

# Problem Set 1: Predicting Income Big Data and Machine Learning

Yenny Rocio Castillo Rodríguez

Fabián Ernesto Vidal Sánchez

Pablo Andrey Rincón Rojas

Carlos David Alape Gamez

07 de septiembre de 2025

## 1. *Introducción*

El análisis de los determinantes salariales resulta esencial para comprender las dinámicas del mercado laboral y las desigualdades que persisten en él. En economías emergentes como la colombiana, las brechas asociadas al género, la informalidad y la educación inciden directamente en la calidad de vida de los hogares y en las oportunidades de movilidad social. Estudiar cómo influyen variables como la edad, el nivel educativo, la formalidad laboral, la ocupación y el tamaño de la empresa en los salarios no solo permite identificar patrones de desigualdad, sino que también aporta evidencia para la formulación de políticas públicas orientadas a reducir dichas disparidades.

En este contexto, la literatura ha documentado ampliamente la relación entre capital humano y salarios. De acuerdo con Becker (1964), la educación y la experiencia laboral son determinantes fundamentales de la productividad y, por ende, de los ingresos. Mincer (1974) incorpora esta visión en su modelo de ingresos, mostrando que la edad y su cuadrado permiten capturar trayectorias no lineales de los salarios a lo largo del ciclo de vida laboral. En Colombia, Núñez y Sánchez (1998) encuentran que cada año adicional de educación aumenta significativamente el salario, confirmando que la escolaridad constituye un factor clave en la explicación de las diferencias salariales.

Por otra parte, la literatura sobre género y mercado laboral ha evidenciado brechas persistentes en perjuicio de las mujeres. Tenjo y Bernat (2012) muestran que, incluso después de controlar por educación y experiencia, las mujeres colombianas perciben ingresos inferiores a los hombres, mientras que Nopo (2009) documenta que en América Latina dichas brechas se relacionan no solo con diferencias en capital humano, sino también con fenómenos de discriminación y segmentación ocupacional. A esto se suma la problemática de la informalidad: de acuerdo con el DANE (2024), cerca del 56 % de los trabajadores en Colombia se encuentran en el sector informal, siendo las mujeres las más afectadas. Maloney (1999) y Carneiro (2002) demuestran que la informalidad no solo implica menores salarios, sino también precarización laboral y menor acceso a protección social.

No obstante, gran parte de los estudios enfrentan limitaciones al restringirse a modelos lineales simples o al incluir variables que podrían inducir sesgos por endogeneidad, como el estrato

socioeconómico o la posición de jefe de hogar. A partir de ello surge la siguiente pregunta de investigación: ¿cómo influyen la edad, el género y otros determinantes laborales en la trayectoria salarial de los individuos ocupados en Bogotá y qué tan bien pueden predecirse sus ingresos mediante modelos econométricos y de machine learning?. El aporte central de este trabajo consiste en evaluar diferentes especificaciones de modelos, identificando qué variables explican mejor los salarios y qué combinación de controles mejora la capacidad predictiva, al tiempo que se discuten los riesgos de incluir controles problemáticos.

Los resultados muestran que la relación edad–salario sigue un patrón no lineal, con un pico salarial alrededor de los 40 a 44 años, mientras que las mujeres alcanzan su máximo de ingresos antes y con trayectorias más cortas que los hombres. Asimismo, se encuentra que, manteniendo constantes variables como educación, ocupación y horas trabajadas, las mujeres perciben en promedio un 25,9 % menos que los hombres. Por otra parte, la inclusión de controles sociodemográficos mejora de forma sustancial la precisión predictiva de los modelos: las especificaciones más simples presentan errores de predicción elevados, mientras que el modelo “género + controles” alcanza el mejor desempeño con un RMSE de 0,629, confirmando la importancia de considerar factores estructurales del mercado laboral más allá de las características básicas.

El presente taller se organiza de la siguiente manera: en la primera sección se describe la base de datos utilizada, así como los procesos de limpieza y construcción de variables; en la segunda se analizan los perfiles edad–salario y la brecha de género bajo distintos modelos econométricos; en la tercera se evalúa el desempeño predictivo de diversas especificaciones, incluyendo técnicas lineales, polinómicas y regularizadas.

## **2. *Datos***

### **2.1. Breve descripción de la base de datos**

Los datos provienen del informe de Medición de Pobreza Monetaria y Desigualdad 2018, elaborado por el DANE con información de la Gran Encuesta Integrada de Hogares (GEIH). Para este ejercicio se seleccionó un subconjunto de individuos residentes en Bogotá, mayores de 18 años y ocupados.

El propósito de usar estos datos es estimar un modelo de predicción de ingresos, con el fin de analizar los salarios individuales. Esto permite apoyar el cálculo de impuestos y reducir riesgos de fraude tributario por omisión de ingresos, además de identificar posibles condiciones de vulnerabilidad. Se incluyen variables de tipo sociodemográfico y laboral, tales como edad, género, nivel educativo, tipo de ocupación, estrato socioeconómico, si corresponde a jefe de hogar, actividad económica y número de horas trabajadas, entre otras, lo cual contribuye a un análisis más preciso de las diferencias entre individuos.

La GEIH es una encuesta implementada por el Departamento Administrativo Nacional de Estadística (DANE) desde el año 2006, tiene cobertura en las 32 capitales de departamentos y 443 municipios tanto en zonas urbanas como rurales. Lo anterior implica que la información de hogares recogida por la GEIH tiene representatividad nacional. La GEIH alimenta las estadísticas económicas de Colombia ya que de esta encuesta se obtiene información como: ingresos, pobreza

monetaria, comportamiento del mercado laboral y cambios demográficos en el país (DANE, 2025).

Para el año 2018 se tuvo un tamaño muestral de 248.028 hogares. Para cada uno de estos hogares la encuesta tiene asignado un factor de expansión que corresponde al peso o representatividad del universo investigado (DANE, 2018).

## 2.2. Proceso de adquisición de los datos

Los datos de la GEIH-2018 fueron obtenidos mediante un proceso de web scraping a la página de GitHub del doctor Ignacio Sarmiento [https://ignaciomsarmiento.github.io/GEIH2018\\_sample/](https://ignaciomsarmiento.github.io/GEIH2018_sample/). A través del lenguaje de programación R, con la librería *rvest*, construimos una función que descarga el contenido HTML de la página de la cuál queremos obtener la información, busca las tablas (*<table>*) en el primer nodo de la página y por último convierte la tabla en un objeto data frame. Posteriormente, con un loop *map\_dfr()*, del paquete purrr, realizamos una iteración sobre las 10 páginas que contienen las tablas de la GEIH de este trabajo.

