**Introducción**

El propósito de este trabajo es construir un modelo predictivo de la pobreza en los hogares colombianos en base a la GEIH de 2018, para ello, se emplearon dos formas de predecir la pobreza, el primero a nivel de clasificación, pobre o no pobre mediante un modelo logit, y el segundo utilizando la predicción de los ingresos de los hogares y el umbral de pobreza, se determinó la clasificación de pobre y no pobre.

La motivación de este taller se basa en el concurso realizado por el Banco mundial “Pover-T Tests: Predicting Poverty”[[1]](#footnote-1), en el cual tuvo 2,307 participantes que buscaron obtener el mejor modelo para medir la pobreza, lo anterior, dado que el objetivo del Banco mundial es acabar la pobreza extrema para 2030, y por ello se requiere como insumo medir la pobreza.

Respecto a los antecedentes del propósito de este taller, se tiene inicialmente el documento del DANE “Medición de Pobreza Monetaria y Desigualdad 2018”[[2]](#footnote-2), en el cual XXXXXXXX.Exponer brevemente el problema y si hay antecedentes

Los datos de este taller fueron suministrados en base a la GEIH de 2018, específicamente de la sección “Medición de Pobreza Monetaria y Desigualdad 2018”, en donde se tienen datos a nivel hogar y a nivel individual. Dado que el enfoque es a nivel hogar, se realizaron varias transformaciones de los datos individuales para obtener la mayor cantidad de variables posibles a nivel hogar, con el fin de utilizar metodologías de machine learning y obtener un buen modelo de predicción y clasificación.

Poner un avance de los resultados y las principales conclusiones

**Datos**

Inicialmente fueron suministradas 4 bases de datos, una de entrenamiento a nivel hogar y una a nivel individual, y una de pruebas a nivel hogar y una a nivel individual. Dichas bases de datos tienen las siguientes dimensiones:

Tabla 1. Cantidad de observaciones y variables bases de datos suministradas

|  |  |  |  |
| --- | --- | --- | --- |
| **Propósito** | **Nivel** | **Núm. Observaciones** | **Núm. Variables** |
| Prueba | Hogar | 66,168 | 16 |
| Individual | 219,169 | 63 |
| Entrenamiento | Hogar | 164,960 | 23 |
| Individual | 543,584 | 135 |

**Base de datos prueba y entrenamiento hogares**

Inicialmente se decide eliminar cuatro variables, la primera contiene el código del departamento, dado que en otra variable se tiene esta información, la segunda “p5100” - Cuánto pagan mensualmente por cuota de amortización-, la tercera “p5130” - Si tuviera que pagar arriendo por

esta vivienda, ¿cuánto estima que tendría que pagar mensualmente? - y la cuarta “p5140” -¿ Cuánto pagan mensualmente por arriendo ? -. Se eliminaron las ultimas 3 variables dado que contienen entre un 38% y 97% de missing values, y por otra parte se considera mas importante reconocer quien paga arriendo en lugar de cuanto paga de arriendo.

Base de datos prueba y entrenamiento individuos

Se eliminaron la mayoría de las variables que contienen entre un 18% y 99% de missing values. Sin embargo, se tuvieron los siguientes tratamientos particulares para:

* Las variables Des - Desocupado -, Ina - Inactivo -, Oc – Ocupados -, y Pet – Población en edad de trabajar –, que toman valor de 1 si el individuo hace parte de dichas categorías, se ve necesario transformar los missing values a la categoría 0, lo cual corresponde a que no hacen parte de esas categorías, permitiendo el uso de dichas variables en posteriores cálculos.
* La variable “p7495” - ¿recibió pagos por concepto de arriendos y/o pensiones? – se imputa eliminando las observaciones que contienen missing values, dicha imputación también se realizó a la variable “p7505” - Durante los últimos doce meses, ¿recibió dinero de otros hogares, personas o instituciones no gubernamentales; dinero por intereses, dividendos, utilidades o por cesantías? -.
* Las variables “p6090” y “p6210”, asociadas a las preguntas ¿Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud? Y ¿Cuál es el nivel educativo más alto alcanzado por .... y el último año o grado aprobado en este nivel?, respectivamente, al contener muy pocos missing values, se imputaron a la categoría no sabe no responde.
* La variable “P6210s1”, que contiene el grado de escolaridad de los individuos, se decide eliminar dado que tiene muchos missing values y la información que nos puede aportar esta variable se encuentra en la variable “P6210”, no con el detalle específico del grado de escolaridad, pero si con la categoría que corresponde.

