Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

El impacto del incremento salarial en México: un análisis con regresión distributiva

---

TESIS para obtener el **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Tesis presentada por:
María Alejandra Leyva Gómez

Asesor de tesis:
Dra. María del Rosario Ruíz Hernández

Tlaquepaque, Jalisco, Julio, 2025

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Formato de Aprobación de la Maestría en Ciencia de Datos

Título de la Tesis: **El impacto del incremento salarial en México: un análisis con regresión distributiva**

Autor(a): **María Alejandra Leyva Gómez**

Tesis aprobada para cumplir con todos los requisitos del grado de Maestría en Ciencia de Datos.

Asesor(a) de tesis, **Dra. María del Rosario Ruíz Hernández**

Lector(a) de tesis, **Dr. Leonardo Adalberto Gática Arreola**

Lector(a) de tesis, **Dra. Alma Nayeli Rodriguez Vazquez**

Asesor(a) académico, **Dra. Rocio Carrasco Navarro**

Tlaquepaque, Jalisco, Julio, 2025

El impacto del incremento salarial en México: un análisis con regresión distributiva

María Alejandra Leyva Gómez

Resumen

Esta investigación analiza el impacto del aumento al salario mínimo en la distribución del ingreso salarial en México, a partir de las políticas implementadas durante el último sexenio. Para ello, se empleó la metodología de regresión distributiva, generando funciones de distribución acumulada (CDF) del ingreso por hora para 2018 y 2024, y construyendo una distribución contrafactual que aísla el efecto del salario mínimo.

Los resultados revelan un efecto positivo en toda la distribución, con mayores aumentos en los percentiles bajos y altos, y aumentos más moderados en los segmentos medios. Esto contradice la idea de que el efecto se limita a los sectores más vulnerables, y confirma un impacto heterogéneo pero generalizado.

Adicionalmente, se imputaron los ingresos faltantes, que representan casi el 30 % de los datos, utilizando modelos de machine learning. El XGBoost Regressor ofreció el mejor desempeño. La imputación incrementó significativamente el ingreso promedio, lo que confirma que los datos faltantes no son aleatorios.

En conjunto, este trabajo muestra cómo herramientas de Ciencia de Datos pueden fortalecer el análisis económico, proporcionando estimaciones más precisas y revelando efectos que de otro modo podrían permanecer ocultos.

Índice general

Índice de cuadros

1 Introducción

Contenidos

1.1 Contexto

El aumento del salario mínimo en México ha sido un tema de discusión relevante, especialmente a partir de las políticas públicas implementadas por el Gobierno Federal en 2019. Durante la década anterior, los ingresos laborales en el país mostraron una caída significativa, con el ingreso mediano disminuyendo drásticamente en términos reales desde 2007 [1]. En particular, tras el aumento del salario mínimo en 2018, la pobreza disminuyó entre 2.6 % y 3 % en la frontera norte del país. Sin embargo, aunque la pobreza general se redujo, la intensidad de la pobreza en la misma región aumentó [2]. Este fenómeno resalta la complejidad de los efectos de las políticas salariales, que pueden tener impactos variados dependiendo del contexto.

Desde 2019, el salario mínimo en México ha experimentado incrementos significativos con el propósito de mejorar el poder adquisitivo de los trabajadores y garantizar un nivel de vida más digno. Es a partir de este año que se implementaron dos zonas salariales diferenciadas: el Salario Mínimo General (SMG) para el resto del país y el Salario Mínimo de la Zona Libre de la Frontera Norte (ZLFN), diseñado para atender las necesidades específicas de esta región colindante con los Estados Unidos de Norte América. De manera adicional, es importante mencionar que estas nuevas políticas salariales coinciden o responden a un cambio en el régimen de gobierno

[1] Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320): 803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003

[2] Raymundo M. Campos-Vazquez and Gerardo Esquivel. The effect of the minimum wage on poverty: Evidence from a quasi-experiment in mexico. *The Journal of Development Studies*, 59(3):360–380, 2023. DOI: 10.1080/00220388.2022.2130056

en el país. A continuación, se mencionan las entidades que componen esta zona diferenciada:

Figura 1.1: Zona Libre de la Frontera Norte

En cuanto a los Salarios Mínimos publicados por la CONASAMI (Comisión Nacional de los Salarios Mínimos), organismo encargado de su revisión, publicación y promoción, para cada año, iniciando en el 2017 a fines de comparación son [3]:

Año	Salario General	Salario ZLFN
2017	$80.04 (9.58 %)	N/A
2018	$88.36 (10.4 %)	N/A
2019	$102.68 (16.2 %)	$176.72 (70 % más alto)
2020	$123.22 (20 %)	$185.56 (5 %)
2021	$141.70 (15 %)	$213.39 (15 %)
2022	$172.87 (22 %)	$260.34 (22 %)
2023	$207.44 (20 %)	$312.41 (20 %)
2024	$241.14 (16.6 %)	$374.44 (16.6 %)

Cuadro 1.1: Evolución de los salarios generales y ZLFN (2017-2024)

En las gráficas siguientes se muestra la evolución del salario y cómo es que a partir de 2019, se implementó el salario diferenciado en la frontera.

[3] Comisión Nacional de los Salarios Mínimos (CONASAMI). Evolución del salario mínimo. https://www.gob.mx/conasami/documentos/evolucion-del-salario-minimo?idiom=es, n.d. Accedido en junio de 2025

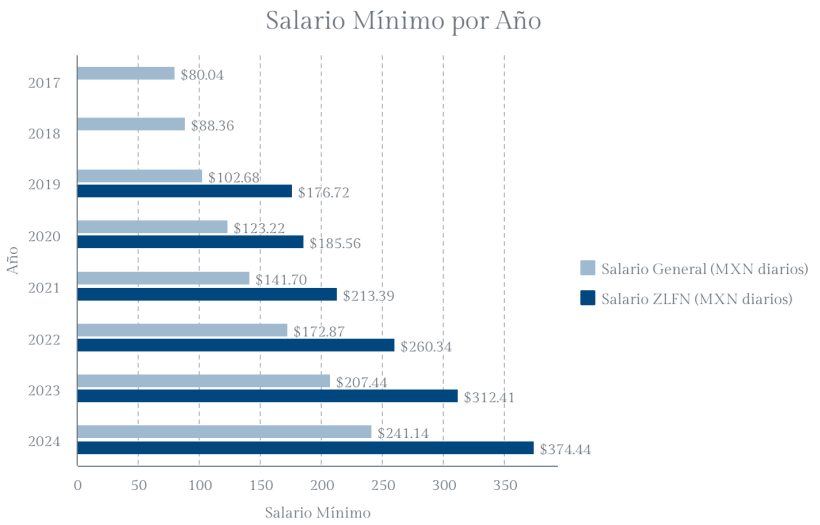

Figura 1.2: Salario Mínimo por Año e Introducción de Salario Mínimo para Frontera Norte

1.2 *Justificación*

La desigualdad de ingresos está relacionada con la baja movilidad social en el país, donde el 50.2 % de los hijos de padres en el quintil más bajo permanece en esa situación, y solo el 2.1 % logra escalar al quintil más alto [4], por lo cual la desigualdad sigue siendo un tema relevante.

En consecuencia, evaluar cómo los aumentos en el salario mínimo han afectado la desigualdad, es necesario para determinar si esta política pública han sido efectiva en mejorar la equidad social y económica en México, o sus efectos se han quedado en la reducción moderada de la pobreza.

De forma adicional, en el caso de que haya existido un impacto en la desigualdad, será importante determinar si este impacto es consecuencia directa de la aplicación de la política social o el cambio observado es unicamente consecuencia de cambios en la estructura de la población.

Por lo tanto, la investigación que se propone es relevante y contribuye positivamente a la literatura que busca explicar la desigualdad de ingresos en México.

[4] Raymundo Campos. Movilidad social, empleo e ingresos laborales en méxico. Technical report, Centro de Estudios Económicos, El Colegio de México, 2021. URL https://movilidadesocial.colmex.mx/wp-content/uploads/2021/10/5.-Raymundo-Campos.pdf. ENOE & IMSS data used

1.3 *Problema*

La forma más común de medir la desigualdad es la propuesta por Corrado Gini en 1912 con su obra "Variabilità e Mutabilità", conocida como el Coeficiente de Gini. Dentro de las razones para seguir utilizando esta métrica al día de hoy destacan su facilidad de interpretación y la capacidad de resumir la desigualdad en un solo

número. Así, el 0 representaría un escenario de perfecta igualdad, mientras que 1 representa la perfecta desigualdad [5]. De forma similar, existen numerosos índices o ratios que nos ayudan a medir la desigualdad del ingreso y hacer posible la comparación entre períodos o países. No obstante, uno de los mayores limitantes de estos coeficientes es que no permiten analizar en qué partes de la distribución se focaliza la desigualdad, es decir, no brinda una comparación de los diferentes estratos poblacionales.

Considerando el problema anterior, el presente trabajo propone utilizar una metodología de Regresión Distributiva, que permita capturar la distribución completa del ingreso en México. El utilizar este tipo de metodologías brinda flexibilidad en el análisis al hacer comparables los distintos estratos que componen la distribución, pero también al poder comparar diferentes años y descomponer sus cambios identificando variaciones en el precio (propiamente en el salario) y en la estructura de la población con ingresos. De forma adicional, esto permitirá identificar fenómenos como el efecto faro y otros que puedan generarse alrededor del Salario Mínimo.

La fuente de información más importante para medir los efectos del incremento en el salario en México es, sin duda, la Encuesta Nacional de Ocupación y Empleo (ENOE), la cual tiene una parte importante de registros incompletos, ya que aproximadamente 30 % de los trabajadores deciden no reportar su ingreso al encuestador. Por lo tanto, para poder trabajar de manera adecuada con la información que proporciona la ENOE, es necesario aplicar técnicas de Ciencia de Datos para preprocesar y modelar los datos faltantes, asegurando que los análisis posteriores sean representativos y confiables.

En muchos estudios, los registros con datos incompletos o faltantes se eliminan sin realizar las correcciones necesarias, lo que puede llevar a resultados sesgados y a una subestimación de indicadores clave. Por ejemplo, un estudio realizado en 2013 [6] demostró que no ajustar por ingresos no reportados resultaba en una subestimación del ingreso individual en un 4 %, y que la pobreza laboral se sobreestimaba en un 25 %. Esto evidencia un segundo problema y subraya la necesidad de aplicar métodos de análisis de datos avanzados, propios de la Ciencia de Datos, para imputar ingresos no reportados y poder realizar un análisis riguroso del impacto del incremento salarial en la distribución del ingreso.

En este sentido, para resolver este problema, en este trabajo se propone el uso de técnicas de preprocesamiento avanzado y modelado predictivo, propios de la ciencia de datos, que pueden mejorar la predicción de ingresos faltantes y, con ello, generar una mayor precisión en el análisis del impacto del incremento salarial en la distribución de los salarios.

[5] Joe Hasell. Measuring inequality: what is the gini coefficient? https://ourworldindata.org/what-is-the-gini-coefficient, 2023. Online resource, accessed June 2025

[6] Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320): 803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003

1.4 Objetivos

1.4.1 Objetivo General

El objetivo de esta investigación es **analizar el impacto** del aumento del salario mínimo en la **distribución del ingreso** salarial en México, a partir de las políticas implementadas durante el último sexenio. Para alcanzar el objetivo general, se explorará cómo la ciencia de datos, a través de las técnicas de preprocesamiento avanzado y modelado predictivo, puede mejorar la precisión del análisis.

1.4.2 Objetivos Particulares

1. Construir y curar una base de datos completa y representativa de la población ocupada en México, incluyendo la imputación de ingresos faltantes mediante modelos predictivos, para garantizar la calidad de los datos y la robustez del análisis.

2. Modelar las distribuciones del ingreso salarial en México, a través de la metodología de Regresión Distributiva, para el último trimestre del 2018 y el segundo trimestre del 2024.

3. Descomponer el impacto del incremento salarial a partir del 2019 en la distribución del ingreso, contra el año 2024, identificando tanto el ocasionado por el aumento del Salario Mínimo, como por la estructura poblacional que percibe salarios.

1.5 Aportación

Esta tesis realiza una aplicación de Ciencia de Datos a un problema económico concreto. El trabajo de investigación que se desarrolla contribuye con dos aportaciones a la literatura existente:

1. La implementación de técnicas de preprocesamiento avanzado y modelado predictivo (con énfasis en el modelo Extreme Gradient Boost Regression) para imputar ingresos no reportados en la ENOE, mejorando la calidad y representatividad de los datos. Esto constituye un ejemplo práctico de cómo la Ciencia de Datos puede aplicarse para generar conocimiento útil en un contexto social y económico, pues hasta donde se tiene conocimiento, no se ha documentado un estudio que aplique esta técnica específica para la imputación de ingresos en México, lo que constituye una aportación novedosa dentro de la investigación económica nacional.