La página de GitHub no tiene restricciones para robots (rastreadores) por lo que podemos concluir que podemos acceder a la información contenida allí.

## 2.3. Proceso de limpieza de los datos

La base de datos original, obtenida por web-scraping, tuvo los siguientes cambios:

1. **Edad:** filtramos la variable de edad para obtener únicamente las personas con edad igual o mayor a 18 años en Bogotá.
2. **Ocu:** seleccionamos a las personas que están clasificadas como ocupados (*ocu = 1*).
3. **totalHoursWorked:** filtramos el número de horas trabajadas, en una semana, para retirar las filas de las personas que tienen 0 horas de trabajo.
4. **horas\_mes\_Worked:** en esta nueva variable registramos el número de horas trabajadas en un mes. Se obtiene al multiplicar **totalHoursWorked** por cuatro.
5. **ingreso\_total:** es la suma entre el *impaes* (ingresos del trabajo principal reportado) y el *isaes* (ingresos de otros posibles trabajos).
6. **salario\_hora:** corresponde a la división entre **ingreso\_total** y **totalHoursWorked**. Representa el salario por hora de todos los individuos previamente filtrados.
7. **salario\_hora (NAs y ceros):** eliminamos todas las filas vacías (NAs) o iguales a cero del salario por hora previamente calculado.
8. **maxEducLevel (NAs):** eliminamos la única fila vacía que hay en la variables de máximo nivel educativo alcanzado por el individuo.

Tras el proceso de depuración, la base de datos final contiene 14.731 filas.

## 2.4. Descripción de las variables usadas en el análisis

### 2.4.1. Tabla descriptiva para variables numéricas

Tabla 1: Edad, salario por hora y logaritmo natural del salario por hora

Variable	N	Media	SD	Mín	Mediana	Máx	Asimetría	Curtosis
Edad	14 731	38.9	13.2	18.0	37.0	91.0	0.48	2.44
ln_wage	14 731	8.7	0.8	5.2	8.6	12.7	0.58	4.84
salario_hora	14 731	9189.3	14361	178.6	5208	312500	7.78	99.7
totalHoursWorked	14 731	47.6	15.1	1.0	48.0	130.0	0.17	5.52

Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Tabla de elaboración propia.

## 2.5. Análisis de variables numéricas

**Edad:** La edad promedio de la muestra es de 39 años, con una desviación estándar de 13 años. La edad mínima es de 18 años, teniendo en cuenta que solo se seleccionaron individuos mayores de edad, y la máxima es de 91 años. La asimetría positiva (0.48) sugiere que los datos están sesgados hacia la derecha, es decir, la mayoría de las edades se concentran en valores relativamente bajos mientras que algunos individuos de edad avanzada alargan la cola de la distribución. La curtosis (2.44), al ser cercana a 3, indica que la distribución se aproxima a una normal, sin evidencia marcada de colas pesadas ni valores extremos fuera de lo común.

**Logaritmo del salario por hora:** El logaritmo natural del salario presenta una media de 8.7 y una mediana de 8.6, con una ligera asimetría positiva (0.58). Esto implica que la distribución es más simétrica y menos influenciada por valores atípicos en comparación con el salario bruto por hora. En términos prácticos, el valor promedio del logaritmo (8.7) equivale aproximadamente a un salario de 6.000 COP por hora ( $\exp 8.7$ ). Este resultado muestra que, al transformar los datos, se obtiene una variable que refleja de manera más adecuada el salario típico en la muestra y que permite realizar inferencias más robustas.

**Salario por hora:** El salario promedio por hora en la muestra es de 9.189 pesos colombianos, con una desviación estándar de 14.361 COP. El valor mínimo registrado es de 178.6 pesos y la mediana es de 5.208 pesos, lo que refleja que la media está fuertemente influenciada por la presencia de valores extremos. El valor máximo, de 312.500 pesos por hora, confirma la existencia de observaciones atípicas. Esto se refleja en la asimetría (7.78) y curtosis (99.7), que son extremadamente elevadas y muestran que la distribución está altamente sesgada hacia la derecha y presenta colas muy pesadas. Este comportamiento justifica el uso del logaritmo natural como una transformación adecuada para el análisis.

**Total de horas trabajadas:** En promedio, los individuos trabajan 47.6 horas a la semana, cifra consistente con la jornada laboral legal vigente en Colombia en el año 2018. La desviación estándar es de 15 horas, lo que sugiere una alta variabilidad en el número de horas trabajadas. El valor mínimo registrado es de 1 hora y el máximo de 130 horas semanales, lo cual constituye un rango amplio y evidencia la presencia de valores atípicos. La curtosis (5.52), al ser mayor que 3, confirma

que la distribución presenta colas pesadas, es decir, que existen individuos con jornadas muy por encima o muy por debajo del promedio. Esto resulta coherente con la realidad del mercado laboral, en el que algunos trabajadores se desempeñan en condiciones atípicas de tiempo.

### 2.5.1. Tabla descriptiva para variables categóricas

Tabla 2: Estadísticas descriptivas de variables categóricas

Variable	N	# Categorías	Moda	Frecuencia Moda	Prop Moda
female	14 731	2	Hombre	7765	53 %
formal	14 731	2	formal	8889	60 %
maxEducLevel	14 731	6	No sabe, no informa	6100	41 %
estrato1	14 731	6	Estrato 2	6311	43 %

Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Tabla de elaboración propia.

## 2.6. Variables categóricas

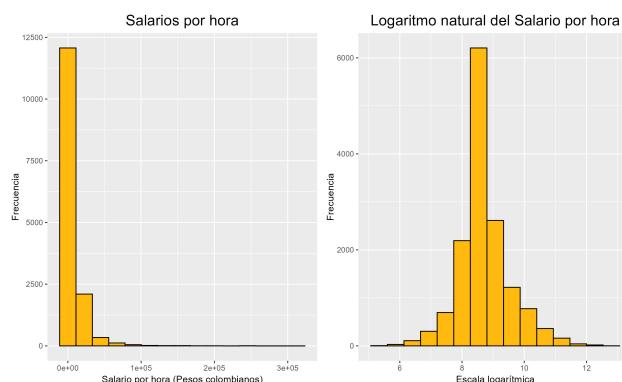
En cuanto a las variables categóricas, los resultados indican que en la muestra se observa una mayor participación de hombres (53 % ) frente a las mujeres. El 60 % de los individuos reporta tener un empleo formal, lo que implica que un 40 % se encuentra en la informalidad, proporción que resulta considerable y consistente con la alta incidencia de empleo informal en Colombia.

