Predicción de Ingresos: Análisis del Mercado Laboral en Bogotá

Harold Stiven Acuña

José David Cuervo
José David Dávila
César Augusto Alfaro

3 de marzo de 2025

Resumen

[Pendiente: Escribir abstract de 100 palabras]

Palabras clave: predicción de ingresos, brecha salarial, economía laboral, aprendizaje automático
Clasificación JEL: J31, C53, J16

Repositorio GitHub:
https://github.com/alfarocesar/BDML_Predicting-income_Equipo8

1. Introducción

En el sector público, la precisión en la declaración de ingresos individuales es fundamental para el cálculo de impuestos. Sin embargo, el fraude fiscal de todos los tipos ha sido siempre un problema importante. Según el Servicio de Impuestos Internos (IRS), aproximadamente el 83,6 % de los impuestos en Estados Unidos se pagan voluntaria y puntualmente. Una de las causas de esta brecha es la subdeclaración de ingresos por parte de los individuos. Un modelo de predicción de ingresos podría potencialmente ayudar a identificar casos de fraude que podrían llevar a la reducción de esta brecha. Además, puede ayudar a identificar a individuos y familias vulnerables que necesitan asistencia adicional.

El análisis del mercado laboral es fundamental para comprender la distribución de los ingresos y la dinámica de las desigualdades económicas. En este contexto, predecir el salario por hora de los trabajadores es una tarea clave para evaluar el impacto de diferentes factores individuales y estructurales en los ingresos laborales.

En Colombia, el Departamento Administrativo Nacional de Estadística (DANE) ha desarrollado metodologías para la medición de pobreza monetaria y desigualdad, utilizando información de la Gran Encuesta Integrada de Hogares (GEIH). Estos esfuerzos proporcionan una base sólida para la construcción de modelos predictivos de ingresos. La Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP) ha realizado un trabajo significativo en la actualización y mejora de las metodologías de medición.

El objetivo principal de este estudio es construir un modelo de predicción del salario horario individual: $w = f(X) + u$, donde w es el salario horario y X es una matriz de variables explicativas o predictores. En este trabajo, nos enfocaremos en la forma funcional $f(X) = X\beta$.

En este contexto, el presente estudio se propone construir un modelo predictivo de ingresos utilizando datos de la GEIH 2018 para Bogotá, con el objetivo de explorar

las relaciones entre características individuales y niveles de ingreso, así como evaluar la capacidad predictiva de diferentes especificaciones estadísticas. La investigación se centra particularmente en el análisis del perfil edad-salario y la brecha salarial de género, temas de gran relevancia tanto para la política fiscal como para las políticas de equidad e inclusión social.

Para abordar este problema, se emplean datos de la GEIH 2018, que proporciona información detallada sobre la situación laboral, ingresos, educación y características sociodemográficas de los individuos. La base de datos incluye 178 variables y más de 32,000 observaciones. Dado que contiene información detallada sobre el nivel educativo, horas trabajadas, informalidad, estrato socioeconómico y otras variables clave, se considera adecuada para responder a la pregunta de investigación sobre los determinantes del ingreso laboral.

2. Datos

En esta sección se presenta el proceso de obtención, limpieza y análisis descriptivo de los datos utilizados para modelar los salarios individuales en Bogotá.

2.1. Descripción de los datos

Los datos utilizados en este análisis provienen de la Gran Encuesta Integrada de Hogares (GEIH) de 2018, específicamente para la ciudad de Bogotá. La GEIH es una encuesta de carácter continuo realizada por el Departamento Administrativo Nacional de Estadística (DANE) con el propósito de proporcionar información básica sobre el tamaño y estructura de la fuerza de trabajo del país, así como las características sociodemográficas de la población colombiana.

El conjunto de datos original contiene información de 32,177 individuos y 178 variables que abarcan aspectos demográficos, educativos, laborales y económicos. Esta riqueza de información permite un análisis detallado de los determinantes del ingreso laboral en la

capital colombiana y facilita la construcción de modelos predictivos robustos que puedan contribuir a la identificación tanto de posibles casos de subdeclaración de ingresos como de hogares en situación de vulnerabilidad.