Unificación bases de datos hogares – individuos

Respecto a las bases de datos de entrenamiento, inicialmente se validó para las variables “p6090” y “p6210” contenían una cantidad relevante de observaciones con bajo la categoría 9 - no sabe no responde -, en dicha validación se encuentra que existen muy pocas observaciones con esta categoría, por lo cual se procede a unir las bases de datos de hogares con la de individuos, tomando como referencia el id de cada hogar. Es importante mencionar que, en la base de datos de individuos, se tomó como referencia las observaciones del jefe de hogar como representación del hogar, y así utilizar esta información para unirla con la base de datos de hogar. Finalmente se eliminaron las variables repetidas en ambas bases de datos “Clase” y “Dominio”.

Respecto a las bases de datos de prueba, se realizó el mismo procedimiento realizado para las bases de datos de entrenamiento.

En este orden de ideas, a continuación, se muestra cuadro con la cantidad de observaciones y variables resultado de la limpieza realizada a las bases de datos iniciales.

Tabla 2. Cantidad de observaciones y variables bases de datos limpias.

|  |  |  |
| --- | --- | --- |
| **Propósito** | **Núm. Observaciones** | **Núm. Variables** |
| Prueba | 66,168 | 24 |
| Entrenamiento | 164,960 | 32 |

A continuación, se presentan tablas con las estadísticas descriptivas de las bases de datos después de su limpieza

Tabla 3. Estadísticas descriptivas - Base de datos prueba

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Estadísticas descriptivas - Base de datos prueba** | | | | |
| Urbano |  |  |  |  |
|  | Si |  | 59,368 | 90% |
|  | No |  | 6,800 | 10% |
| Tipo vivienda | |  |  |  |
|  | Propia - pagada |  | 25,235 | 38% |
|  | Propia - están pagando |  | 2,148 | 3% |
|  | Arriendo - Subarriendo |  | 25,310 | 38% |
|  | Usufructo |  | 10,259 | 16% |
|  | Posesión sin título |  | 3,178 | 5% |
|  | Otra |  | 38 | 0% |
| Afiliado a salud | |  |  |  |
|  | Si |  | 62,479 | 94% |
|  | No |  | 3,665 | 6% |
|  | No sabe |  | 23 | 0% |
| Recibe pagos arriendo/pensión | |  |  |  |
|  | Si |  | 11,257 | 17% |
|  | No |  | 54,910 | 83% |
| Recibe otras ayudas | |  |  |  |
|  | Si |  | 18,594 | 28% |
|  | No |  | 47,573 | 72% |
| Ocupado | |  |  |  |
|  | Si |  | 46945 | 71% |
|  | No |  | 19222 | 29% |
| Desocupado | |  |  |  |
|  | Si |  | 3095 | 5% |
|  | No |  | 63072 | 95% |
| Inactivo |  |  |  |  |
|  | Si |  | 16127 | 24% |
|  | No |  | 50040 | 76% |
|  |  | Media | Min | Max |
| Línea de indigencia | | 120,142 | 99,545 | 131,126 |
| Línea de pobreza | | 270,779 | 167,222 | 303,817 |

