**Problem Set 1: Predicting Income**

1. **Introducción**

La sub-declaración de los ingresos por parte de los ciudadanos representa uno de los mayores retos del sistema fiscal, pues conocer el valor real de la renta individual es esencial para el cálculo los impuestos. De acuerdo con estudios realizados para Estados Unidos por el Servicio de Impuestos Internos (IRS), alrededor del 83,6 % de los impuestos se pagan de manera voluntaria y oportuna, con el reporte de ingresos más bajos que los reales como la principal causa de esta brecha. Ahora bien, resulta fundamental conocer el contexto colombiano, en donde el mercado laboral y el ingreso están permeados por variables sociales complejas. Así, a lo largo de este *Problem Set* se desarrolla un modelo de predicción de ingresos basado en características individuales usando datos de 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.

Las principales conclusiones indican que XXXX. Considerar estos resultados es crucial para la implementación de políticas fiscales efectivas y para mejorar la eficiencia del sistema de recaudo de impuestos.

**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/Repositorio_Taller1>

1. **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).

La evasión y elusión de impuestos ocasionan una pérdida para el gobierno de Colombia de entre COP $50 y $80 billones anuales. En 2021, según el Banco Interamericano de Desarrollo (BID), estos dos fenómenos representaron aproximadamente US$17 mil millones al año (o cerca de COP $68 mil millones) y, de acuerdo con Fedesarrollo, se pierden alrededor de 5,4 puntos porcentuales del PIB anual debido a estos factores. Sobre esto, es relevante mencionar que la mayor pérdida de recaudo se da por la evasión del impuesto de renta de las empresas, con un 3,4% del PIB; seguida de la evasión del IVA, que representa el 1,3% del PIB, y finalmente la evasión por concepto de impuesto de renta a personas, con cerca de un 0,7% del PIB (La República, 2022).

1. **Datos**

***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) 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 evasión y elusión de impuestos 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 declarados y los reales.

***Adquisición de los datos***

Para la obtención de los datos de la GEIH se utilizaron técnicas de web scraping. El conjunto de datos contiene todos los individuos muestreados en Bogotá y está disponible en el siguiente sitio web [https://ignaciomsarmiento.github.io/GEIH2018 sample/](https://ignaciomsarmiento.github.io/GEIH2018%20sample/). En este caso, la página web fuente de los datos divide la información en 10 “*chunks*” de datos. Para conocer la estructura de cada enlace fue necesario analizar el código HTML de la página web, logrando identificar su naturaleza dinámica. El hecho de ser dinámica plantea un reto para la extracción, pues el enlace original no es donde realmente reposa la información. Es necesario esperar a que cargue por completo para inspeccionarlo y encontrar enlace particular de la tabla a extraer, de lo contrario, no es posible detectarla mediante el código de R.

El web scraping se realizó usando el paquete *rvest*. En este caso, se identificó un patrón en los enlaces de cada data chunk, permitiendo importar la información por medio del loop a continuación y uniendo las 10 porciones de los datos en un solo data frame:

df\_list <- list()

for (i in 1:10) {

html\_i <- read\_html(paste0("https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_", i, ".html")) %>%

html\_table()

df\_i <- as.data.frame(html\_i)

df\_list[[i]] <- df\_i

}

GEIH <- do.call(rbind, df\_list)

***Descripción del proceso de limpieza de datos***

La base de datos extraída contiene un total de 32.177 observaciones de todos los individuos muestreados en la GEIH en Bogotá y 178 variables. Para la limpieza de datos, se eliminan las variables que no resultan útiles para el modelo de predicción. Además, el análisis se centra únicamente en las personas empleadas mayores de 18 años, por lo que la muestra se limita a personas que cumplen estas características.

De acuerdo con la literatura económica, algunas de las variables más relevantes para la predicción de la renta individual son: edad, género, años de educación, experiencia, tipo de ocupación. A continuación, se justifica de manera detallada la inclusión de cada una de las variables en el modelo:

Relab

Mincer: Esta fórmula sugiere que el salario de un individuo depende de su nivel educativo y su experiencia laboral, teniendo en cuenta otras variables relevantes que lo describan. Con esta ecuación, se pueden hacer ajustes para considerar las variaciones individuales.

* AAA
* BBB
* CCC

1. **Perfil Edad – Salarios**

****

1. **Brecha salarial de género**
2. **Predicción de ingresos**
3. **Bibliografía**