La GEIH resulta particularmente adecuada para los objetivos de este estudio debido a su diseño metodológico riguroso y a la amplitud de las variables recolectadas. Por un lado, la encuesta captura información detallada sobre los diversos componentes del ingreso laboral (salarios, bonificaciones, horas extras, etc.), lo que permite una caracterización precisa de nuestra variable dependiente. Por otro lado, incluye un amplio conjunto de características individuales y laborales que la literatura económica ha identificado como determinantes clave de los ingresos, facilitando la especificación de modelos con alto poder explicativo.

2.2. Proceso de adquisición de datos

Los datos analizados en este estudio provienen de la Gran Encuesta Integrada de Hogares (GEIH) de 2018, específicamente para la ciudad de Bogotá. La GEIH es una encuesta que realiza el Departamento Administrativo Nacional de Estadística (DANE) de Colombia con el propósito de proporcionar información básica sobre el tamaño y estructura de la fuerza de trabajo del país (empleo, desempleo e inactividad), así como de las características sociodemográficas de la población.

Para la adquisición de los datos, desarrollamos un algoritmo automatizado que identifica y extrae las tablas contenidas en cada una de las diez páginas HTML disponibles en el sitio https://ignaciomsarmiento.github.io/GEIH2018_sample/. El proceso consistió en realizar solicitudes HTTP a cada URL, procesar el contenido HTML recibido y extraer las estructuras tabulares utilizando las bibliotecas `httr` y `rvest` en R.

Una consideración importante durante la extracción fue la necesidad de respetar las políticas del sitio web, incluyendo pausas entre solicitudes para evitar sobrecargar el servidor. No se identificaron restricciones explícitas para el acceso a estos datos, ya que se encuentran disponibles públicamente con fines académicos. Sin embargo, aplicamos prác-

ticas éticas de web scraping, limitando la frecuencia de solicitudes y utilizando únicamente los datos necesarios para nuestro análisis.

El conjunto de datos original contiene información de 32,177 individuos y 178 variables que abarcan aspectos demográficos, educativos, laborales y económicos. Esta amplia gama de variables permite un análisis detallado de los determinantes del ingreso y facilita la construcción de modelos predictivos robustos. Las variables incluyen datos sobre edad, sexo, nivel educativo, estado laboral, horas trabajadas, tipo de ocupación, tamaño de la empresa, y diversos componentes del ingreso laboral.

El código desarrollado para esta tarea se encuentra disponible en nuestro repositorio de GitHub, permitiendo la reproducibilidad completa del proceso de adquisición de datos.

La GEIH es particularmente adecuada para abordar el problema planteado en este estudio por varias razones. Primero, contiene mediciones detalladas de ingresos laborales desagregados por fuente (salario base, horas extras, bonificaciones, etc.), lo que permite una caracterización precisa de la variable dependiente. Segundo, incluye un amplio conjunto de características individuales y laborales que la literatura ha identificado como determinantes clave de los ingresos. Tercero, su diseño muestral garantiza representatividad para la ciudad de Bogotá, permitiendo generalizaciones válidas para la población urbana de la capital colombiana.

2.3. Proceso de limpieza de datos

El proceso de limpieza de datos comprendió varias etapas orientadas a obtener una muestra analítica adecuada para la modelación de ingresos laborales. Partimos de los datos crudos obtenidos mediante web scraping, que constaban de 32,177 observaciones y 178 variables.

En primer lugar, realizamos una exploración exhaustiva de los valores faltantes en el conjunto de datos. El análisis reveló un patrón estructural en los datos ausentes, donde aproximadamente el 60 % de los datos están ausentes. Este patrón es esperado en encues-

tas de hogares como la GEIH, ya que muchas preguntas son aplicables solo a subgrupos específicos de la población (por ejemplo, las preguntas sobre características laborales solo aplican a personas empleadas).