Respecto al nivel educativo, el 41 % de los individuos no reporta información sobre su máximo nivel alcanzado. En términos de estrato socioeconómico, el 43 % de la muestra pertenece al estrato 2, lo que refleja una mayor concentración en estratos bajos. Este resultado es coherente con la estructura de la población urbana en Colombia, donde los estratos bajos y medios-bajos concentran la mayor proporción de hogares.

### 2.6.1. Gráficas de las variables usadas en los modelos predictivos

#### 2.6.1.1 Variable dependiente: salarios por hora

Gráfica 1



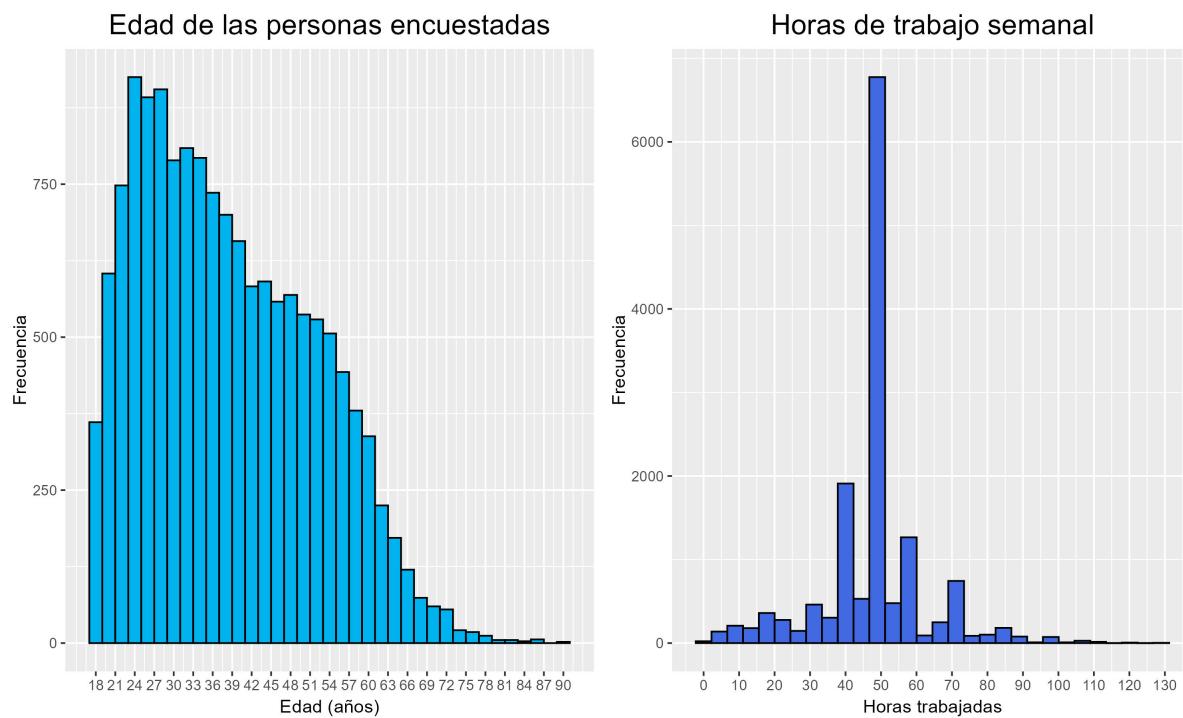
Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Gráficas de elaboración propia.

En la **Gráfica 1** tenemos los histogramas de los salarios por hora, medidos en pesos colombianos, y el logaritmo natural de los salarios por hora. En la primera gráfica podemos observar una distribución asimétrica a la derecha en la que la mayoría de los salarios por hora se concentran en valores bajos, esto lo podemos corroborar con las medidas de tendencia central y el coeficiente de asimetría de la **Tabla 1**. Además, hay presencia de valores atípicos, ubicados a la derecha del histograma, que representan unos pocos individuos que declararon tener elevados salarios para el año 2018.

El segundo histograma representa la transformación logarítmica de los salarios por hora. Esta gráfica tiene una forma aproximadamente simétrica, leptocúrtica (un pico pronunciado) y con una menor cantidad de datos atípicos respecto al primer histograma. La mayoría de datos se concentran en el centro de la gráfica, aproximadamente en el valor de 8.6, lo cual también podemos corroborar en la **Tabla 1** observando que los valores de la media y la mediana son similares.

### 2.6.1.2 Variables independientes

**Gráfica 2**

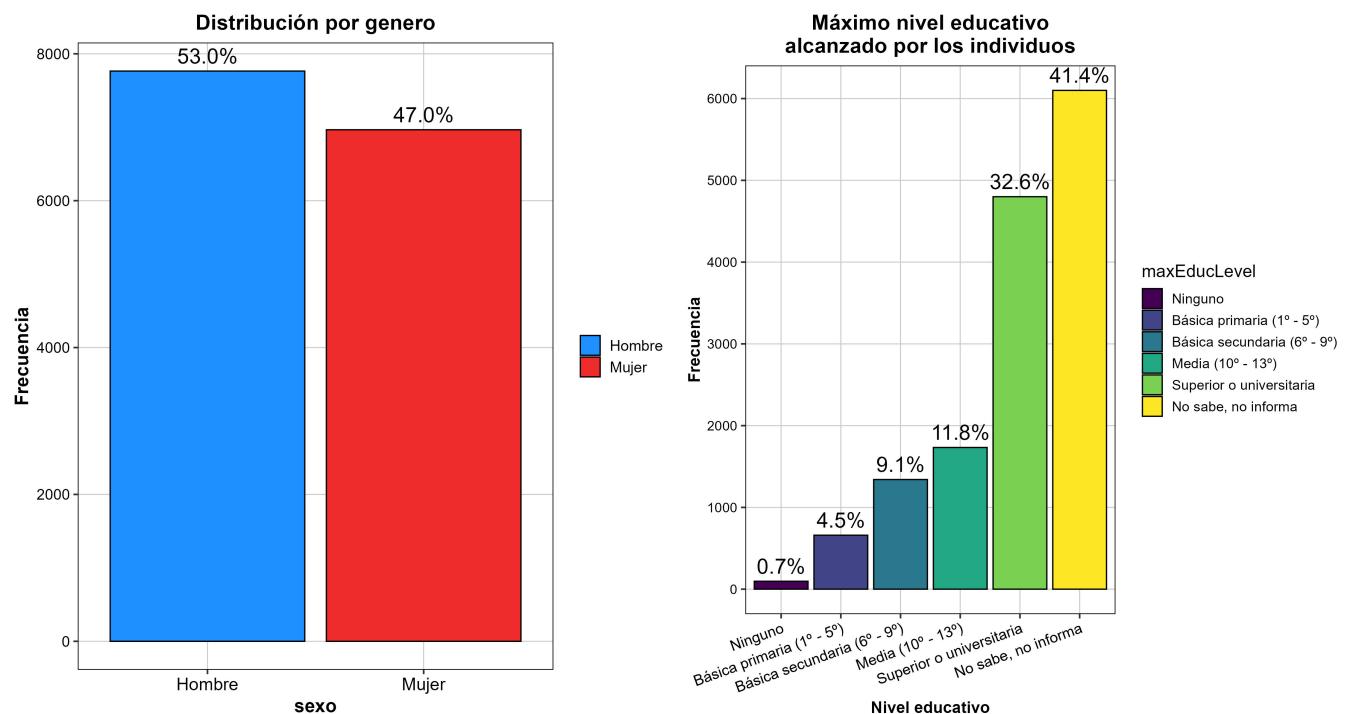


Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Gráficas de elaboración propia.

