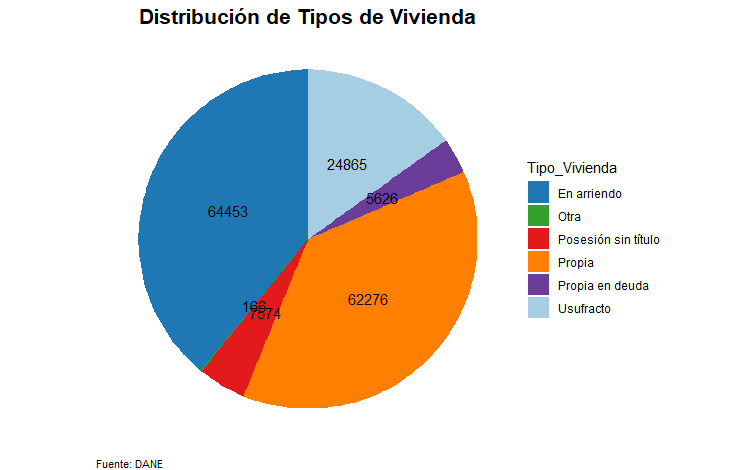
* Número de cuartos

La distribución del número de cuartos en el hogar es similar a la distribución de salario vista anteriormente, ya que, a mayor salario en el hogar existe la posibilidad de tener casas con espacios más amplios.



* Tipo de vivienda

Se puede observar que la mayor parte de la muestra vive en casa en arriendo, seguido de casa propia, lo cual se puede relacionar con que un porcentaje más bajo de la población cuenta con los recursos suficientes para tener casa propia. Llama la atención que una gran cantidad de hogares vive en usufructo.



* Número de personas por cuarto

En este caso, se identifica que una gran cantidad de hogares tiene en promedio 2 y 3 personas por cuarto. Esta variable es de interés ya que, mientras más personas vivan por cuarto la probabilidad de que sean hogares pobres es más alta, mientras que si solo vive una persona en el cuarto, la probabilidad de que sea un hogar pobre es menor.

Gráfico, Histograma

Descripción generada automáticamente

* Edad promedio del hogar

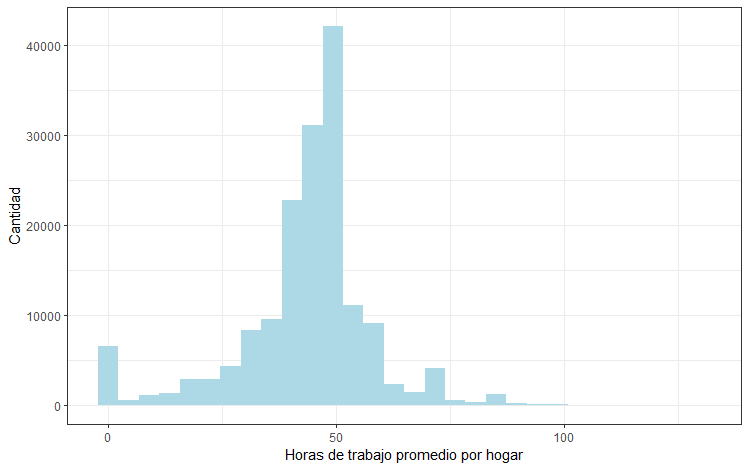
En la edad promedio por hogar, se puede observar que también se concentran los datos en valores bajos, a saber, la mayor cantidad de hogares tiene una edad promedio de 25 años, por lo que son hogares jóvenes. Mientras que existen pocos hogares en donde la edad promedio sea de más de 50 años. La edad puede ser un factor relevante a la hora de predecir la pobreza ya que, a mayor edad existe la posibilidad de tener más experiencia y por lo tanto más salario.

Gráfico, Histograma

Descripción generada automáticamente

* Horas trabajadas en promedio del hogar

Esta variable muestra que la mayoría de los hogares trabaja entre 47 y 48 horas, los que concuerda con lo establecido por la ley. La distribución de esta variable parece ser normal; sin embargo, se ve una cantidad alta de hogares en 0, estos pueden ser aquellos hogares en línea de pobreza, ya que no reciben ingresos por su trabajo.



* Máximo nivel educativo alcanzado en un hogar

A continuación, se presenta la gráfica de distribución del nivel de educación máximo alcanzado por algún miembro de un hogar. Allí se identifica que la mayoría de los hogares tiene algún individuo que terminó la universidad. Seguido de educación media y básica secundaria. Es de resaltar que casi 40.000 hogares en la muestra solo alcanzaron un nivel de educación de básica primaria y básica secundaria, lo cual puede influir directamente en si el hogar está por debajo de la línea de pobreza o no.