La exploración de correlación entre valores faltantes nos permitió identificar grupos de variables que sistemáticamente presentan datos ausentes para los mismos individuos. Este análisis fue crucial para comprender la estructura de los datos y desarrollar una estrategia adecuada de manejo de valores faltantes.

Siguiendo los criterios establecidos en las instrucciones del problema, filtramos la muestra para incluir únicamente a individuos mayores de 18 años y que estuvieran empleados al momento de la encuesta. Esta restricción redujo la muestra a aproximadamente 16,682 observaciones, que representan adecuadamente a la población objetivo de nuestro estudio: la fuerza laboral adulta de Bogotá.

Posteriormente, creamos las variables clave para nuestro análisis:

1. **Salario por hora (hourly_wage)**: Calculado como el ingreso laboral mensual dividido por el número de horas trabajadas mensualmente (considerando 4.345 semanas por mes).

2. **Variable indicadora de género (female)**: Recodificada para que tome el valor de 1 para mujeres y 0 para hombres.

3. **Variables para el análisis del perfil edad-salario**: Incluimos tanto la edad como su término cuadrático.

4. **Logaritmo del salario por hora (log_hourly_wage)**: Transformación que permite normalizar la distribución de los salarios y facilita la interpretación de los coeficientes como elasticidades o cambios porcentuales.

Para garantizar la calidad de los datos, implementamos un proceso de detección y manejo de valores atípicos (outliers) en los salarios. Calculamos los z-scores para la variable de salario por hora y excluimos las observaciones con valores absolutos superiores a 3, lo

que corresponde a salarios que se desvían más de tres desviaciones estándar de la media. Este filtro nos permitió eliminar casos extremos que podrían distorsionar los resultados del análisis, manteniendo al mismo tiempo una muestra representativa.

Finalmente, eliminamos las observaciones con valores faltantes en las variables clave para el análisis, asegurando así que el conjunto de datos final estuviera listo para la modelación estadística. El conjunto de datos limpio y procesado fue almacenado en formato RDS para su uso en las etapas posteriores del análisis.

La siguiente tabla resume el tamaño de la muestra después de cada etapa del proceso de limpieza, destacando la proporción de observaciones que se conservaron en cada paso:

Cuadro 1. Resumen de la muestra después de cada etapa del proceso de limpieza

Etapa	**Observaciones**	**Porcentaje**
Muestra original	32,177	100.0 %
Filtro por edad (>18) y ocupación	16,682	51.8 %
Después de la creación de variables	16,682	51.8 %
Después de filtrar outliers	16,015	49.8 %
Muestra analítica final	16,015	49.8 %

Fuente: Elaboración propia con datos de la GEIH 2018 para Bogotá.

2.4. Variables incluidas en el análisis

Las variables seleccionadas para nuestro análisis fueron cuidadosamente elegidas con base en la literatura económica sobre determinantes de ingresos y considerando su relevancia para los objetivos específicos del estudio. A continuación, presentamos las principales variables incluidas:

2.4.1. Variable dependiente

- **log_hourly_wage**: Logaritmo natural del salario por hora, calculado a partir del ingreso laboral mensual reportado (y_ingLab_m) dividido por las horas trabajadas mensualmente. Esta transformación logarítmica es estándar en la literatura econó-

mica, ya que normaliza la distribución típicamente sesgada de los ingresos y permite interpretar los coeficientes como cambios porcentuales aproximados.

2.4.2. Variables independientes clave

- **age**: Edad del individuo en años, variable fundamental para el análisis del perfil edad-salario.
- **age_squared**: El cuadrado de la edad, incluido para capturar la relación no lineal (cóncava) entre edad e ingresos documentada en la literatura.
- **female**: Variable binaria que toma el valor de 1 para mujeres y 0 para hombres, utilizada para analizar la brecha salarial de género.

2.4.3. Variables de control

- **educ_level**: Nivel educativo máximo alcanzado, categorizado en niveles que van desde sin educación formal hasta posgrado.
- **formal_work**: Indicador de formalidad laboral, que distingue entre trabajadores formales e informales.
- **totalHoursWorked**: Total de horas trabajadas, que captura la intensidad de la participación laboral.
- **sizeFirm**: Tamaño de la empresa donde trabaja el individuo, variable que aproxima características no observables del empleo como tecnología y productividad.