En la primera gráfica tenemos las edades de las personas encuestadas en nuestra base de la GEIH y que tienen residencia en Bogotá. Podemos observar que el histograma tiene una forma asimétrica a la derecha y que la mayor parte de individuos tienen entre 24 y 37 años reportados, por lo que podemos inferir que las personas de esta muestra pertenecen al bono demográfico. También vemos algunos datos atípicos de adultos mayores ( $> 70$  años) que reportan estar ocupados y recibiendo salario por sus actividades laborales.

El histograma de la derecha muestra el número total de horas que la persona dedica a su trabajo. Vemos que tiene una forma leptocúrtica, en donde los datos se concentran alrededor de 48 horas trabajadas en la semana. Además, hay una fuerte presencia de datos atípicos con personas que han reportado trabajar más de 60 horas semanales.

**Gráfica 3**



Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Gráficas de elaboración propia.

El primer diagrama de barras contiene las categorías de la variable *sex*: hombre y mujer. De esta gráfica inferimos que los hombres son la categoría con mayor proporción respecto a las mujeres en nuestra base de datos. Esto puede tener dos explicaciones: al quitar las filas con presencia de NAs, en la variable de salario por horas, es posible que muchas mujeres también fueran removidas de nuestra base de datos y en segundo lugar, al filtrar por la variable de ocupación, tendrímos más hombres ya que el desempleo afecta con mayor intensidad a las mujeres.

El segundo diagrama presenta el máximo nivel educativo que declararon tener los individuos encuestados en el año 2018. La mayoría de personas refirieron que no saben o no informan su educación, sin embargo, después están las personas con educación superior que representan el 32,6 % de la muestra. Además, la suma de las personas que tienen como máximo un título de bachillerato es de 26,1 % y esto puede tener origen en que las personas con mayores calificaciones académicas tienen mayor probabilidad de encontrar trabajo.

### 3. Perfil de edad y salario

En los tres modelos que estimamos observamos que los coeficientes de  $Age$  son positivos y estadísticamente significativos, mientras que los de  $Age^2$  resultan negativos. Esto confirma que el salario cambia de manera no lineal: al inicio de la vida laboral los ingresos crecen con la edad hasta alcanzar un salario máximo en determinado momento, a partir del cual comienzan a decrecer.

**Tabla 3:** Perfil edad–salario (comparativo)

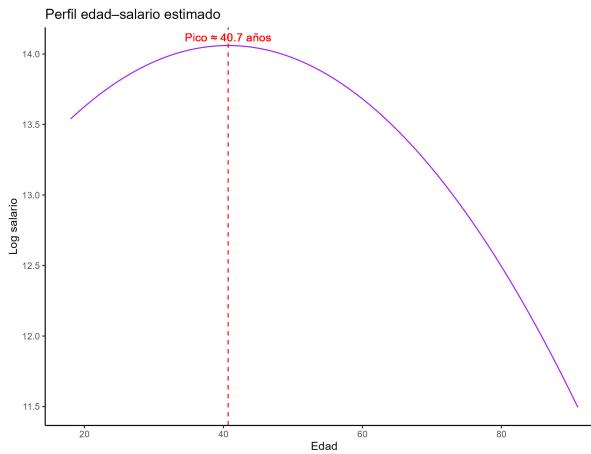
	Variable dependiente: Log salario		
	Básico	Con controles	Sin malos controles
	(1)	(2)	(3)
Edad	0,082*** (0,003)	0,051*** (0,002)	0,058*** (0,002)
Edad <sup>2</sup>	-0,001*** (0,00004)	-0,001*** (0,00003)	-0,001*** (0,00003)
Observaciones	14,732	14,731	14,731
R <sup>2</sup>	0.050	0.566	0.479
R <sup>2</sup> ajustado	0.050	0.565	0.479

Nota: \*  $p < 0,1$ ; \*\*  $p < 0,05$ ; \*\*\*  $p < 0,01$

En el modelo simple, en el que solo regresamos  $Age$  y  $Age^2$ , el coeficiente de edad refleja que un año adicional representa un aumento del 8,2 % en los salarios. Esta tendencia se sostiene hasta la edad de 40,7 años. Sin embargo, este modelo explica tan solo el 5 % de la variación en los salarios, lo cual posiblemente está asociado a la ausencia de otros determinantes relevantes que en este caso son variables omitidas.

$$\log(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

**Gráfica 4**

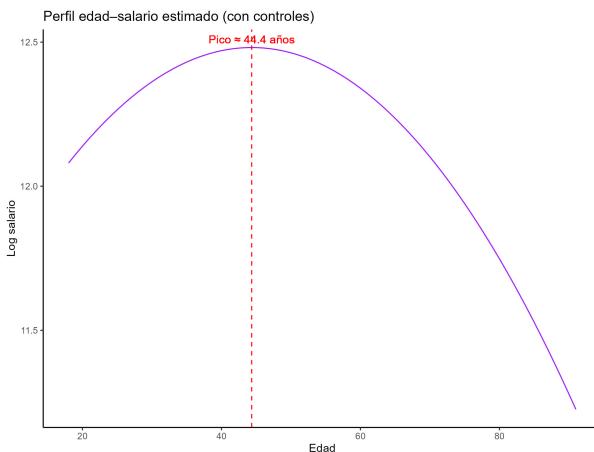


Relación salario–edad en el modelo simple. Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018.  
Elaboración propia.

Al incluir controles como horas trabajadas, género, formalidad, nivel educativo, relación laboral, jefe de hogar, estrato socioeconómico y tamaño de empresa, el coeficiente de  $Age$  cae a 5,1 %, lo que

indica que un año adicional de edad se traduce en dicho aumento salarial. Este cambio desplaza la edad máxima de crecimiento de ingresos a los 44,4 años, a partir de los cuales los salarios comienzan a decrecer. El modelo mejora sustancialmente en su bondad de ajuste, explicando el 56,6 % de la variación de los salarios.