Gráfico, Gráfico de barras

Descripción generada automáticamente

* Afiliación seguridad social

Por último, se utilizó esta variable para identificar si alguna persona en el hogar está afiliada a seguridad social o no, esta variable nos puede brindar información que permita establecer si el hogar es pobre o no, en la media en que, si al menos una persona está afiliada puede significar que tiene ingresos suficientes para no estar debajo de la línea de pobreza. A saber. Se observa que el 3% de los hogares no cuenta con ningún individuo afiliado a seguridad social, esta es una distribución similar a la vista anteriormente de línea de pobreza.

Gráfico, Gráfico circular

Descripción generada automáticamente

1. **Modelo y resultados**
   1. **Clasificación**

Se utilizarán modelos de clasificación binarios para realizar la predicción de aquellos hogares que son pobres y aquellos que no lo son. Para ello se utilizarán al menos tres modelos con diferentes variables predictivas hasta alcanzar el mejor resultado; además, se utilizarán distintos métodos de predicción como Logit, Lasso (tomando como métrica la sensibilidad o el ROC y haciéndolo upsample o downsample) y Elastic Net.

Para hacer este ejercicio se ha dividido la muestra de entrenamiento en tres partes; la primera, es una mini muestra de training, la cual contiene el 70% de la base de datos principal (training de hogares) y ha sido utilizada para la estimación de los modelos; la segunda, que es de evaluación, la cual ha sido útil para desarrollar técnicas de post procesamiento, evaluando el punto de quiebre óptimo de los modelos; finalmente, el tercer modelo es de testeo, cuyo objetivo es el de predecir la pobreza a partir de los modelos estimados. Para cada modelo y especificación se presentarán los resultados, de acuerdo con las métricas de ROC, sensibilidad[[4]](#footnote-5), la especificidad[[5]](#footnote-6), precisión[[6]](#footnote-7) y el coeficiente kappa. Dado que no se trata de una muestra balanceada, la precisión no será la única variable importante por considerar.

* **Modelo 1**

Teniendo en cuenta lo anterior, el primero de los modelos elegidos es el siguiente:

Donde:

* : Es una variable dummy que es 1 si la persona es pobre y 0 en caso contrario
* : Es una variable categórica que señala el número de personas por unidad de gasto
* : Es una variable continua que refleja la línea de pobreza que aplica al hogar, de acuerdo con su sitio de residencia
* : Es una variable categórica que refleja el tipo de vivienda y toma los siguientes valores. (a: Propia, totalmente pagada; b: Propia, la están pagando; c: En arriendo o subarriendo; d: En usufructo; e: Posesión sin título; f: Otra)
* : Es una variable que refleja la ciudad donde reside la persona
* : Es una variable categórica que refleja el número de cuartos del hogar

Con este modelo, y teniendo presente los métodos de predicción a utilizar, se han obtenido los siguientes resultados:

Tabla 1. Resultados para el primer modelo predictivo

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Especificación** | **Alpha** | **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** | **Kappa** |
| **Logit** | N.A. | N.A. | 0.7733385 | 0.2024085 | 0.9686044 | 0.8148902 | 0.2285727 |
| **Lasso (Sensibilidad)** | 0 | 0.0094356945 | 0.7734394 | 0.1762496 | 0.9740862 | 0.8140242 | 0.2058859 |
| **Lasso (ROC)** | 0 | 0.0098838153 | 0.7734398 | 0.1753862 | 0.9742054 | 0.8139462 | 0.2050074 |
| **Lasso Upsample** | 0 | 0.0136766552 | 0.7735692 | 0.7016445 | 0.7067255 | 0.7041850 | 0.4083700 |
| **Lasso Downsample** | 0 | 0.0130565713 | 0.7718717 | 0.6991289 | 0.7059481 | 0.7025385 | 0.4050769 |
| **Elastic net** | 0.1 | 0.0192128642 | 0.7735979 | 0.13519821 | 0.9822547 | 0.8123181 | 0.1671635 |

A continuación se muestran las tablas de contingencia para identificar falsos positivos y falsos negativos en cada especificación:

Tabla 2. Tablas de contingencia para el primer modelo predictivo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Especificación** | **Predicción / Dato real** | | | |
| **SI/NO** | **SI/SI** | **NO/NO** | **NO/SI** |
| **Logit** | 5196 | 1335 | 25709 | 752 |
| **Lasso (Sensibilidad)** | 5372 | 1159 | 25844 | 617 |
| **Lasso (ROC)** | 5378 | 1153 | 25845 | 616 |
| **Lasso Upsample** | 1927 | 4604 | 18761 | 7700 |
| **Lasso Downsample** | 1936 | 4595 | 18801 | 7660 |
| **Elastic net** | 5659 | 872 | 26019 | 442 |

