TRABAJO PRÁCTICO 1

Cantero, Lara Sofia

Temossi, Francisco

Vollmer, Candela

CONTEXTO: En el sector público, un reporte preciso de los ingresos de los individuos es indispensable para el cálculo de los impuestos. Sin embargo, el fraude fiscal siempre ha sido un problema importante. De acuerdo con el IRS, un 83,6% de los impuestos son pagados voluntariamente y a tiempo en Estados Unidos. Una de las causas de esa brecha es la subdeclaración de ingresos por parte de los individuos.

Un modelo de predicción de ingresos podría ayudar potencialmente a detectar casos de fraude que podrían conducir a la reducción de la brecha. Además, un modelo de predicción de ingresos puede ayudar a identificar a personas y familias vulnerables que puedan necesitar más asistencia.

El objetivo principal del trabajo es construir un modelo de salario horario individual:

$$w = f(X) + u$$

donde X es una matriz que incluye potenciales variables explicativas/predictoras.

I. INTRODUCCIÓN: Exponer brevemente el problema y si existen antecedentes de ello. Describir brevemente los datos y su idoneidad para abordar el problema planteado. Debe contener una vista previa de los resultados y las principales conclusiones.

La obtención de información precisa sobre los ingresos es un desafío, especialmente en economías en desarrollo como las de América Latina. Los impuestos sobre la renta constituyen una fuente importante de recursos para los países; sin embargo, es común que algunos individuos subdeclaren sus ingresos para evadir parte de sus obligaciones fiscales. Aunque existen agencias especializadas en detectar estos casos de evasión, los procesos de auditoría suelen ser costosos. En este contexto, han comenzado a implementarse técnicas de *machine learning*, las cuales son eficaces para predecir un resultado sin requerir una comprensión detallada de la relación causal entre la variable dependiente, como ingresos o beneficios declarados, y las variables utilizadas para la predicción, como características socioeconómicas. En términos generales, los modelos predictivos pueden contribuir a una recaudación más eficiente.

En cuanto al uso de *machine learning* para la predicción de ingresos, la evidencia aún es incipiente. No obstante, estudios recientes, como los de Jo (2024), Matkowski (2021) y Wang (2022), han mostrado ciertos avances en este ámbito.

Para el caso colombiano, De Roux et al. (2018) utilizaron datos de declaraciones de impuestos de la ciudad de Bogotá y aplicaron una técnica de aprendizaje no supervisado para identificar subdeclaraciones en los impuestos de construcción. Por su parte, Battiston et al. (2024) emplearon métodos de clasificación supervisada para predecir el subreporte de ingresos por parte de trabajadores independientes en Italia. JOE utilizaron datos administrativos para identificar qué casos de evasión permiten maximizar los ingresos de la agencia tributaria tras una auditoría.

Sobre la adopción de estas técnicas por los organismos tributarios, la OCDE (2021) estima que aproximadamente el 75% de las administraciones fiscales ya implementan *machine learning* con el fin de mejorar la eficiencia y optimizar la asignación de recursos.

Los datos por utilizar se obtuvieron del reporte "Medición de Pobreza Monetaria y Desigualdad" que extrae información de las bases de microdatos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 que elabora el Departamento Administrativo Nacional de Estadística (DANE). La GEIH es una integración de las tres más importantes Encuestas a Hogares de Colombia: la Encuesta Continua de Hogares (ECH), la Encuesta Nacional de Ingresos y Gastos (ENIG) y la Encuesta de Calidad de Vida (ECV). Se propuso en 2005 con el propósito de ampliar el alcance temático de la investigación y reducir el costo de la aplicación.

La recolección de la GEIH empezó el 7 de agosto de 2006 en su módulo central de mercado laboral e ingresos y, a partir del 11 de septiembre, con su módulo de gastos de los hogares.

Las características más importantes de esta encuesta son:

- ✓ Unidad de observación: Hogares.
- ✓ Muestra (total anual): 240.000 hogares, aproximadamente.
- ✓ Periodicidad: trimestral y anual.
- ✓ Temas: algunos son ingresos, propiedad, ahorro/inversión, empleo y desempleo, situación económica, economía rural, vivienda, fertilidad y uso de servicios públicos.
- ✓ Cobertura: cobertura nacional que permite obtener resultados por zona urbana y rural, cinco grandes regiones (Región Atlántica, Región Oriental, Región Central, Región Pacífica y Región Bogotá) y total por departamentos: 23 departamentos, las 13 grandes ciudades con sus áreas metropolitanas y 11 ciudades intermedias.

