

# Problem Set 2 – GRUPO 6

# Víctor Dulio Chique

# Víctor Iván Sánchez

# Natalia Castro

# PREDICCIÓN DE PROBREZA EN COLOMBIA

1. Introducción:

La medición de la pobreza busca identificar los grupos de la población más vulnerables para realizar intervenciones de política pública que permitan mejorar sus carencias. Si bien en la actualidad el concepto de pobreza no aborda solamente el ingreso y se busca entender también las carencias sociales de los hogares, la medida de pobreza monetaria continúa siendo muy utilizada ya que es un indicador de un mínimo de subsistencia alimentaria y no alimentaria.

El ingreso usualmente se mide a través de encuestas las cuales requieren de un gran esfuerzo logístico, de diseño y de implementación. De igual manera exige un trabajo de medición de canastas alimentarias y no alimentarias que no son fáciles de generalizar pues cada territorio y cada familia tienen costumbres propias. Adicionalmente, las familias pueden esconder algunos de sus ingresos si saben que sus respuestas favorecerán subsidios o ser incluidos en programas sociales. Las variables que deben ser incluidas en los estudios de medición de pobreza monetaria son entonces numerosas y utilizarlas todas en un modelo puede resultar ser una tarea muy compleja.

El objetivo de este trabajo es precisamente buscar modelos y variables diferentes al ingreso que logren clasificar correctamente a los hogares colombianos entre aquellos que se encuentran en situación de pobreza o no, con el fin de tener herramientas de clasificación más precisas, menos complejas y más efectivas.

OJO FALTA: QUÉ SE OBTUVO DE LOS MODELOS FINALES.

2. Base de datos:

La base de datos que se utilizó en este trabajo fue la de la Medición de Pobreza Monetaria “Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE del Departamento Administrativo Nacional de Estadística – DANE. Los datos se encuentran divididos en dos grandes muestras: la de entrenamiento y la de testeo. A su vez cada una se divide en dos grupos: hogares y personas. Esta base de datos es conveniente porque provee información no sólo sobre los diferentes tipos de ingreso de los hogares sino sobre las características de contexto propias de cada hogar. Lo anterior permite proponer modelos con variables diferentes al ingreso total de cada hogar y seleccionar aquellos que estiman las predicciones con las mejores métricas de clasificación.

La base de datos de entrenamiento tenía 117.156 observaciones y la de testeo 68.168. Como primer paso se unieron las bases de hogares y personas para entrenamiento y posteriormente se hizo lo mismo para las bases de testeo. Las variables que se utilizaron en los modelos fueron:

|  |  |  |
| --- | --- | --- |
| Cabecera\_resto | Hombre | Horas de trabajo |
| Ingreso | Edad | Más tiempo de trabajo |
| Cuartos\_hogar | Entidad\_salud | Tamaño de la empresa |
| Vivienda\_propia | Nivel Educativo |  |
| Personas\_hogar | Tiempo de trabajo |  |

Cuadro 1: Variables seleccionadas 1

Debido a que existían datos faltantes se imputaron las variables: *entidad\_salud, mas\_trabajo, tipo\_de\_trabajo, tiempo\_trabajando, horas\_trabajo*, tamaño de la empresa con la media y moda de cada una de estas.

3. Estadísticas Descriptivas

A continuación se realiza una breve descripción de las variables que se incluyeron en los modelos de clasificación y predicción. Comenzamos con la variable objetivo del estudio: *Pobre*. El cuadro 2 muestra que la muestra tiene un desbalance importante que se puede clasificar como moderado (1%-20%). El problema con la prevalencia del resultado *pobre* es que los algoritmos tenderán a categorizar las predicciones dentro de esta clase mayoritaria. De esta manera la medida de *Accuracy* (proporción de predicciones correctas dentro del total de predicciones) tenderá a ser muy alta pero el modelo no necesariamente predecirá con exactitud a aquellos que no lo son. Esto es un problema porque si solamente se toman en cuenta estas predicciones para realizar intervenciones de política pública, seguramente se estaría beneficiando a personas para las cuales no están destinados esos recursos.

|  |  |
| --- | --- |
| Pobre | No pobre |
| 81.88% | 18.11% |

Cuadro 2: Porcentaje personas pobreza 1

|  |
| --- |
| Chart, bar chart  Description automatically generated |

Gráfica 1: Personas en situación de pobreza

A continuación se presentan varios gráficos con la proporción de las variables categóricas que se utilizaron en los modelos. Estas muestran una población en la que más del 60% son hombres, más del 85% tiene acceso a salud, además de las horas de trabajo 92% quiere trabajar más, la mayoría son trabajadores por cuenta propia o son empleados u obreros de empresas particulares, el 52% sólo cuenta con educación primaria y/o secundaria, cerca del 50% vive en arriendo y en su mayoría son hogares de cabeceras de Colombia.

|  |  |
| --- | --- |
| Gráfica 2: Distribución por género | Chart  Description automatically generated  Gráfica 3: Personas con servicio salud 1 |

|  |  |
| --- | --- |
| Gráfica 4: Quiere trabajar más horas? | Gráfica 5: Nivel Educativo |
| Gráfica 6: Tipo de trabajo | Gráfica 7: Tipo de vivienda |

|  |
| --- |
| Gráfica 9: Cabecera o resto 1 |

En cuanto a las variables continuas se observa una muy baja correlación (aunque significativa) entre ingreso, edad y tiempo trabajado. Inclusive la correlación es negativa entre los años que lleva trabajando una persona en su empresa y el ingreso para quienes se encuentran en situación de pobreza. Lo mismo sucede entre la variable edad y el ingreso.

Chart, scatter chart

Description automatically generated

MODELOS:

1. Modelo de clasificación:

Chart, line chart

Description automatically generated

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart, bar chart  Description automatically generated |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logit |  |  |
| False Positive Rate | 0.03171 |  |  |
| True Positive Rate | 0.2027 |  |  |

MODELOS DE PREDICCION DE INGRESOS:

1. Modelo de regresión lineal para predicción de ingreso: se utilizó la variable ingreso como variable dependiente y las variables independientes mencionadas sin interacciones. No sorprende encontrar que educación superior es la variable más importante en la predicción de ingreso. Ser empleador y trabajar para el gobierno también aparecen como significativas.

Table

Description automatically generated

Chart, funnel chart

Description automatically generated

2. Ridge: Nuevamente prediciendo la variable ingreso:

Chart

Description automatically generated

A picture containing diagram

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| OLS | RIDGE | LASSO | EN |
| 2293922 | 2296688 | 2294358 | 2297956 |