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.

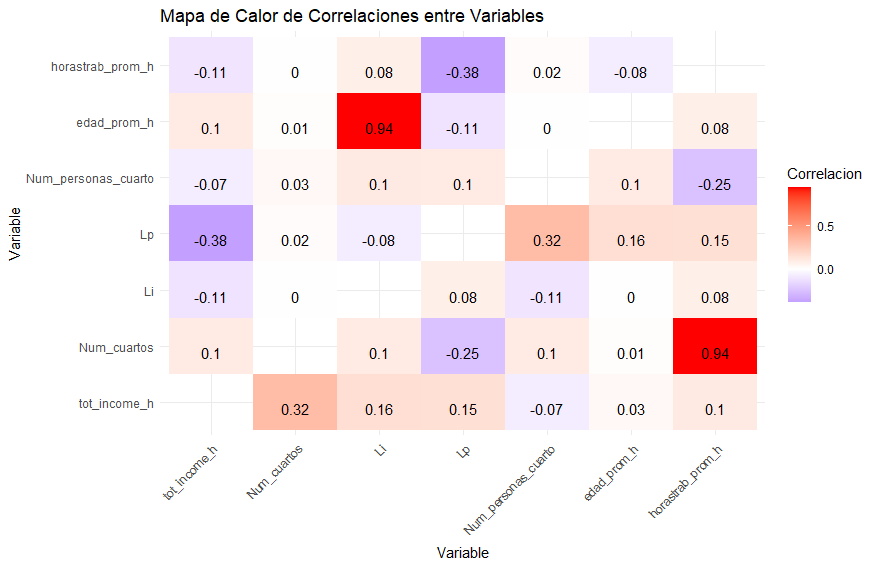
1. **Datos**

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

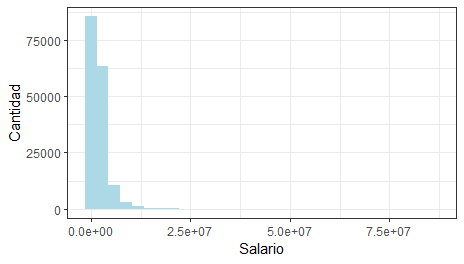
Fdsfdsfds



* Ingreso

Dsadadsad

|  |  |
| --- | --- |
| Medida | Valor |
| Min. | $0 |
| 1st Qu. | $800,000 |
| Median | $1,400,000 |
| Mean | $2,102,586 |
| 3rd Qu. | $2,518,242 |
| Max. | $85,833,333 |