2. El análisis del impacto del incremento del salario mínimo en la distribución del ingreso mediante regresión distributiva, utilizando

los datos imputados. Este análisis, aunque más estadístico que predictivo, se beneficia directamente de la imputación realizada con técnicas de Ciencia de Datos, demostrando la conexión entre modelado predictivo y análisis económico. Asimismo, se destaca que no se ha utilizado previamente la regresión distributiva para evaluar el efecto del aumento al salario mínimo en México, un tema de debate relevante en la política pública y la economía del país.

Es importante destacar que, a lo largo del desarrollo de la tesis, se reconoció que los datos deben interpretarse en su contexto económico y social. Aplicar herramientas de Ciencia de Datos no es un fin en sí mismo; su valor radica en permitir un estudio riguroso de fenómenos que afectan a los hogares mexicanos. Esta perspectiva asegura que la metodología y los resultados estén orientados a generar conocimiento relevante y aplicable en la realidad.

2 Revisión de literatura

Contenidos

2.1 Medidas de desigualdad

2.1.1 Índice de Gini e Índice de Theil

La medida más común para medir la desigualdad ha sido el Índice de Gini. Este coeficiente sigue siendo utilizado de forma recurrente para medir la concentración de la riqueza, por ejemplo, estudios recientes para México describen un Índice de Gini de 0.454 para el 2020 [1]. Como anteriormente se menciona, este coeficiente engloba en una sola cifra la situación de la desigualdad del país, representando esto tanto ventajas como desventajas. Los resultados se encuentran entre 0, que representa la perfecta igualdad, y 1 que representa la perfecta desigualdad.

Existen diversos métodos para determinar este coeficiente, siendo la Curva de Lorenz la más utilizada. El coeficiente de Gini representará el área entre la línea de igualdad perfecta (la cual representa el escenario en el que toda la población obtuviera el mismo ingreso) y la Curva de Lorenz, esta última, representa la distribución acumulada del ingreso [2]. Entre más alejada esté la Curva de Lorenz de la línea de igualdad perfecta, mayor será la desigualdad para una población. Si bien es una herramienta útil para obtener un panorama general de la desigualdad, no nos permite analizar en qué subgrupos de la población se focaliza la misma o cuál es la estructura general del ingreso en la población.

El Índice de Theil es un poco más sensible a las colas que el Índice de Gini, pues gracias a su fórmula da más peso a las diferencias entre los

[1] Merari Cortés Sánchez, Adriana Zambrano-Reyes, and Tomás Gómez-Rodríguez. Evolución de la desigualdad salarial en méxico 2016-2020, un problema para el desarrollo económico. *Boletín Científico de las Ciencias Económico Administrativas del ICEA*, 12(23):6–13, 2023. ISSN 2007-4913. DOI: 10.29057/icea.v12i23.11572. URL https://doi.org/10.29057/icea.v12i23.11572. Recibido el 4 de septiembre de 2023; aceptado el 23 de octubre de 2023; publicado el 5 de diciembre de 2023

[2] Merari Cortés Sánchez, Adriana Zambrano-Reyes, and Tomás Gómez-Rodríguez. Evolución de la desigualdad salarial en méxico 2016-2020, un problema para el desarrollo económico. *Boletín Científico de las Ciencias Económico Administrativas del ICEA*, 12(23):6–13, 2023. ISSN 2007-4913. DOI: 10.29057/icea.v12i23.11572. URL https://doi.org/10.29057/icea.v12i23.11572. Recibido el 4 de septiembre de 2023; aceptado el 23 de octubre de 2023; publicado el 5 de diciembre de 2023

ingresos. No tiene un valor límite, pero entre más cercano se encuentre a 0, más se acerca a la igualdad perfecta. La popularidad de este índice se debe a que puede descomponerse en subgrupos. Es decir, maneja dominios "within.o "dentroz "between.o .entre"; esto significa que es capaz de medir la desigualdad dentro de un mismo grupo de la población, así como, la variación del ingreso entre grupos diferentes de la población [3]. Si bien se muestra un grado mayor de profundidad en el análisis, nuevamente obtenemos una cifra que refleja de forma general la desigualdad.

Según un informe reciente del Banco Mundial, el coeficiente de Gini para México en 2024 se ubica en 0.435, indicando una alta desigualdad en la distribución de ingresos dentro del país [4]. Este valor muestra una ligera mejora en comparación con años anteriores, como el 2018 (0.467). Sin embargo, México aún se clasifica entre las economías con mayores niveles de desigualdad, superando el umbral de 40 que define a los países con alta inequidad [5].

[3] Stefano Marchetti and Nikos Tzavidis. Robust estimation of the theil index and the gini coeffient for small areas. *Journal of Official Statistics*, 37(4):955–979, 2021. DOI: 10.2478/jos-2021-0041. URL https://doi.org/10.2478/jos-2021-0041

[4] Tadeo Campoy. México se ubica entre las economías con alta desigualdad, según informe del banco mundial. https://www.elimparcial.com/dinero/2024/10/15/mexico-se-ubica-entre-las-economias-con-alta-desigualdad-segun-informe-del-banco-mundial/, 2024. Publicado el 15 de octubre de 2024, accedido en junio de 2025

[5] Index Mundi. Mexico - indice de gini. https://www.indexmundi.com/es/datos/mexico/indicador/SI.POV.GINI, 2025. Accedido en junio de 2025

2.1.2 *Método DiNardo, Fortin y Lemieux*

Publicado en 1995, el método DiNardo, Fortin y Lemieux introduce una descomposición semiparamétrica, donde se pueden analizar las variaciones en el ingreso y ver su relación con diversos factores como las características de la población. Debido a que en su momento se buscaba analizar la importancia de los sindicatos y salarios mínimos frente a la distribución del ingreso, el modelo propuesto logra analizar cómo sería la distribución del ingreso de un año si se mantuvieran las características (en este caso niveles de sindicalización) de años anteriores[6].

Esta metodología utiliza toda la distribución salarial y, a través de la construcción de una contrafactual, logra analizar distribuciones salariales completas considerando sus características, pero con precios distintos. Por ejemplo, Rodríguez, Castro y Mendoza [7]construyen una contrafactual para la distribución salarial de los trabajadores informales, considerando las características de este grupo, pero remunerándolos de acuerdo al ingreso de los trabajadores formales. Esto hace al método muy útil para analizar el tipo de comparaciones que se propone el presente estudio.

[6] John DiNardo, Nicole M. Fortin, and Thomas Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Technical Report w5093, National Bureau of Economic Research, 1995. URL https://doi.org/10.3386/w5093

[7] Ricardo E. Rodríguez Pérez, Deyanira Castro Lugo, and Martín Mendoza López. Desigualdad salarial y trabajo informal en regiones de méxico. *Región y Sociedad*, 31:1–23, 2019. DOI: 10.22198/rys2019/31/1062. URL https://doi.org/10.22198/rys2019/31/1062

2.1.3 *Regresión Distributiva*

La regresión distributiva es una técnica utilizada para analizar y modelar la distribución de una variable dependiente (generalmente el ingreso, la riqueza o alguna medida de bienestar) en función de varias variables explicativas. Este tipo de regresión es especialmente útil en estudios de desigualdad económica, donde se busca entender cómo

diferentes factores (como el nivel educativo, la ocupación o la ubicación geográfica) contribuyen a las diferencias en los ingresos entre distintos grupos de la población. A diferencia de las regresiones convencionales que buscan una relación directa entre variables, la regresión distributiva pone énfasis en cómo los cambios en las variables independientes afectan a los diferentes percentiles de la distribución de la variable dependiente.

Este enfoque ofrece ventajas importantes frente a los índices de desigualdad tradicionales, como el Gini o Theil, que resumen la distribución en un único valor. La regresión distributiva no solo considera múltiples variables explicativas, sino que permite analizar cómo cada factor impacta distintos segmentos de la población, ofreciendo una comprensión más completa y matizada de la desigualdad. De esta manera, se pueden identificar patrones que pasarían desapercibidos si solo se emplearan coeficientes agregados, y al mismo tiempo es posible visualizar la distribución de manera gráfica, observando su evolución y dinámica a lo largo de toda la curva de ingresos.

De forma general, este modelo, que se basa en lo estipulado por Redmond, Doorley y McGuinness en 2020, realiza una serie de modelos probit en varios puntos de la distribución salarial de un año 0, antes de los pronunciados aumentos en el salario mínimo. La variable dependiente es binaria y toma el valor de 1 si el salario de un individuo está por debajo de un umbral específico y 0 en caso contrario. Este umbral se establece secuencialmente en diferentes puntos de la distribución salarial. Al promediar las probabilidades predichas en cada intervalo de salario, se obtiene una estimación de las distribuciones marginales de salario. Los coeficientes obtenidos en el año 0 se aplican a los datos del año 1, para construir una distribución salarial contrafactual, que muestra cómo habría sido la distribución salarial en el año 1, si los trabajadores hubieran recibido los salarios de acuerdo con la estructura salarial del año 0. Finalmente, se realiza una descomposición similar a las técnicas estándar de Oaxaca (1973) y Blinder (1973) sobre las distribuciones salariales observadas, analizando dos efectos: un efecto precio y un efecto de composición.[8]

Este enfoque es útil para separar los efectos del cambio en la política de salario mínimo sobre los salarios, permitiendo identificar qué parte de la diferencia en las distribuciones salariales es atribuible al cambio en el salario mínimo (efecto precio) y qué parte se debe a otros factores relacionados con la composición de la población laboral (efecto de composición).

En lugar de enfocarse únicamente en el ingreso promedio, la regresión distributiva permite ver cómo ciertos factores afectan de manera desigual al ingreso. La regresión distributiva es un enfoque muy

[8] Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland

valioso en estudios de economía y políticas públicas, ya que proporciona una visión más amplia y detallada de cómo las características socioeconómicas afectan a los diferentes segmentos de la población. La regresión distributiva proporciona un análisis más rico y pertinente que los índices tradicionales, al considerar múltiples variables explicativas y al permitir una evaluación granular de cómo distintos factores afectan a toda la distribución del ingreso. Incluso nos permite una visualización gráfica completa de la curva de distribución acumulada, abriendo con ello la puerta a conclusiones más pertinentes.

Por todas estas razones, la regresión distributiva se presenta como la herramienta más adecuada para este estudio. A diferencia de los índices tradicionales, permite considerar múltiples variables explicativas, evaluar efectos en distintos percentiles de la distribución y construir distribuciones contrafactuales que facilitan el análisis de cambios estructurales y de políticas. Su capacidad de generar visualizaciones detalladas y de revelar patrones que quedarían ocultos con indicadores agregados la convierte en un enfoque más riguroso, completo y pertinente para comprender la desigualdad salarial en México y evaluar el impacto del incremento del salario mínimo de manera precisa y contextualizada.

2.2 *Incrementos salariales y su impacto en la desigualdad*

Las implicaciones del aumento al salario mínimo han sido objeto de un amplio análisis en la literatura económica, con resultados que varían significativamente según el país, año, contexto político, y estructura social. La polémica que existe alrededor de este tema y, sobre todo, el hecho de que exista numerable evidencia tanto a favor como en contra sugiere que de forma genérica no es posible señalar el aumento al salario mínimo como una medida intrínsecamente positiva o negativa.

Por citar algunos ejemplos, en Irlanda, el aumento del salario mínimo en 2015 mostró resultados positivos amplios. Tras el incremento, hubo una disminución de cuatro puntos porcentuales en el número de trabajadores que ganaban el salario mínimo o menos. Este aumento no solo benefició a quienes ganaban justo el salario mínimo, sino que también mejoró los salarios de los trabajadores que ganaban por encima del mínimo, lo que se conoce como el "efecto faro". En este caso, el impacto positivo en los salarios se extendió hasta el 30° percentil de la distribución salarial, y la desigualdad entre los salarios más altos y más bajos disminuyó significativamente [9]

En contraste, el análisis de los efectos del aumento del salario mínimo en Ecuador durante el periodo de 2007 a 2019 mostró que el incremento llevó a un ligero aumento en el empleo informal, especialmente en los deciles más bajos de ingreso. Los resultados respaldan el "efecto faro"

[9] Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland

para los trabajadores informales, quienes pudieron haber negociado mejores salarios a raíz del incremento. Sin embargo, los autoempleados experimentaron un efecto negativo en sus ingresos. Así, aunque el aumento del salario mínimo en Ecuador llevó a una menor reducción de la pobreza e inequidad, el impacto general del aumento fue limitado debido a los altos niveles de informalidad laboral [10]. Un análisis más amplio que explora el efecto del incremento en el salario mínimo en las seis economías más grandes de América Latina durante el periodo 2001-2018 encontró que los efectos positivos del incremento en el salario mínimo disminuyeron en la década de 2010 para todos los países, coincidiendo con un mercado laboral más debilitado a nivel regional [11].