2.4.4. Análisis descriptivo de las variables clave

Observamos que la edad promedio de los trabajadores en la muestra es de aproximadamente 34.4 años, con una desviación estándar considerable de 20.9 años, lo que indica una fuerza laboral diversa en términos etarios. La proporción de mujeres en la muestra

es del 47.9 %, lo que refleja una participación laboral femenina ligeramente inferior a la masculina en Bogotá.

Respecto a las variables laborales, el promedio de horas trabajadas semanales es de aproximadamente 47.2 horas, superior a la jornada laboral estándar de 40 horas, lo que sugiere una prevalencia de horas extras o empleos con jornadas extendidas. La tasa de formalidad laboral es del 58.2 %, indicando que más de la mitad de los trabajadores en la muestra tienen empleos formales.

Un hallazgo notable es la variabilidad en los ingresos laborales. El salario por hora presenta una distribución altamente heterogénea, con un coeficiente de variación elevado, lo que justifica tanto nuestra transformación logarítmica como nuestro enfoque en la modelación de los determinantes de esta variable.

Es importante destacar que, tras el proceso de limpieza y filtrado descrito en la sección anterior, nuestra muestra analítica final conserva la representatividad de la población ocupada de Bogotá mayor de 18 años, permitiendo inferencias válidas sobre los patrones salariales en esta población.

3. Perfil Edad-Salario

Los resultados de la regresión de $\log(w) = \beta_1 + \beta_2 * Age + \beta_3 * Age^2 + u$, permiten interpretar que la constante de 13.060 será el valor esperado del logaritmo del ingreso, mientras que, al analizar la parte lineal de esta regresión, es decir, el componente Age, este es estadísticamente significativo y denota que, ante aumentos marginales de la edad, el salario incrementará en un 3.8 % cuando los efectos no lineales son mínimos. Ahora bien, respecto del componente cuadrático de la ecuación, se observa que el coeficiente es negativo y significativo, lo cual indica que la función es cóncava y, en ese sentido, que los ingresos crecerán hasta llegar a un punto de inflexión, para luego decrecer.

Para hallar ese punto de inflexión, se calcula la siguiente derivada para la ecuación:

$$\frac{d(log(w))}{dAge} = \beta_1 \ + \beta_2 \ * Age + \beta_3 \ * Age^2 + u$$

Lo que resulta en:

$$\frac{d(log(w))}{dAge} = \beta_2 \ + 2\beta_3 \ * Age$$

Igualando a cero:

$$\beta_2 \ + 2\beta_3 \ * Age = 0$$

Encontramos el punto de inflexión que será:

$$Age_{max} = -\frac{\beta_2}{2\beta_3}$$

Por lo que:

$$Age_{max} = -\frac{0{,}038}{2(-0{,}0004)} = 47{,}5$$

Como resultado de reemplazar los coeficientes beta obtenidos en la regresión, se encuentra que los ingresos aumentan hasta la edad aproximada de 48 años y luego disminuyen. Es decir, 48 años es la edad donde el individuo alcanza su salario máximo.

Al respecto, Becker (1964) en el capítulo de su libro, denominado Age, Earnings, Wealth, and Human Capital, perfila la relación entre la edad y salario, y allí logra evidenciar que los ingresos tienden a ser relativamente bajos en los primeros años laborales; luego se incrementan gradualmente, en virtud del desarrollo de habilidades y la experiencia adquirida, hasta llegar a una edad pico, en el intervalo de 45 a 54 años, donde los individuos alcanzan su máximo salario. En ese sentido, el resultado de los coeficientes obtenidos en la regresión es apropiado y encuentra sustento en lo señalado por Becker (1964), tal y como se indicó.

