**Problem Set 2: Predicting Poverty**

1. **Introducción**

Este estudio se basa en la competencia “Pover-T Tests: Predicting Poverty” realizada por el Banco Mundial, la cual tuvo como objetivo la predicción de la pobreza en Colombia. Este análisis partió de la hipótesis de que los altos costos y el tiempo generan dificultades para la medición de esta, lo que complejiza obtener resultados apropiados. Sin embargo, al desarrollar mejores modelos, se pueden realizar encuestas con menos preguntas y más específicas que midan de manera rápida y económica la efectividad de las nuevas políticas e intervenciones. Por esta razón, es importante tener mayor precisión para orientar estas intervenciones e iterar las políticas, maximizando su impacto y su rentabilidad.

Diseñar una política pública es fundamental para la construcción de una sociedad justa y en la que todos tengan oportunidades. Por ello, es importante la ejecución acertada de modelos que puedan predecir correctamente la población objetivo y que dicha política pueda ser aplicada para realizar intervenciones a las familias adecuadas.

En este documento se presentan dos modelos de predicción de la pobreza en los hogares colombianos para conocer con exactitud cuáles hogares son pobres y cuáles no, con el fin de que la política pueda ser aplicada a quienes se encuentran realmente en condición de pobreza, y así, puedan recibir las ayudas del Gobierno, planteadas en la política, evitando casos en donde algunos hogares no sean identificados como pobres y con ello, no reciban los subsidios respectivos. El primer modelo se ejecutará por medio de clasificación de hogares pobres, y el segundo, se realizará por medio de una regresión, en donde se toman los ingresos y se comparan con la línea de pobreza para definir si son pobres o no.

Así, a lo largo de este *Problem Set,* se realizará un modelo de predicción de cuáles hogares son los más pobres, basado en características individuales con datos adquiridos por la Gran Encuesta Integrada de Hogares (GEIH) de 2018 del Departamento Administrativo Nacional de Estadística (DANE), para conocer información sociodemográfica de la población y señalar casos de fraude que podrían conducir a la reducción de la brecha en el país e identificar a las familias vulnerables que podrían necesitar mayor asistencia por parte del Gobierno. Para ello, se aplicarán dos metodologías: Clasificación y Regresión. Además, se identificarán parámetros como ROC, 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.

Las principales conclusiones indican que las variables de sexo y edad, tipo de ocupación, nivel de educación, experiencia y estrato socioeconómico, tienen impacto significativo en el salario de los incluidos estudiados y son clave para predecir sus ingresos por hora.

**Nota:**

La base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: <https://github.com/Yilap/Taller2-BDML.git>

**Contexto**

En el mercado colombiano se presentan dos grandes fenómenos que afectan el recaudo de impuestos por parte de las entidades del Estado: i) la **evasión de impuestos** *(tax evasion)* que implica actos ilícitos por parte de los contribuyentes los cuales violan los deberes derivados de la relación jurídica tributaria - tales como presentar declaraciones verdaderas o mantener los libros comerciales regulares – y ii) la **elusión de impuestos** *(tax avoidance),* conocida como la práctica de actos mediante los cuales se influencian los canales de conexión para evitar la aplicación de ciertos gravámenes tributarios (Sentencia C-360 de 2016 Corte Constitucional de Colombia, 2016). La evasión de impuestos se genera en múltiples formas, entre las que se desatacan la omisión de ingresos, declaración de costos, deducciones y descuentos inexistentes, subvaloración de activos, mimetización ilegal de ingresos, entre otros. Estas acciones dificultan el ejercicio de las autoridades estatales para recaudar y hacer cumplir las obligaciones tributarias de los usuarios. Adicionalmente, se considera el fraude fiscal como delito, toda vez que el contribuyente tiene la intención de evadir el pago de sus obligaciones fiscales, como consecuencia de su comportamiento premeditado (Sentencia C-360 de 2016 Corte Constitucional de Colombia, 2016).