**Gráfica 5**



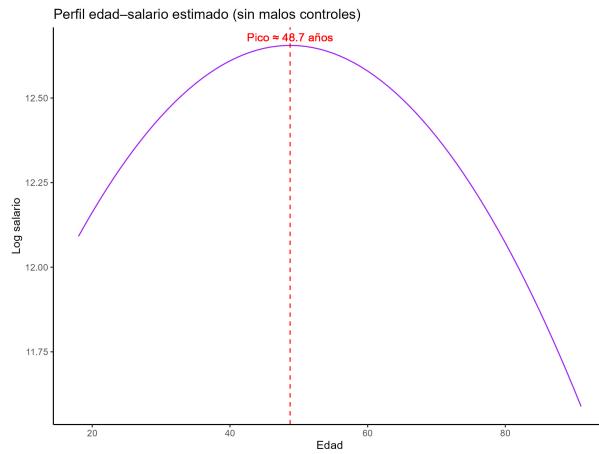
Perfil edad–salario con controles. Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018.  
Elaboración propia.

Es importante señalar que la elección de nuestros controles está asociada a la relación de cada una de estas variables con la variable dependiente, análisis que se realizó a partir del conocimiento intuitivo del mercado laboral. Así, el salario depende directamente de la cantidad de horas ofrecidas al mercado (*totalHoursWorked*); las brechas salariales entre hombres y mujeres están documentadas (*female*); el estatus formal o informal es un determinante claro de los ingresos (*formal*); la educación tiene efectos directos sobre los salarios en la medida en que un mayor nivel educativo puede incidir positivamente en ellos (*nivel\_educativo*); las diferentes modalidades de relación laboral conllevan estructuras salariales distintas (*relab*); y el tamaño de la empresa puede relacionarse con la capacidad del empleador de ofrecer mejores salarios en función de sus beneficios (*sizeFirm*).

No obstante, evidenciamos que dos de los controles usados pueden considerarse malos controles. Por un lado, el estrato socioeconómico puede estar correlacionado con el ingreso laboral, ya que el nivel salarial determina en gran medida el acceso a viviendas de ciertos estratos. Por otro lado, ser jefe de hogar más que un determinante puede ser un resultado del nivel de ingresos, dado que suele asociarse esa posición a quien más gana en el hogar. Al incluir estas dos variables, nuestro segundo modelo podría inducir sesgos por endogeneidad.

Al excluir estrato socioeconómico y jefe de hogar, el coeficiente de *Age* aumenta, de modo que en este modelo un año adicional de edad representa un incremento del 5,8 % en los salarios. Asimismo, la edad máxima de crecimiento salarial se ubica en 48,7 años. Este modelo reduce su bondad de ajuste, explicando el 47,9 % de la variación en los salarios.

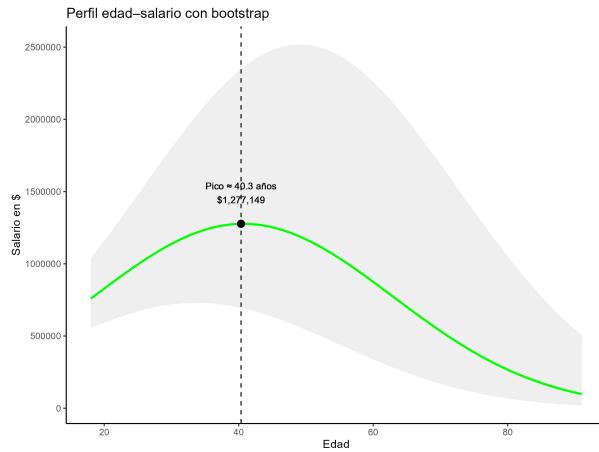
**Gráfica 6**



Perfil edad-salario excluyendo malos controles. Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Elaboración propia.

Finalmente, el *bootstrapping* con 10.000 repeticiones aplicado al modelo inicial confirmó la solidez de la relación entre salario y edad. La edad pico estimada mediante este procedimiento (40,34 años) es muy cercana a la obtenida directamente a partir de los coeficientes del primer modelo (40,7 años). Esto significa que, incluso bajo múltiples remuestreos, el patrón de nuestro primer modelo se mantiene estable y refuerza la evidencia de que alrededor de los 40 años se alcanza el salario máximo esperado en la trayectoria laboral, sin incluir los controles de los siguientes modelos.

**Gráfica 7**



Perfil edad-salario estimado con *bootstrapping* (10.000 repeticiones) y punto de máximo salarial.  
Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Elaboración propia.

## 4. La brecha salarial entre géneros

En nuestro primer modelo que relaciona la variable categórica de género, siendo la categoría mujeres 1 y comparativa con hombres

$$\log(w) = \beta_1 + \beta_2 \text{Female} + u$$

se evidencia que las mujeres ganan en este modelo 23.2 % menos que los hombres. Sin embargo, es de aclarar que este modelo explica tan solo un 1.7 %, posiblemente asociado a problemas de variable omitida.

En nuestro segundo modelo incorporamos variables de control como condiciones iguales entre hombres y mujeres, buscando desvirtuar que a trabajos iguales corresponde igual remuneración. Dentro de los controles usados retomamos los ya definidos anteriormente, excluyendo los malos controles que ya identificamos: el estrato y ser jefe de hogar. Además, incluimos dos controles más: por un lado, el oficio, dado que la ocupación determina ingresos, y la variable categórica *microEmpresa*, que define a las personas que trabajan o no en una empresa de estas características. Para esta variable, el conocimiento intuitivo explica que, al igual que el tamaño de la empresa, las empresas más pequeñas pagan menos.

$$\begin{aligned} \log(w) = & \beta_1 + \beta_2 \text{Female} + \beta_3 \text{nivel\_educativo} \\ & + \beta_4 \text{formal} + \beta_5 \text{oficio} + \beta_6 \text{relab} \\ & + \beta_7 \text{totalHoursWorked} + \beta_8 \text{sizeFirm} \\ & + \beta_9 \text{age} + \beta_{10} \text{agesqr} + \beta_{11} \text{microEmpresa} \end{aligned} \quad (1)$$

Al incorporar los controles, el coeficiente de *female* aumenta a -0.259, sugiriendo que, manteniendo constantes las demás características de nuestros controles, las mujeres ganan 25.9 % menos que los hombres. Este modelo aumenta su bondad de ajuste explicando el 49.8 % de la variación de los salarios. Entre los controles que resultaron significativos en este modelo, el nivel educativo (0.273), la formalidad laboral (0.364) y la ocupación (0.436) se asocian positivamente con el salario. Por otro lado, las horas totales de trabajo (-0.369) y el tamaño de empresa medido de manera continua (-0.005) presentan efectos negativos, lo que puede relacionarse con condiciones precarias o menores ingresos por hora en ocupaciones de largas jornadas o empresas pequeñas.