En cuanto al ajuste del modelo, se observa que este posee un R^2 de 0.022, lo cual indica que el modelo solo es explicado en una pequeña proporción y ello sugiere que, en efecto, la edad es un valor determinante a la hora de estimar los ingresos. Sin embargo, se evidencia la necesidad de incluir más factores que afectan la estimación del salario, que pueden ser variables observables y no observables, tales como educación, experiencia, habilidades, entre otras. Es decir, para realizar un modelo que explique el nivel de salario, además de la edad, resulta necesario incluir otras variables que permitan obtener un resultado más acertado.

Finalmente, se realizó un test para medir la heterocedasticidad, del cual se obtuvo un p-valor menor a 0.05 y este indica la existencia de heterocedasticidad, frente a lo cual, entonces, se evidencia la necesidad de utilizar otra metodología para abordar el problema en la estimación del salario.

4. Brecha Salarial de Género

4.1. Identificación de la Variable Dependiente

Este análisis corresponde a un modelo de regresión en el que la variable dependiente es el **logaritmo del salario** (log_salario).

4.2. Interpretación de los Coeficientes

modelo_lm <− lm(log_salario ∼ age + I(age^2) + sex + p6100 + fex_c + p6426 + relab, data = train_data)

Cada coeficiente refleja el **impacto de la variable independiente en el logaritmo del salario**, manteniendo constantes las demás variables.

Cuadro 2. Resultados de la regresión del modelo de brecha salarial

Variables	**Coeficiente**	**Error estándar**	**Valor t**	**Pr(>|t|)**
(Intercept)	12.915	0.079	163.51	< 2e-16 ***
age	0.077	0.004	19.83	< 2e-16 ***
I(age^2)	-0.001	0.000	-19.81	< 2e-16 ***
sex	0.131	0.014	9.02	< 2e-16 ***
p6100	-0.023	0.019	-1.24	0.217
fex_c	-0.000	0.000	-0.29	0.773
p6426	0.003	0.000	22.92	< 2e-16 ***
relab	-0.098	0.017	-5.78	8.6e-09 ***

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.3. Evaluación de la Significancia Estadística

- **Variables significativas:** age, I(age^2), sex, p6100, p6426, relab.

4.4. Conclusiones sobre la Brecha Salarial

1. **El salario tiene una relación cuadrática con la edad**, es decir, aumenta con la edad, pero a un ritmo decreciente.
2. **El género influye en el salario**, con los hombres ganando más que las mujeres, lo que sugiere la presencia de una brecha salarial de género.
3. **La variable p6100 (Nivel de educación):** tiene un impacto negativo en el salario, mientras que p6426 (tiempo en el trabajo actual) tiene un impacto positivo.
4. **La relación laboral (relab) también impacta negativamente en el salario**, lo que podría estar relacionado con el tipo de contrato o estabilidad del empleo.
5. **El factor de expansión (fex_c) no parece tener un efecto significativo**, por lo que podría eliminarse del modelo sin afectar su interpretación.

Cuadro 3. Interpretación de los coeficientes del modelo

Variables	Interpretaciones
Intercepto ((Intercept))	- Tiene un valor alto y significativo, indicando el punto de referencia cuando todas las variables explicativas son cero. - Su intervalo de confianza es el más amplio, lo cual es común en modelos con escalas transformadas.
Edad (age)	- Su coeficiente es positivo, lo que indica que el salario tiende a aumentar con la edad. - Su intervalo de confianza es estrecho, sugiriendo que la estimación es precisa y significativa.
Edad al cuadrado (I(age^2))	- Su coeficiente es negativo, lo que confirma que la relación entre edad y salario no es lineal, sino cuadrática. - Indica que el salario crece con la edad, pero a un ritmo decreciente.
Sexo (sex)	- Su coeficiente es positivo y significativo. - Esto sugiere que hay una diferencia salarial entre géneros, donde los hombres (referencia) tienen un salario mayor en promedio en comparación con las mujeres.
Factor de expansión (fex_c)	- Su coeficiente es cercano a cero y su intervalo de confianza es amplio. - Esto sugiere que el factor de expansión no tiene un impacto significativo en el salario.
p6100 (Nivel de educación)	- Su coeficiente es negativo, lo que indica que esta variable reduce el salario. - Tiene un intervalo de confianza estrecho, lo que sugiere alta precisión.
p6426 (Tiempo en el trabajo actual)	- Su coeficiente es positivo, indicando un impacto positivo en el salario. - Su intervalo de confianza es estrecho, indicando alta precisión.
Relación laboral (relab)	- Su coeficiente es negativo, sugiriendo que ciertos tipos de relación laboral pueden estar asociados con menores salarios. - Tiene un intervalo de confianza estrecho, indicando que la estimación es robusta.