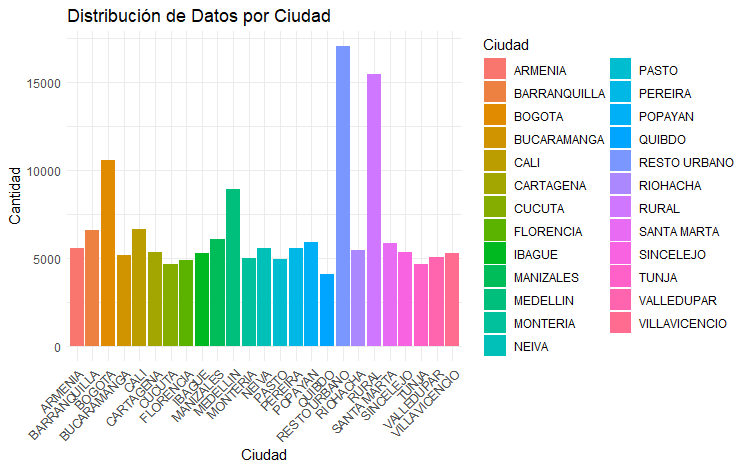


* Pobreza

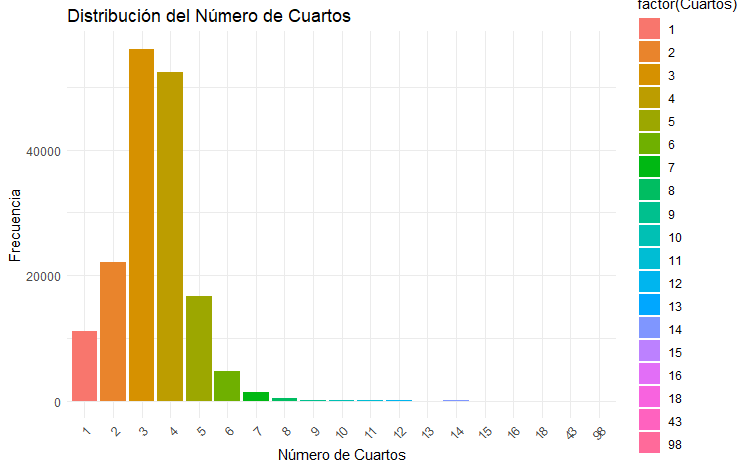
Gráfico, Gráfico circular

Descripción generada automáticamente

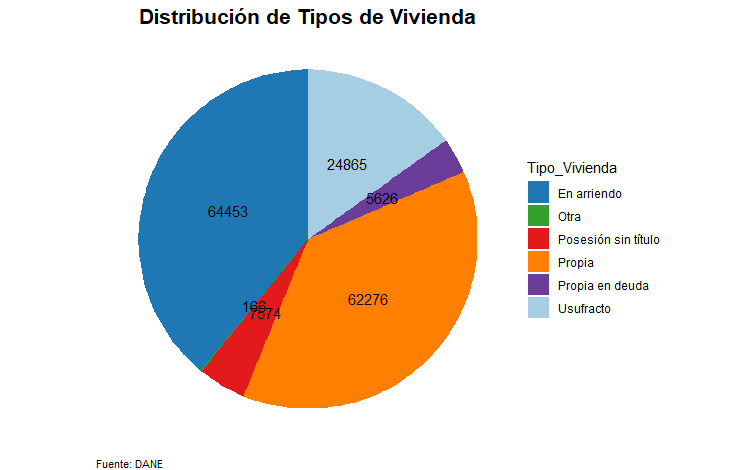
* Dominio



* Número de cuartos



* Tipo de vivienda



* Línea de indigencia

Gráfico, Histograma

Descripción generada automáticamente

* Línea de pobreza

Gráfico, Histograma

Descripción generada automáticamente

* Número de personas por cuarto

Gráfico, Histograma

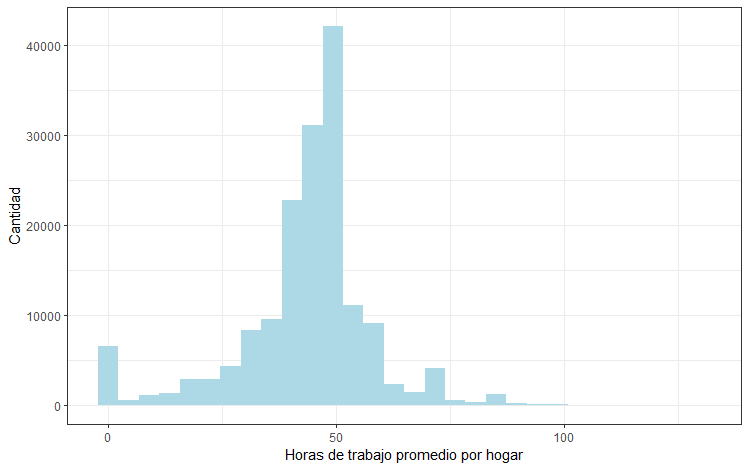
Descripción generada automáticamente

* Edad promedio del hogar

Gráfico, Histograma

Descripción generada automáticamente

* Horas trabajadas en promedio del hogar



* Máximo nivel educativo alcanzado en un hogar

Gráfico, Gráfico de barras

Descripción generada automáticamente

* Afiliación seguridad social

Gráfico, Gráfico circular

Descripción generada automáticamente

1. **Modelo y resultados**

* **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.

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **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 | 0.7736501 | 0.7003770 | 0.7084263 | 0.7044017 | 0.4088033 |
| **Lasso Downsample** | 0 | 0.0130565713 | 0.7718717 | 0.6991289 | 0.7059481 | 0.7025385 | 0.4050769 |

A continuación se muestran las tablas de contingencia para identificar falsos positivos y falsos negativos de los cinco modelos utilizados

logit

Pobre No Si

Si 5196 1335

No 25709 752

lasso\_SENS

Pobre No Si

Si 5378 1153

No 25845 616

lasso\_ROC

Pobre No Si

Si 1810 4721

No 18359 8102

lasso\_upsample

Pobre No Si

Si 1929 4602

No 18780 7681

lasso\_downsample

Pobre No Si

Si 1936 4595

No 18801 7660

Entiendo que podríamos hacer esto mismo para otros modelos (cambiando los predictores que se eligieron)… también podríamos incluir Elastic net, etc)

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)