* **Modelo 2**

El segundo de los modelos elegidos es el siguiente:

Sumado a las variables antes explicadas, se incluyen:

* : Es una variable categórica que refleja el máximo nivel educativo alcanzado por el hogar, donde se encuentran las opciones de ninguno, preescolar, primaria, secundaria, media, superior, o no sabe.
* : Es una variable categórica que refleja el número de habitaciones en las que duermen las personas del hogar, donde se encuentran las opciones de ninguno, preescolar, primaria, secundaria, media, superior, o no sabe.
* : Es una variable continua que refleja el número de personas por habitación.

Con este modelo, y teniendo presente los métodos de predicción a utilizar, se han obtenido los siguientes resultados:

Tabla 3. Resultados para el segundo modelo predictivo

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Especificación** | **Alpha** | **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** | **Kappa** |
| **Logit** | N.A. | N.A. | 0.8080366 | 0.2842526 | 0.9599701 | 0.8244076 | 0.3084832 |
| **Lasso (Sensibilidad)** | 0 | 0.0124646013 | 0.8078311 | 0.254985565 | 0.9673044 | 0.8243990 | 0.288747289 |
| **Lasso (ROC)** | 0 | 0.0118994705 | 0.8078318 | 0.254985565 | 0.9673044 | 0.8243990 | 0.288747289 |
| **Lasso Upsample** | 0 | 0.0157192623 | 0.8082528 | 0.7252616 | 0.7381536 | 0.7317076 | 0.4634152 |
| **Lasso Downsample** | 0 | 0.0157192623 | 0.8068945 | 0.7258057 | 0.7349560 | 0.7303808 | 0.4607616 |
| **Elastic net** | 0.1 | 0.0025040422 | 0.8080862 | 0.2754898 | 0.9622343 | 0.8244596 | 0.3028093 |

A continuación se muestran las tablas de contingencia para identificar falsos positivos y falsos negativos en cada especificación:

Tabla 4. Tablas de contingencia para el segundo modelo predictivo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Especificación** | **Predicción / Dato real** | | | |
| **SI/NO** | **SI/SI** | **NO/NO** | **NO/SI** |
| **Logit** | 4677 | 1854 | 25517 | 944 |
| **Lasso (Sensibilidad)** | 4871 | 1660 | 25683 | 778 |
| **Lasso (ROC)** | 4871 | 1660 | 25683 | 778 |
| **Lasso Upsample** | 1783 | 4748 | 19688 | 6773 |
| **Lasso Downsample** | 1771 | 4760 | 19646 | 6815 |
| **Elastic net** | 4733 | 1798 | 25568 | 893 |

* **Modelo 3**

El segundo de los modelos elegidos es el siguiente:

Sumado a las variables antes explicadas, se incluyen:

* : Es una variable dummy que toma el valor de 1 si el hogar está afiliado a salud
* : Es una variable continua que muestra la edad promedio del hogar
* : Es una variable continua que muestra la edad promedio del hogar

Con este modelo, y teniendo presente los métodos de predicción a utilizar, se han obtenido los siguientes resultados:

Tabla 3. Resultados para el segundo modelo predictivo

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Especificación** | **Alpha** | **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** | **Kappa** |
| **Logit** | N.A. | N.A. | 0.8632603 | 0.4629197 | 0.95516 | 0.8564067 | 0.4828833 |
| **Lasso (Sensibilidad)** | 0 | 0.0118994705 | 0.8634538 | 0.424544538 | 0.9631227 | 0.8550731 | 0.461762106 |
| **Lasso (ROC)** | 0 | 0.0124646013 | 0.8634541 | 0.424415033 | 0.9631443 | 0.8550644 | 0.461678243 |
| **Lasso Upsample** | 0 | 0.0157192623 | 0.8648494 | 0.7789851 | 0.7903712 | 0.7846781 | 0.5693563 |
| **Lasso Downsample** | 0 | 0.0157192623 | 0.8645168 | 0.7785116 | 0.7879647 | 0.7832381 | 0.5664763 |
| **Elastic net** | 0.1 | 0.0002504042 | 0.8632949 | 0.4599844 | 0.9558750 | 0.8563894 | 0.4815270 |

A continuación se muestran las tablas de contingencia para identificar falsos positivos y falsos negativos en cada especificación:

Tabla 4. Tablas de contingencia para el segundo modelo predictivo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Especificación** | **Predicción / Dato real** | | | |
| **SI/NO** | **SI/SI** | **NO/NO** | **NO/SI** |
| **Logit** | 3476 | 3055 | 25374 | 1087 |
| **Lasso (Sensibilidad)** | 3704 | 2827 | 25592 | 869 |
| **Lasso (ROC)** | 3704 | 2827 | 25592 | 869 |
| **Lasso Upsample** | 1415 | 5116 | 21036 | 5425 |
| **Lasso Downsample** | 1414 | 5117 | 21033 | 5428 |
| **Elastic net** | 3532 | 2999 | 25433 | 1028 |

1. **Modelo de predicción de ingreso y resultados**

Para poder tener una predicción más precisa de la variable pobre, se procedió a realizar un modelo de predicción del ingreso para la base de datos test, utilizando la información de ingresos tomada de la base de datos train\_hogares. Para esto, se realizaron distintos modelos y se eligió el que tuviera un mejor desempeño.

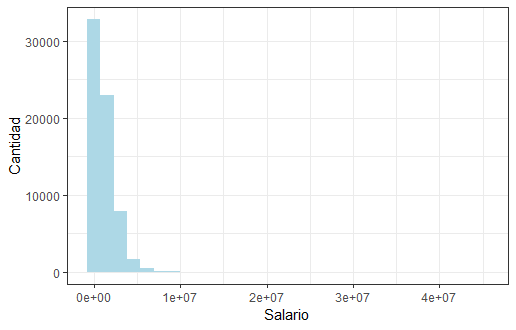
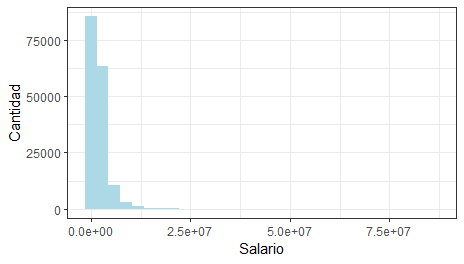
El modelo elegido fue el Elastic Net.

A continuación, se presentan las estadísticas descriptivas de la variable ingreso para la base de *train* y la base *test*.

|  |  |  |
| --- | --- | --- |
| Medida | Train | Test |
| Min. | 0 | 21,850 |
| 1st Qu. | 800,000 | 380,434 |
| Median | 1,400,000 | 785,012 |
| Mean | 2,102,586 | 1,242,944 |
| 3rd Qu. | 2,518,242 | 1,762,270 |
| Max. | 85,833,333 | 44,872,887 |

Se puede observar que, si bien existe una diferencia en todas las medidas de ambas bases de datos. Su comportamiento y distribución es similar; en ambos casos la mayor cantidad de observaciones se encuentran en niveles de ingreso bajo, también existen datos altos de ingresos que sesgan la muestra, lo cual se puede observar en las siguientes gráficas.

Distribución de salario (*train*) Distribución de salario (*test*)



1. **Modelo de clasificación final y resultados**

Para realizar el modelo final, se añadieron las variables:

* tot\_income\_h: Variable resultado de la predicción anterior.
* Num\_personas\_cuarto: Variable que representa cuantas personas viven en un cuarto en cada hogar.
* edad\_prom\_h: Variable continua que mide la edad promedio de un hogar.
* horastrab\_prom\_h; Variable continua que mide cuantas horas se trabaja en promedio en el hogar en una semana.
* max\_educ\_h: Variable categórica que mide cual es el nivel educativo máximo alcanzado por un individuo en un hogar.
* max\_health\_h; Variable dummy que mide si en el hogar hay al menos una persona afiliada a seguidad social.

El modelo es el siguiente:

1. **Conclusiones y recomendaciones**

* Las ponemos al final

1. La curva ROC (Receiver Operating Characteristic) y el área AUC (Area Under the Curve) son dos métricas comúnmente utilizadas para evaluar y comparar la calidad del rendimiento de un modelo de clasificación [↑](#footnote-ref-2)
2. De acuerdo con los valores de salario mínimo y subsidio de transporte aplicable a la fecha. Fuente: <https://www.portafolio.co/economia/empleo/salario-minimo-colombia-2017-109538> [↑](#footnote-ref-3)
3. Fuente: [Boletín técnico Pobreza Monetaria en Colombia 2018 (dane.gov.co)](https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf) [↑](#footnote-ref-4)
4. La cual hace referencia a la capacidad de detectar verdaderos positivos. [↑](#footnote-ref-5)
5. La cual hace referencia a la capacidad de detectar verdaderos negativos. [↑](#footnote-ref-6)
6. La cual hace referencia a la proporción de predicciones correctas. [↑](#footnote-ref-7)