Finalmente, al utilizar el teorema FWL para estimar la brecha neta de género, controlando por las mismas variables que en el modelo anterior, el coeficiente así como el error estándar son exactamente los mismos, teniendo una disminución en el  $R^2$ . Este modelo explica el cambio de salario en un 3.6 %, lo cual posiblemente está asociado a la evaluación que hacemos solo de la relación entre residuos, no del modelo completo. El resultado obtenido en nuestro segundo modelo, así como el estimado aplicando el teorema de FWL, muestra consistencia: al hacer un bootstrap de este modelo en 10.000 repeticiones, el coeficiente de la variable *female* se mantiene en -0.259 y el error estándar en 0.011.

Si analizamos nuestro modelo inicial de edad–salario diferenciado para hombres y mujeres

$$\log(w) = \beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + u$$

se observa que la edad en la que se alcanza el salario máximo difiere significativamente entre géneros. En el caso de los hombres, el pico se alcanza a los 43,7 años, mientras que para las

**Tabla 4:** Comparación de modelos de brecha salarial

	<i>Variable dependiente:</i>		
	ln_wage (Sin controles)	ln_wage (Con controles)	res_salario (FWL)
Mujer	-0,232*** (0,014)	-0,259*** (0,011)	
res_mujer			-0,259*** (0,011)
Variable dependiente	Log salario	Log salario (con controles)	Residuales del salario
Observaciones	14,732	14,731	14,731
R <sup>2</sup>	0.017	0.498	0.036
R <sup>2</sup> ajustado	0.017	0.497	0.036

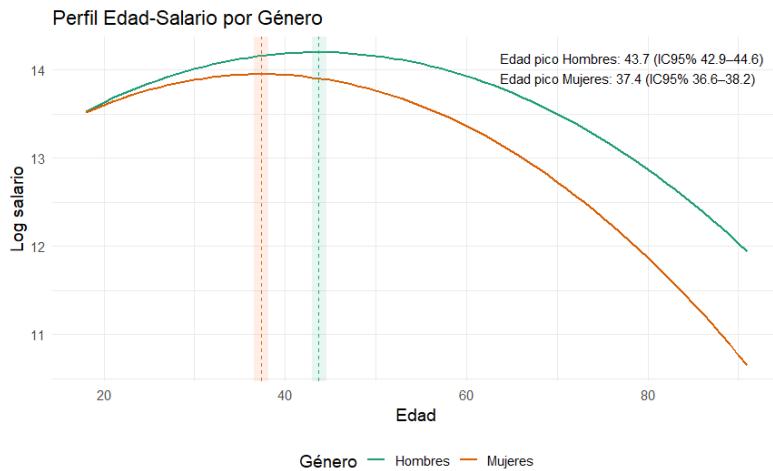
*Nota:* \* $p < 0,1$ ; \*\* $p < 0,05$ ; \*\*\* $p < 0,01$

mujeres ocurre de manera anticipada, a los 37,4 años. Esta diferencia de 6,3 años evidencia que la trayectoria salarial de las mujeres es más corta, es decir, alcanzan su techo de ingresos antes y enfrentan un descenso más temprano en comparación con los hombres.

A esto se suma que, de acuerdo con nuestros modelos anteriores, las mujeres perciben en promedio un 25,9 % menos que los hombres, incluso manteniendo constantes factores como educación, ocupación u horas trabajadas. Esto refuerza la idea de que la brecha salarial de género no puede explicarse únicamente por un problema de selección. En un escenario de pura selección, cabría esperar que la brecha desapareciera una vez controladas variables como el oficio, pues la diferencia se debería a la sobrerepresentación de mujeres en ocupaciones o sectores de menores ingresos. Sin embargo, el hecho de que la brecha persista e incluso se amplíe al incluir controles apunta a la existencia de un componente de discriminación, entendido como diferencias en la remuneración que no se explican por las variables observadas.

No obstante, es importante reconocer que ciertos factores asociados al rol social de las mujeres sí pueden influir en la selección laboral, como la maternidad, las responsabilidades de cuidado o las interrupciones en la trayectoria profesional.

**Gráfica 8**



Perfil edad–salario diferenciado por género. Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018.  
Elaboración propia.

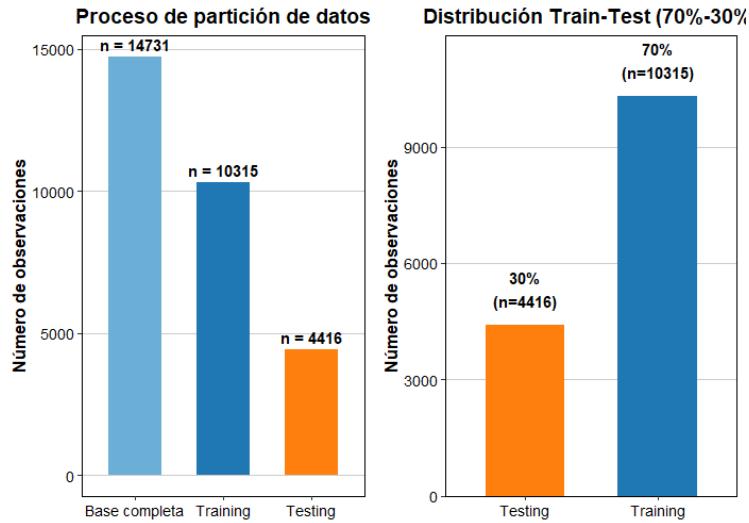
## 5. Predicción de salarios

### 5.1. Muestra de entrenamiento y prueba

En primer lugar, se procedió a dividir la base de datos en dos subconjuntos: un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %). Esta partición garantiza que el modelo pueda ajustarse utilizando una parte representativa de la muestra, mientras que la porción reservada se emplea exclusivamente para evaluar su capacidad predictiva. El procedimiento permitió obtener una distribución equilibrada y verificar que no existieran valores atípicos o inconsistencias en la variable dependiente, lo que asegura mayor validez en las comparaciones posteriores de desempeño entre modelos.

La intuición detrás de este proceso es que entrenar y evaluar con los mismos datos conduce a un sobreajuste, es decir, un modelo con bajo error en la muestra original pero con baja capacidad de generalización. Al reservar un conjunto de prueba independiente, se logra una medida más realista del error de predicción. Los resultados mostraron una separación clara entre entrenamiento y prueba, con proporciones muy cercanas a las esperadas (70 % y 30 %), permitiendo avanzar con modelos robustos y comparables en términos de su error cuadrático medio (RMSE).