1. **Datos**
2. ***Descripción de las fuentes de datos***

Para el desarrollo de este *Problem Set* se utilizarán los datos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 del Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta contiene información sobre las condiciones de empleo de las personas (si trabajan, en qué trabajan, cuánto ganan, fuentes de ingresos, si tienen seguridad social en salud o si están buscando empleo), adicional a las características generales de la población como sexo, edad, estado civil y nivel educativo (DANE, 2018). La GEIH consolida información no solo a nivel nacional sino a nivel regional, departamental, cabecera y ciudades capitales.

La GEIH es una fuente de información pertinente para el análisis de la pobreza en Colombia, ya que, al condensar información sociodemográfica sobre los individuos, los niveles de ingresos laborales y no laborales y la estructura de la fuerza de trabajo en el territorio nacional (tasas de ocupación, desempleo e informalidad laboral), permite identificar las características más relevantes que influyen sobre la renta personal. Usando estos datos como insumo, es posible predecir diferencias significativas entre los ingresos de los hogares y poder determinar si se encuentran en situación de pobreza o no.

1. ***Modelo de Clasificación***

La pobreza puede estar dada por diferentes variables. Sin embargo, es fundamental contar con las variables relevantes para que este modelo sea robusto, pero no se incurran en gastos que entorpezcan la investigación. La variable Npersug (No. personas en la unidad de gasto) evidencia aquellas personas que dentro del hogar están dentro de la unidad de gasto. De acuerdo con el análisis, la moda de esta variable es 3, es decir, 3 personas por unidad de gasto es el valor más común entre unidades de gasto por familia. También, se evidencia que el rango va de 1 a 28 personas por UG. La línea de pobreza (Lp) establece el límite de ingresos por debajo del cual un hogar es considerado pobre. El valor mínimo es COP 167.222; el máximo es COP 303.8107; la media es COP 271.605; la moda es COP 281.549,3 (DANE, 2019).

De acuerdo con el DANE (2019) se evidencia que la línea de pobreza monetaria nacional fue de $257.433 pesos. La variable Dominio es una variable categórica que indica en donde vive el hogar. La variable categórica P5090 (OcViv) hace referencia al tipo de ocupación que tiene el hogar en la vivienda, es decir, arriendo, propia, entre otros. Por otra parte, la variable numérica P5000 hace referencia a la cantidad de habitaciones que cuenta la vivienda que tiene el hogar, evidenciando que el mínimo es 1 habitación, máximo 98 y la cantidad de habitación más común es 3.

1. ***Modelo de Regresión***

Para el modelo de regresión se utilizarán las variables del modelo test personas que se encuentren simultáneamente en la de train personas. En el literal d se presentan las variables elegidas y su correspondiente análisis. Por un lado, la variable Edad es una de las más importantes a ser considerada en la regresión del ingreso, incluyendo su variante cuadrática. Se tiene que la edad mínima es 0 años, la máxima 110 años y la media 33 años para la base train personas y para la base test.

La variable Dominio, tiene la misma descripción que en la sección anterior. Para la variable Sexo, que es categórico, que toma valores de 1 para hombre y 2 para mujer, donde se encuentra mayor cantidad de mujeres en la base train personas. La variable Educ hace referencia al nivel educativo máximo, que va desde 1 (Ninguno) hasta 7 (terciaria o posgrado) y 9 (No reporta). Se imputó a las variables NA el número 9. El dato que más se repite es 3, es decir, nivel de primaria incompleta. Para la variable Ocup se tiene que la mayoría de las personas no presenta ocupación.

1. ***Análisis descriptivo de los datos (estadísticas descriptivas)***

***xxxxxxxxxxxxx***

1. **Modelos y resultados**
2. ***Clasificación de modelos***
3. ***Modelo de regresión de los ingresos***
4. **Modelo Final**
5. **Conclusiones y recomendaciones**
6. **Bibliografía**

DANE., (2019). Boletín Técnico Pobreza Monetaria en Colombia Año 2018. Bogotá D.C., mayo3 de 2019. Tomado de: <https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf>