TALLER 1

BIG DATA Y MACHINE LEARNING PARA ECONOMÍA APLICADA

MECA

# INTRODUCCIÓN

La evasión tributaria tiene grandes repercusiones sobre el desarrollo y crecimiento de los países, limitando las capacidades del gobierno como gestor de política pública. Los desafíos relacionados con la regresividad en la recaudación de impuestos aumentan la brecha de desigualdad y afectan directamente el bienestar de los ciudadanos, esto implica que, aquellos individuos cuyo nivel de ingresos es alto, su tasa efectiva de impuestos pagada disminuye, afectando la equidad social y haciendo que los individuos con menos recursos tengan cargas impositivas mayores. Esto puede resultar en una disminución de los ingresos fiscales totales del gobierno y afectando significativamente al presupuesto del gobierno dificultando la posibilidad de tener un gasto eficiente y en alguna medida austero. (Moller, 2012)

La evasión tributaria se ve fuertemente afectada por la informalidad y el desempleo, la informalidad laboral es la proporción de ocupados que reciben ingresos al margen del control tributario y bajo condiciones que no están dentro del aval institucional. El trabajador informal, normalmente, no paga sus cargas correspondientes a la obra social y a la jubilación, además, no dispone de seguro médico(entre otras cosas)**.** El problema de la informalidad laboral en Colombia repercute significativamente en la seguridad social, especialmente porque los trabajadores informales no tienen la posibilidad de realizar las cotizaciones que constituyen una fuente de financiación en el sistema de pensiones(Montenegro, 2018).

Así mismo, la ausencia de seguridad social es una característica que define a las personas ocupadas dentro del sector informal, en este sentido, se puede hablar de una relación inversa entre los subsidios o ayudas por parte del Estado y el incentivo de participar en el mercado laboral, donde a mayor ayuda del gobierno la probabilidad de pertenecer a la informalidad es mayor. Estas ayudas pueden generar incentivos perversos sobre la decisión de participar en el mercado laboral por parte de los integrantes de un hogar beneficiario del programa, generando mayor comodidad en pertenecer al régimen subsidiado que al contributivo y a su vez, propiciando a que las personas prefieran trabajos informales (Farné, David, & Paola, 2016).

Según la teoría convencional de ingreso de los hogares, la entrega de una transferencia produce un puro efecto ingreso al igual que por alquileres, intereses, sueldos y salarios, entre otros. Por lo que, “si la entrega del subsidio está sujeta a los (bajos) ingresos del hogar, puede que algunos miembros de la familia decidan trabajar menos, o simplemente no trabajar, para no exponerse a ser reportados como receptores de remuneraciones por encima de los umbrales máximos establecidos para ser beneficiarios” (Moffit, 1996). Atendiendo a estas consideraciones, se entiende que los subsidios pueden desestimular la permanencia en el mercado del trabajo o la formalización en el mismo.

En retrospectiva, es imperativo el estudio de ingresos individuales en el contexto de una reforma estructural para la evasión de impuestos, considerando las diversas condiciones salariales de la población mayor a 18 años, empleando métodos técnicos que ayuden a predecir el nivel de ingresos asertivamente. Por lo que es de interés en este trabajo de investigación, estimar un modelo econométrico a partir de los datos obtenidos en la Gran Encuesta Integrada de Hogares para 2018 realizada por el Departamento Administrativo Nacional de Estadística (DANE). El trabajo de investigación tendrá cinco secciones, de la cual la primera fue esta introducción, en la segunda sección se explicará el proceso de adquisición, limpieza y descripción de los datos y las variables utilizadas. PLANTEAR HIPÓTESIS O BREVE DESCRIPCIÓN DE LOS RESULTADOS. La tercera establece modelos perfil Edad-Salario con diferentes metodologías empleada. La cuarta analiza la brecha salarial de género. La quinta y última parte, trata una estimación de ganancias por modelo, las conclusiones y limitaciones de esta investigación.

# DATOS

En esta sección se hace una descripción de los datos utilizados para el análisis y su finalidad. Posteriormente, se muestra el proceso realizado para obtener los datos y realizar su respectiva limpieza y procesamiento. Por último, se presenta un reporte de estadísticas descriptivas de las variables a utilizar para el desarrollo de este problem set.

## Descripción de los datos

La información utilizada proviene del informe de Medición de Pobreza Monetaria y Desigualdad del año 2018, realizado por el Departamento Nacional de Estadística (DANE) a partir de la Gran Encuesta Integrada de Hogares (GEIH). Esta encuesta proporciona información estadística sobre el tamaño y estructura de la fuerza de trabajo (empleo, desempleo y población fuera de la fuerza de trabajo), los ingresos laborales y no laborales de los hogares, la pobreza monetaria y la pobreza monetaria extrema de la población residente en el país.

Las temáticas por las cuales se indagan en la GEIH permiten caracterizar a la población según sexo, edad, parentesco con el jefe del hogar, nivel educativo, afiliación al sistema de seguridad social en salud, grupos poblacionales como etnias, campesinos, LGBT o con algún tipo de discapacidad, otras formas de trabajo como producción de bienes y servicios para autoconsumo, trabajo en formación y voluntariado, entre otras.

Actualmente, la GEIH cuenta con una muestra anual aproximada de 315.000 hogares a nivel nacional, lo que hace que sea la encuesta de mayor cobertura a nivel nacional. De modo que, la GEIH permite obtener indicadores confiables y series continuas para analizar la fuerza de trabajo del país y los principales indicadores del mercado laboral, considerados como información fundamental para la toma de decisiones de política pública.

Respecto al diseño estadístico, la GEIH tiene cobertura nacional con diferentes niveles de desagregación temporal y geográfica: total nacional, total de cabeceras de ciudades (con o sin áreas metropolitanas), grandes agrupaciones (cabeceras, centros poblados y rural disperso) y departamentos. Además, tiene desagregación anual, semestral, trimestral y mensual. Su unidad de observación son las viviendas, los hogares y las personas,

## Proceso de adquisición de los datos

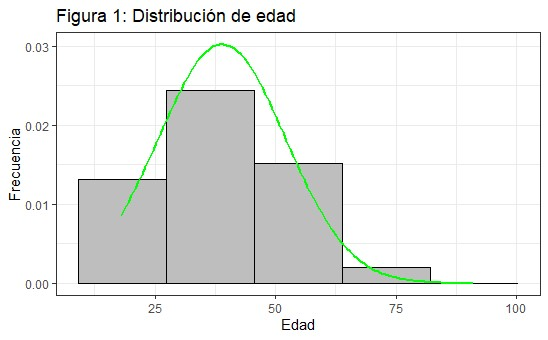
## Limpieza de los datos

Para obtener la base de datos final para estimar los modelos solicitados en los siguientes puntos, se realizó un conteo de los datos faltantes en la base de datos. Se mantienen las variables que tienen el 15% de datos faltantes o menos, para que los modelos estimados no pierdan poder de predicción.

Con los datos restantes, se revisaron las variables que pueden explicar mejor los cambios en el ingreso como la educación, el estrato socioeconómico, el sexo de nacimiento, la edad, si se encuentra ocupado o no, entre otras.

Por último, se filtró la base de datos para quedar con las observaciones de personas mayores de 18 años y que se encuentran ocupadas.

## Análisis descriptivo de los datos



Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Histograma, Gráfico de cajas y bigotes

Descripción generada automáticamente

# PERFIL EDAD-SALARIO

Regresión

Interpretación de los coeficientes

Ajuste del modelo

Gráfico del perfil estimado

# BRECHA SALARIAL DE GÉNERO

# PREDICCIÓN DE LAS GANANCIAS POR MODELO

Comparar los modelos vistos en clase

Repositorio GirHub:

# Referencias

Moller, C. L. (2012). *¿Por qué Colombia necesita un sistema tributario más progresivo?* Banco Mundial. Obtenido de https://www.bancomundial.org/es/news/opinion/2012/12/17/why-colombia-needs-a-more-progressive-tax-system