La GEIH es especialmente útil para abordar problemas relacionados con la estructura del mercado laboral y la distribución del ingreso. Es por esto por lo que se considera idónea para un modelo de predicción de salarios.

II. DATOS:

a) Describa brevemente los datos, incluyendo su propósito y cualquier otra información relevante.

La Gran Encuesta Integrada de Hogares es una encuesta mediante la cual se solicita información sobre las condiciones de empleo de las personas en Colombia (si trabajan, en qué trabajan, cuánto ganan, si tienen seguridad social en salud o si están buscando empleo), además de las características generales de la población como sexo, edad, estado civil y nivel educativo, se pregunta sobre sus fuentes de ingresos. La GEIH proporciona al país información a nivel nacional, cabecera - resto, regional, departamental, y para cada una de las capitales de los departamentos.

En el presente trabajo se seleccionaron unas pocas variables de todas incluidas en la encuesta porque se creyeron pertinentes para poder formular un modelo de predicción de salarios. Algunas de ellas son: el ingreso horario nominal, edad, horas trabajadas, condición del trabajador, etc.

b) Describa el proceso de adquisición de los datos y si existen restricciones para acceder o extraer estos datos.

Los datos fueron obtenidos mediante la técnica de web scraping de una página web que contenía información detallada sobre 32,177 observaciones correspondientes a la Gran Encuesta Integrada de Hogares (GEIH) del año 2018. Como el acceso a los microdatos anonimizados de uso público es de carácter gratuito y está disponible en la página Web del DANE, no se nos presentó ninguna restricción para su uso.

c) Describa el proceso de limpieza de datos y

La base de datos final se ha conformado mediante la filtración de observaciones que incluyan únicamente a individuos, que claramente hayan contestado la encuesta, con ingresos superiores a cero y empleados mayores de 18 años.

d) Un análisis descriptivo de los datos.

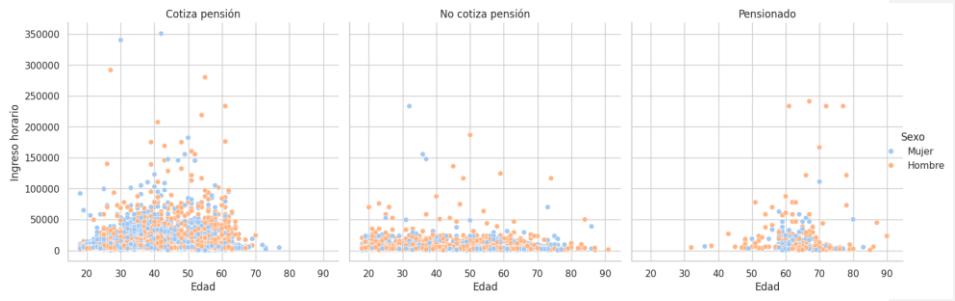
	Ingreso horario nominal	Edad	Sexo	Hs trabajadas (usualm.)	Hs trabajadas (total)	Micro Empresa	Antig.	Formal
Obs.	14764	14764	14764	14764	14764	14764	14764	14477
Media	8541.87	38.89	0.53	47.2	47.6	0.43	61.97	0.6
S.E.	13866.13	13.2	0.5	15.06	15.16	0.49	88.03	0.49
Mín.	0.47	18.0	0.0	1.0	1.0	0.0	0.0	0.0
25%	3796.53	28.0	0.0	40.0	40.0	0.0	7.0	0.0
50%	4837.49	37.0	1.0	48.0	48.0	0.0	24.0	1.0
75%	7899.31	49.0	1.0	50.0	50.0	1.0	72.0	1.0
Máx.	350583.34	91.0	1.0	130.0	130.0	1.0	720.0	1.0

La submuestra seleccionada, compuesta por personas empleadas mayores de 18 años con salario horario positivo, incluye 14764 observaciones, representando aproximadamente el 46% de la muestra original. La media y la mediana del ingreso horario confirman la esperada asimetría positiva (o hacia la derecha) de la distribución. La edad promedio de la población analizada es de 39 años, y el 53% de los individuos son hombres. En cuanto a las horas trabajadas, el promedio semanal asciende a 47 horas, lo cual sugiere una jornada laboral promedio de 8 horas diarias durante 6 días a la semana, superando el estándar de 40 horas semanales.

La variable "microempresa" permite observar que el 43% de los trabajadores están empleados en establecimientos con cinco trabajadores o menos, incluyendo a los cuentapropistas. En relación con la antigüedad, esta presenta una gran dispersión, con un promedio de cinco años. Cabe destacar que el 25% de los encuestados tiene una permanencia de siete meses o menos en su empresa actual.

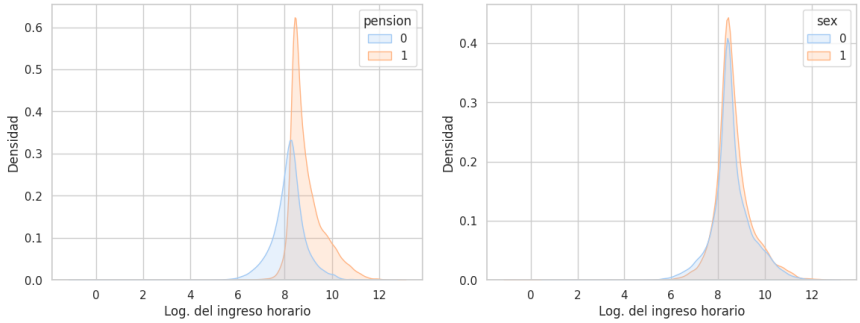
Por último, se ha creado una variable "formal" que toma valor 1 si el individuo cotiza pensión y 0 en caso contrario (excluyendo a los pensionados), lo cual permite aproximar el nivel de informalidad de la economía. Al considerar la submuestra completa, la informalidad alcanza aproximadamente el 40%, pero disminuye al 23% al excluir a cuentapropistas y empleadores.

La Figura X ilustra la relación entre ingreso horario y edad, desagregada por sexo y condición laboral (formal, informal o pensionado).



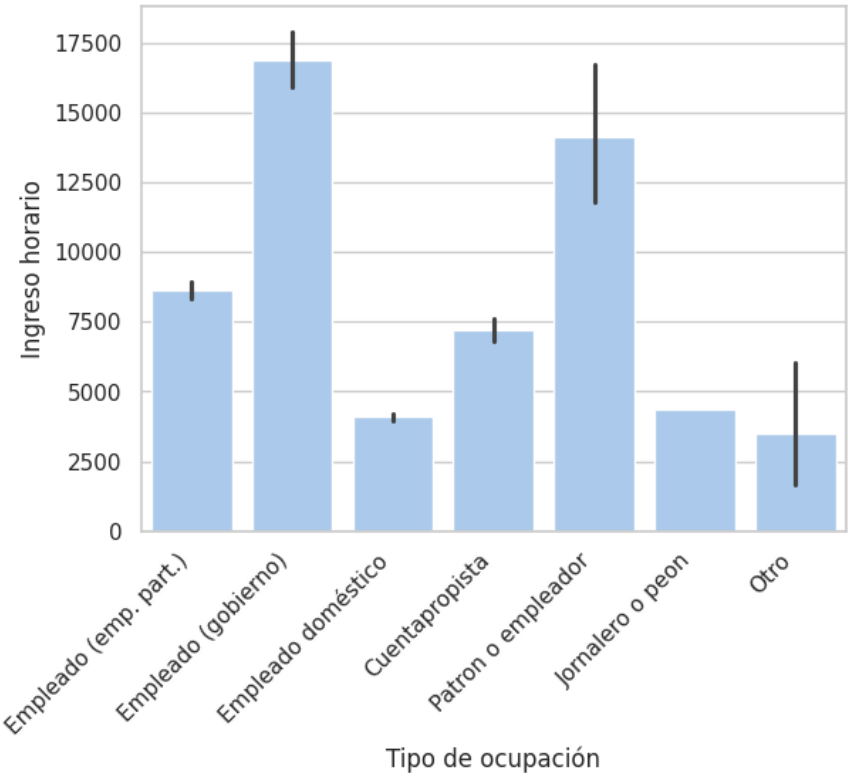
La Figura X muestra la densidad del logaritmo del ingreso horario, desglosada por condición laboral (formal o informal) y por sexo. En el panel A, se observa una diferencia notable entre trabajadores formales e informales. Es plausible que los trabajadores formales se vean influenciados por una medida de salario mínimo, lo que contribuye a que sus ingresos sean, en promedio, más altos que los de los trabajadores informales. En cuanto a las diferencias de género, aunque ambos grupos presentan distribuciones similares, los hombres alcanzan un pico más alto en la densidad de ingresos y acumulan menor densidad en los niveles de ingreso más bajos.

Comentado [S1]: Pensión 1 indica formal y sexo 1 indica hombre. Tengo que cambiar las leyendas.



La Figura X muestra el ingreso promedio reportado por individuos según el tipo de ocupación, junto con sus respectivos intervalos de confianza. Esto sugiere que el tipo de

ocupación puede ser un predictor relevante del nivel de ingresos.



III. PREDICCIÓN DE SALARIOS:

Referencias

Battaglini, M; Guiso, L.; Lacava, C.; Miller, D. L.; Patacchini, E. (2024). Refining public policies with machine learning: The case of tax auditing. Journal of Econometrics. https://doi.org/10.1016/j.jeconom.2024.105847.

Battiston, P.; Gamba, S.; Santoro, A. (2024). Machine learning and the optimization of prediction-based policies. Technological Forecasting and Social Change 199. https://doi.org/10.1016/j.techfore.2023.123080.

de Roux, D.; Perez, B.; Moreno, A.; Villamil, M.; Figueroa, C. (2018). Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 215–222. https://doi.org/10.1145/3219819.3219878

Jo, K. (2024). Income Prediction Using Machine Learning Techniques. UCLA. ProQuest ID: Jo_ucla_0031N_22838. Merritt ID: ark:/13030/m54c47kf. https://escholarship.org/uc/item/6d01c9v7

Matkowski, M. (2021). Prediction of Individual Level Income: A Machine Learning Approach [Honors Thesis, Bryant University]. Archivo digital. https://digitalcommons.bryant.edu/honors_economics/39/

OCDE (2021), Tax Administration 2021: Comparative Information on OECD and other Advanced and Emerging Economies, OECD Publishing, Paris, https://doi.org/10.1787/cef472b9-en.

Wang, J. (2022). Research on Income Forecasting based on Machine Learning Methods and the Importance of Features. ICIDC – EAI. https://eudl.eu/doi/10.4108/eai.17-6-2022.2322745

Variable	Pregunta/contenido	Valor
y_total_m_ha	income salaried + independents total - nominal hourly	
age	Edad	
sex		=1 male, =0 female
clase	Área	=1 urban; =0 rural
cotPension	¿está cotizando actualmente a un fondo de pensiones?	=1 cotiza pensión; 2=no cotiza pensión; =3 pensionado; =9 N/A
hoursWorkUsual	usual weekly hours worked - principal occ	
totalHoursWorked	total hours worked previous week	
maxEducLevel	max. education level attained	
microEmpresa =1	if 5 workers or less in firm; =0 otherwise; incluye cuenta propia	
sizeFirm	size of the firm by categories	=1: self-employed; =2: 2-5 workers; =3: 6-10 workers; =4: 11-50 workers; 5: >50 workers
p6426	¿cuánto tiempo lleva trabajando en esta empresa, negocio, industria, oficina	
relab	type of occupation	Por ejemplo 1 "Obrero o empleado de empresa particular"; 2 "Obrero o empleado del gobierno";
oficio	occupation	

Comentado [S2]: Entiendo que es el salario horario tanto de asalariados como de independientes, construido a partir de las variables de ingreso y hs trabajadas. No aparece en la documentación, sólo en el diccionario.

Comentado [S3]: La agregaría como indicador de formalidad (salarios formales son típicamente más altos que los informales)

Comentado [S4]: Esta tiene la ventaja de que incluye horas de la ocupación secundaria, pero puede haber un poco más de ruido al ser únicamente de la semana previa

Comentado [S5]: Elegir una variable de horas trabajadas. De igual manera tienen características muy parecidas (ver tabla estadística descriptiva)

Comentado [S6]: Elegir microempresa O sizeFirm

Comentado [S7]: La agregaría como indicador de antigüedad

Comentado [S8]: Tiene muchas categorías, capaz es un montón

Comentado [S9]: Elegir relab U oficio