**Gráfica 9**



Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Gráficas de elaboración propia.

Adicionalmente, se generaron representaciones gráficas del proceso de partición, lo que permitió visualizar tanto la distribución proporcional entre entrenamiento y prueba como la secuencia de división desde la base completa. Estas gráficas son útiles porque hacen evidente la correspondencia entre los porcentajes teóricos y los datos efectivos, además de facilitar la comunicación de los resultados. En conjunto, este paso metodológico sienta las bases para la posterior estimación de modelos comparativos y la aplicación de técnicas de validación cruzada, reforzando la transparencia y robustez del análisis predictivo.

## 5.2. Comparación de modelos y desempeño predictivo

Para evaluar la capacidad predictiva de los salarios se estimaron nueve modelos alternativos, que combinan relaciones lineales, controles sociodemográficos y extensiones no lineales. Los primeros dos modelos se enfocan en la relación edad–salario: (1) un modelo simple lineal con edad y edad al cuadrado; y (2) la misma relación, pero adicionando controles de horas trabajadas, género, formalidad, nivel educativo, posición laboral y tamaño de la empresa. Los modelos (3) y (4) capturan la brecha salarial por género, primero en forma simple y luego incorporando controles adicionales.

Posteriormente, se implementaron modelos más flexibles para explorar posibles no linealidades. El modelo (5) utiliza un polinomio cúbico de la edad; el modelo (6) considera la interacción entre edad y género; mientras que el modelo (7) explora la interacción entre edad (en forma cuadrática) y nivel educativo. Finalmente, el modelo (8) emplea *splines* en la edad para permitir mayor flexibilidad en la pendiente, y el modelo (9) aplica una versión regularizada de ridge, que controla el sobreajuste al penalizar la magnitud de los coeficientes.

Los resultados (Tabla 5 y Grafica 10) muestran que los modelos más simples, como edad–salario y género, son los menos precisos ( $RMSE \approx 0,86$ ). En contraste, los modelos con controles reducen sustancialmente el error ( $RMSE \approx 0,63$ – $0,64$ ). Los modelos no lineales y regularizados alcanzan

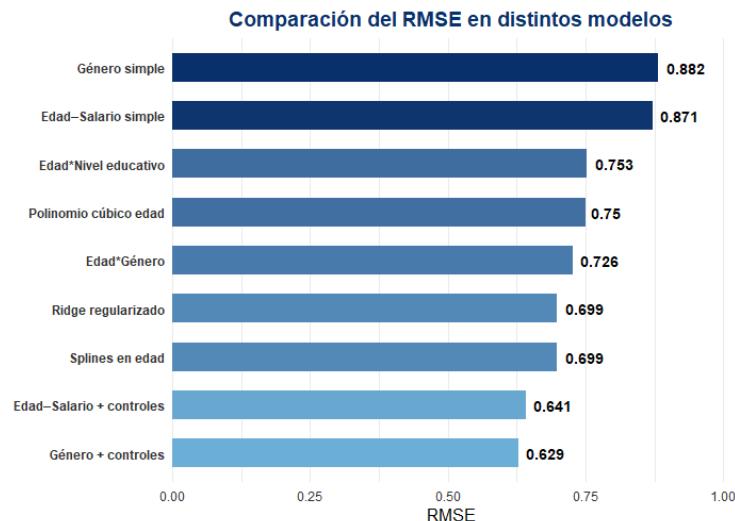
mejoras moderadas (RMSE entre 0,70 y 0,75), pero no superan el desempeño de los modelos con controles. En conclusión, la inclusión de características observables resulta más relevante que aumentar la complejidad funcional del modelo.

**Tabla 5:** Especificaciones de los modelos y resultados de RMSE

Modelo	Forma funcional	RMSE
1. Edad–Salario simple	$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \epsilon$	0.871
2. Edad–Salario + controles	$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 X + \epsilon$	0.641
3. Género simple	$\ln(wage) = \beta_0 + \beta_1 female + \epsilon$	0.882
4. Género + controles	$\ln(wage) = \beta_0 + \beta_1 female + \beta_2 X + \epsilon$	0.628
5. Polinomio cúbico edad	$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 Z + \epsilon$	0.750
6. Edad*Género	$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \beta_4 (age * female) + \beta_5 W + \epsilon$	0.726
7. Edad*Nivel educativo	$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 nivel\_educativo + \beta_4 (age * nivel\_educativo) + \epsilon$	0.753
8. Splines en edad	$\ln(wage) = \beta_0 + f(age) + \beta_1 female + \beta_2 formal + \beta_3 nivel\_educativo + \epsilon$	0.700
9. Ridge regularizado	$\ln(wage) = \beta_0 + f(age) + \beta_1 female + \beta_2 formal + \beta_3 nivel\_educativo + \epsilon, \min \ y - X\beta\ ^2 + \lambda\ \beta\ ^2$	0.700

La gráfica presenta la comparación del error cuadrático medio (RMSE) entre distintas especificaciones de modelos de predicción salarial. Se observa que las versiones más simples, como el modelo de género y el de edad-salario sin controles, presentan los errores más altos, lo que evidencia su limitada capacidad explicativa. En contraste, al incluir controles adicionales o formas funcionales más flexibles, como polinomios, splines o regularización ridge, el RMSE disminuye de manera importante, alcanzando sus valores más bajos en los modelos con género y controles (0.629) y edad-salario con controles (0.643). Esto demuestra que la incorporación de mayor complejidad y factores explicativos mejora la precisión predictiva del modelo.

**Gráfica 10**



Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Gráficas de elaboración propia.

### 5.3. Discusión sobre los resultados

#### (i) Desempeño general de los modelos

El desempeño de las distintas especificaciones muestra que los modelos más simples, como *Edad-Salario simple* y *Género simple*, presentan errores de predicción elevados, con valores de RMSE cercanos a 0.87. En contraste, al incorporar controles adicionales y formas funcionales más flexibles, el error disminuye de manera notable, alcanzando valores de RMSE entre 0.63 y 0.70. Esto confirma que la inclusión de covariables relevantes —como educación, formalidad laboral o tamaño de la empresa— permite capturar mejor la heterogeneidad en los salarios, mientras que las formas demasiado restrictivas resultan insuficientes para reflejar la variabilidad real de los datos.

#### (ii) Especificación con menor error de predicción

El modelo con mejor desempeño fue *Género + controles*, con un RMSE de 0.6286. Este resultado indica que las diferencias salariales no pueden explicarse únicamente por características demográficas básicas, sino que requieren la incorporación de determinantes estructurales del mercado laboral. El hecho de que esta especificación supere tanto a modelos más simples como a otros más complejos sugiere que existe un balance adecuado entre parsimonia y capacidad predictiva.