Tabla 4. Estadísticas descriptivas - Base de datos entrenamiento

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Estadísticas descriptivas - Base de datos entrenamiento** | | | | |
| Urbano |  |  |  |  |
|  | Si |  | 149,606 | 91% |
|  | No |  | 15,354 | 9% |
| Tipo vivienda | |  |  |  |
|  | Propia - pagada |  | 62,120 | 38% |
|  | Propia - están pagando |  | 5,616 | 3% |
|  | Arriendo - Subarriendo |  | 64,344 | 39% |
|  | Usufructo |  | 25,000 | 15% |
|  | Posesión sin título |  | 7,717 | 5% |
|  | Otra |  | 163 | 0% |
| Afiliado a salud | |  |  |  |
|  | Si |  | 155,089 | 94% |
|  | No |  | 9,789 | 6% |
|  | No sabe |  | 81 | 0% |
| Recibe pagos arriendo/pensión | |  |  |  |
|  | Si |  | 29,154 | 18% |
|  | No |  | 135,805 | 82% |
| Recibe otras ayudas | |  |  |  |
|  | Si |  | 47,103 | 29% |
|  | No |  | 117,856 | 71% |
| Ocupado | |  |  |  |
|  | Si |  | 117234 | 71% |
|  | No |  | 47725 | 29% |
| Desocupado | |  |  |  |
|  | Si |  | 7645 | 5% |
|  | No |  | 157314 | 95% |
| Inactivo |  |  |  |  |
|  | Si |  | 40080 | 24% |
|  | No |  | 124879 | 76% |
| Pobre |  |  |  |  |
|  | Si |  | 33024 | 20% |
|  | No |  | 131936 | 80% |
| Indigente | |  |  |  |
|  | Si |  | 8023 | 5% |
|  | No |  | 156937 | 95% |
|  |  | Media | Min | Max |
| Línea de indigencia | | 120,416 | 99,545 | 131,126 |
| Línea de pobreza | | 271,605 | 167,222 | 303,817 |
| Ingreso Total antes de arriendos | | 2,089,017 | 0 | 85,833,333 |
| Ingreso Total con arriendos | | 2,305,640 | 0 | 88,833,333 |

Respecto a las tablas anteriores, se resalta que todas las variables categóricas conservan prácticamente las mismas proporciones, de igual manera las líneas de indigencia y pobreza son casi idénticas.

Adicionalmente estas tablas nos permiten observar que en Colombia la mayoría de personas viven en una zona urbana, solo el 38% tienen casa propia pagada en su totalidad, el 94% de los hogares cuentan con afiliación al servicio de salud, cerca del 18% de los hogares reciben ingresos por pensión y/o arriendos, cerca del 29% de los hogares reciben ayudas económicas de otros hogares, personas, instituciones no gubernamentales, intereses por cesantías y/o dividendos-utilizades por cesantías, También llama notoriamente la atención los ingresos por hogares, donde hay un máximo de aproximadamente 85 millones de pesos, mientras se presentan mínimos de cero pesos, para la variable de ingreso total antes de arriendos, la mediana es de $1,400,000, siendo este inferior a la media, así mismo, la mediana del ingreso con imputación de arriendo es de $1,582,735, siendo inferior a la media.

Pendiente grafica y parla

**Modelos y resultados**

**Modelos de Clasificación**

*## presentar los modelos que se usan y argumentar porqué se escogen y porqué pueden ser los mejores. Los modelos escogidos deben tener el mejor desempeño en términos de F2 score, pues los falsos negativos son más importantes, dado que debemos intentar no dejar pobres clasificados como no pobres, pues estaríamos sub estimando el número de pobres.*

*Confusion Matrix and Statistics*

*Reference*

*Prediction No Si*

*No 80853 15888*

*Si 11529 7201*

*Accuracy : 0.7626*

*95% CI : (0.7601, 0.765)*

*No Information Rate : 0.8*

*P-Value [Acc > NIR] : 1*

*Kappa : 0.2013*

*Mcnemar's Test P-Value : <2e-16*

*Sensitivity : 0.31188*

*Specificity : 0.87520*

*Pos Pred Value : 0.38446*

*Neg Pred Value : 0.83577*

*Prevalence : 0.19995*

*Detection Rate : 0.06236*

*Detection Prevalence : 0.16221*

*Balanced Accuracy : 0.59354*

*'Positive' Class : Si*

*# explicar el modelo que se escoge*

*# comparación de los modelos usando ROC, AUC, fn y fp*

*# describir las variables usadas en el modelo y una medida de su importancia relativa*

*# describir abordajes de sub-sampling para los desbalances de clase*

**Modelos de regresión de ingresos**

*# explicación detallada del modelo escogido*

*# comparación de 5 modelos en términos de MSE*

*# Convertir predictores a binarios para mostrar desempeño en términos de ROC ..*

*# describir las variables usadas en el modelo y una medida de su importancia relativa*

1. <https://www.drivendata.org/competitions/50/worldbank-poverty-prediction/page/97/> [↑](#footnote-ref-1)
2. <https://microdatos.dane.gov.co/index.php/catalog/608/datafile/F1#page=F2&tab=data-dictionary> [↑](#footnote-ref-2)