1. INTRODUCCCIÓN

En Colombia, de acuerdo con la Dirección de Impuestos y Aduanas Nacionales (DIAN) la evasión de impuestos es cercana a los $65 billones de pesos, estimaciones indican que la evasión esta cercana a 5,4 puntos del PIB, de los cuales 0,7 puntos corresponden evasión de impuestos de personas naturales. [[1]](#footnote-1) Algunos trabajos han estimado el subreporte de ingresos en 1.66 veces mas que el reporte inicial. (Rocha (2014)). Por lo anterior, usando metodologías aprendizaje estadístico, aun incipientes en economía, estimamos un modelo de predicción de salarios basado en información de la GEIH para el año 2018 en Bogota. Este modelo predictivo es potencialmente útil para encontrar casos de fraude fiscal y adicionalemente, apoyar la toma de decisiones de política para intervenir familias vulnerables.

1. DATOS

Usamos datos de la Gran Encuesta Integrada de Hogares GEIH en Bogota para el año 2018, la cual contiene información sobre las condiciones de empleo de las personas, características generales de la población como sexo, edad, estado civil y nivel educativo. Contiene también información sobre fuentes de ingreso.

El proceso de adquisición de la información se realizo mediante el scraping la información contenida en el sitio web <https://ignaciomsarmiento.github.io/GEIH2018_sample/>, sobre esta información no existe ningún tipo de restricción. El sitio web contiene información extraída y clasificada de la GEIH con variables originales y construidas que proporcionan datos relevantes para el análisis. Se realizo un web scrapping de los 10 bloques de información los cuales estaban particionados en diferentes URL´s. Estos bloques de información se consolidaron para terminar con una sola base de datos integrada de la información relevante.

* 1. *Limpieza de datos*

Considerando que la GEIH en lo concerniente a los procedimientos de conformación, depuración, imputación y empalme del ingreso contiene 137 variables con las características de los individuos pero sobre todo aproximadamente 40 variables que contienen información relacionada a ingreso, es necesario realizar un procedimiento de limpieza de la base en el cual se restringe el universo de datos en aquellos relevantes para los objetivos del trabajo, en particular usamos procedimientos para limpiar la base y dejar únicamente las siguientes variables: *i) Sexo*; Variable dicótoma que toma valor 1 si es hombre y 0 si es mujer. Usamos esta variable con la finalidad de identificar brechas salariales entre hombres y mujeres. *ii) Edad;* contiene la edad en años del individuo, la muestra se analiza para aquellos mayores de 18 años. Esta variable se usa para identificar patrones predictivos entre la edad y el salario de las personas *iii) Educación*; contiene e máximo grado escolar alcanzado

1. Ver:<https://www.larepublica.co/especiales/reforma-tributaria-2022/segun-la-dian-y-el-minhacienda-la-evasion-de-impuestos-es-cercana-a-80-billones-3422523> [↑](#footnote-ref-1)