Cuadro 4. Comparativa de desempeño de modelos

Modelo	**RMSE**
Regresión Lineal	0.6610
Ridge	0.6689
Lasso	0.6609
Árbol de Decisión	0.6405
Random Forest	0.6172

4.5. Análisis de Resultados de los Modelos

- **El modelo con mejor desempeño es el Random Forest** (RMSE = **0.6172**), lo que sugiere que captura mejor las relaciones en los datos.
- **Los árboles de decisión también tienen un buen rendimiento** (RMSE = **0.6405**), superando a la regresión lineal y regularizada (Ridge y Lasso).
- **La regresión Ridge tiene el peor RMSE** (0.6689), lo que indica que no aporta una mejora significativa en comparación con la regresión lineal estándar.
- **La regresión Lasso es casi idéntica a la regresión lineal** (0.6609 vs. 0.6610), lo que sugiere que la penalización no está eliminando muchas variables o que las seleccionadas son las más relevantes.

4.6. Interpretación de los Resultados del Bootstrap

Cuadro 5. Resultados del análisis bootstrap

Parámetro (t*)	**Valor original**	**Sesgo (Bias)**	**Error estándar**
t1*	129.156	-0.0005547	0.07899
t2*	0.07746	0.002186	0.00390
t3*	-0.0009846	-7.18e-08	0.0000498
t4*	0.1309	0.00908	0.01449
t5*	-0.02287	-0.000236	0.01850
t6*	-3.53e-06	1.40e-06	0.0000121
t7*	0.00298	-1.58e-06	0.0001300
t8*	-0.09775	-2.48e-05	0.01686

- **El primer coeficiente (t1*)** tiene un valor alto (12.9156), pero con un error estándar relativamente pequeño (0.07899), lo que sugiere que es un estimador confiable.
- El sesgo en general es bajo, lo que indica que los coeficientes no están sobreajustados y se mantienen estables en diferentes muestras.
- **El coeficiente t8*** (-0.09775) tiene un error estándar algo alto (0.01686), lo que sugiere que su estimación podría ser menos precisa.
- **Los coeficientes más pequeños (t6* y t7*)** tienen errores estándar muy bajos, lo que indica que su variabilidad es mínima y su estimación es confiable.

El bootstrap confirma que la mayoría de los coeficientes son **estables y con bajo sesgo**. Sin embargo, algunas variables pueden tener menor precisión en su estimación (como t8*).

4.7. Interpretación del Intervalo de Confianza (IC)

Cuadro 6. Intervalos de confianza para los coeficientes

Variables	Lower	Upper
(Intercept)	127.621098540	130.685510906
age	0.0699493513	0.0851710103
I(age^2)	-0.0010824361	-0.0008892900
sex	0.1017960768	0.1587800722
p6100	-0.3166915835	-0.2440927380
fex_c	-0.0002464454	0.0002437665
p6426	0.0027246483	0.0032378946
relab	-0.1308652327	-0.0644208559

- **Variables con efecto significativo:** age, I(age^2), sex, p6100, p6426, y relab, ya que sus intervalos de confianza no incluyen el cero.
- **Variable no significativa:** fex_c, porque su intervalo sí incluye el cero, lo que sugiere que su impacto en el salario no es concluyente.
- Se confirma la relación cuadrática de la edad con el salario.

- El género (sex) influye de manera positiva en el salario, lo que podría indicar una brecha salarial de género.

5. Predicción de Ingresos

[Modelos predictivos y validación]

6. Conclusiones

[Conclusiones principales]

Tablas y Figuras

A. Apéndice: Tablas y Figuras