**Tabla 5.** Modelo con menor error de predicción

Especificación	RMSE
Género + controles	0.6288

Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Resultados de elaboración propia.

#### (iii) Distribución de errores y observaciones atípicas

El análisis de los errores de predicción del mejor modelo muestra que la mayoría de las observaciones se concentran en torno a cero, lo que refleja un buen ajuste global. No obstante, se detectaron 206 observaciones como outliers, identificadas mediante el criterio del rango intercuartílico (IQR). Estas corresponden a individuos con salarios significativamente diferentes a los estimados, lo que se refleja en colas pronunciadas de la distribución del error.

**Tabla 6.** Ejemplos de outliers detectados en el modelo ganador

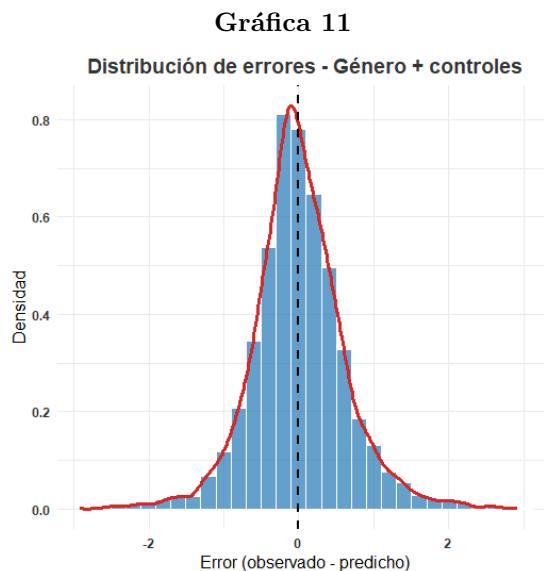
Salario observado (ln)	Salario predicho (ln)	Error
11.00	13.07	-2.07
12.61	14.37	-1.76
16.70	14.81	1.90
15.20	13.53	1.68
10.60	12.95	-2.36

Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Resultados de elaboración propia.

La interpretación de estos outliers es doble: por un lado, pueden reflejar situaciones reales y atípicas del mercado laboral (salarios muy altos o muy bajos en comparación con la media); por otro,

pueden ser consecuencia de limitaciones en la especificación del modelo. En el marco de políticas públicas, estos casos podrían considerarse de interés para la DIAN, pero se recomienda un análisis complementario antes de concluir que se trata de inconsistencias tributarias.

La gráfica de la distribución de errores del modelo “Género + controles”, que presentó el menor RMSE (0.6288), muestra un comportamiento bastante simétrico alrededor de cero, lo cual indica que el modelo no presenta sesgos sistemáticos en la predicción: en promedio, no sobrestima ni subestima los salarios de manera consistente. La curva de densidad en rojo refuerza esta normalidad aproximada, aunque se observan colas más largas, lo que evidencia la presencia de outliers. En efecto, se identificaron 206 observaciones atípicas, correspondientes a casos en los que el modelo predijo salarios muy diferentes a los observados. Estos valores extremos pueden responder tanto a individuos con características inusuales en el mercado laboral como a limitaciones del modelo para capturar comportamientos no lineales. En cualquier caso, el patrón general refleja que el modelo logra una predicción robusta para la mayoría de la muestra, aunque con cierta fragilidad frente a datos atípicos.



Fuente: Gran Encuesta Integrada de Hogares (GEIH) 2018. Gráficas de elaboración propia.

## 5.4. Validación Leave-One-Out-Cross-Validation (LOOCV)

Para evaluar la robustez de los modelos con mejor desempeño en el conjunto de prueba, se aplicó la técnica de *Leave-One-Out Cross-Validation* (LOOCV). Esta metodología consiste en entrenar el modelo con todas las observaciones menos una, predecir para la observación excluida y repetir el proceso hasta recorrer toda la muestra. Su ventaja es que utiliza la máxima cantidad de datos para entrenamiento en cada iteración, lo que reduce la varianza de la estimación del error predictivo en comparación con una única partición *train-test*.

**Tabla 8:** Comparación del error de predicción en Test set y LOOCV

Modelo	RMSE Test	RMSE LOOCV
Género + controles	0.6286	0.6227
Edad–Salario + controles	0.6415	0.6333

Los resultados de la Tabla 8 muestran que los valores de RMSE bajo LOOCV son muy cercanos a los obtenidos en el *validation set approach*, con ligeras reducciones en ambos casos. Esto indica que los modelos son estables y no dependen fuertemente de una partición específica de los datos. En particular, el modelo *Género + controles* sigue siendo el más preciso, confirmando su capacidad de generalización.

Finalmente, el vínculo con la estadística de influencia es relevante: LOOCV penaliza de manera más visible a las observaciones influyentes, ya que cada punto se convierte en el “caso de prueba” en algún momento. En consecuencia, las diferencias entre RMSE en test y LOOCV reflejan la sensibilidad del modelo a outliers e influencias individuales. La cercanía de ambos valores sugiere que, aunque se identificaron observaciones atípicas previamente, su impacto en el error promedio no es lo suficientemente grande como para alterar el ranking de desempeño de los modelos.

Enlace al repositorio de GitHub: [GitHub - Problem Set 1 Predicting Income](#)

## Referencias

- [1] Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.
- [2] Carneiro, F. G. (2002). The changing pattern of wage determination in Brazil. *World Bank Policy Research Working Paper Series*, (2754).
- [3] DANE. (2018). *Colombia - Gran Encuesta Integrada de Hogares (GEIH) - 2018*. Recuperado de <http://microdatos.dane.gov.co/index.php>
- [4] DANE. (2025). *Colombia - Gran Encuesta Integrada de Hogares (GEIH) - 2025*. Recuperado de <http://microdatos.dane.gov.co/index.php>
- [5] Maloney, W. F. (1999). Does informality imply segmentation in urban labor markets? Evidence from sectoral transitions in Mexico. *World Bank Economic Review*, 13(2), 275–302. <https://doi.org/10.1093/wber/13.2.275>
- [6] Mincer, J. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research; Columbia University Press.
- [7] Núñez, J., & Sánchez, F. (1998). *Educación y salarios en Colombia*. Archivos de Macroeconomía, Departamento Nacional de Planeación.

- [8] Nopo, H. (2009). *The gender wage gap in Latin America: Evidence from the 1990s*. Inter-American Development Bank.
- [9] Tenjo, J., & Bernat, L. F. (2012). Brechas salariales de género en Colombia: Una aproximación con descomposiciones de Oaxaca-Blinder. *Revista de Economía del Rosario*, 15(2), 123–157.