El impacto del aumento del salario mínimo en México ha sido analizado en diversos estudios, los cuales destacan tanto efectos positivos como desafíos persistentes. Kaplan y Arceo [12] encontraron que los cambios en el salario mínimo real en México inducen ajustes en los ingresos laborales de manera homogénea, es decir, el aumento afecta a todas las categorías salariales, pero de forma más débil cuando se considera el ingreso total de los trabajadores. Sin embargo, este efecto ha ido perdiendo fuerza con el tiempo, especialmente durante los años 1994-2001. Esto sugiere que el impacto del salario mínimo en los ingresos laborales podría estar disminuyendo en años recientes, lo que pone en evidencia la necesidad de considerar otros factores para explicar las disparidades salariales.

En estudios más recientes, el realizado por Campos y Esquivel en 2023 resalta los efectos positivos del aumento salarial en la reducción de la pobreza, especialmente en la frontera norte de México, donde el incremento fue más significativo. En esta región, la pobreza disminuyó en un 2.6 %, mientras que el resto del país experimentó una caída más modesta debido a un aumento del 16 % en el salario mínimo. Sin embargo, la elasticidad de la pobreza respecto al salario mínimo muestra una relación inversa, sugiriendo que si bien el salario mínimo contribuye a la disminución de la pobreza, la persistencia de las desigualdades salariales sigue siendo un desafío.[13]

En términos de desigualdad de género, Campos [14] señala que, a pesar de los avances en la mejora salarial, las brechas de ingresos entre hombres y mujeres persisten, con las mujeres ganando entre un 13 % y un 15 % menos que los hombres. Esta disparidad resalta la necesidad de abordar las desigualdades estructurales que afectan a ciertos grupos, incluso cuando el salario mínimo se ha ajustado para mejorar las condiciones generales de los trabajadores[15].

En conjunto, estos estudios sugieren que, aunque el aumento del salario mínimo ha tenido un efecto positivo en algunos aspectos, las desigualdades estructurales, como la brecha salarial de género, siguen

[10] Susana Herrero Olarte. The minimum wage in ecuador. *Revista de Economía y Trabajo*, 2023. Based on ENEMDUM and INEC data, Panel Data Regression methodology

[11] Carlo Lombardo, Lucía Ramírez-Leira, and Leonardo Gasparini. Does the minimum wage affect wage inequality? a study for the six largest latin american economies. *Latin American Economic Review*, 2024. URL https://laer-journal.springeropen.com/articles/10.1186/s40503-024-00103-7. EPH, PNAD, CASEN, GEIH, ENIGH, ENAHO; comparative cross-country analysis

[12] David S. Kaplan and Francisco Pérez Arce Novaro. El efecto de los salarios mínimos en los ingresos laborales de méxico. *El Trimestre Económico*, 73(289): 139–173, 2006. URL https://doi.org/10.20430/ete.v73i289.556. Accedido en julio de 2025

[13] Raymundo M. Campos-Vazquez and Gerardo Esquivel. The effect of the minimum wage on poverty: Evidence from a quasi-experiment in mexico. *The Journal of Development Studies*, 59(3):360–380, 2023. DOI: 10.1080/00220388.2022.2130056

[14] Raymundo Campos. Movilidad social, empleo e ingresos laborales en méxico. Technical report, Centro de Estudios Económicos, El Colegio de México, 2021. URL https://movilidadesocial.colmex.mx/wp-content/uploads/2021/10/5.-Raymundo-Campos.pdf. ENOE & IMSS data used

[15] Raymundo M. Campos-Vazquez and Gerardo Esquivel. The effect of the minimum wage on poverty: Evidence from a quasi-experiment in mexico. *The Journal of Development Studies*, 59(3):360–380, 2023. DOI: 10.1080/00220388.2022.2130056

siendo desafíos significativos.

2.3 *Ingresos no reportados y su imputación*

La imputación de datos de ingresos no reportados es crucial en los estudios económicos y laborales, especialmente en el contexto mexicano, donde un porcentaje significativo de trabajadores no reporta sus ingresos de manera precisa. El número de personas que no reportan sus ingresos ha ido en incremento y para el año 2024 es cerca del 30 %. Según Campos[16], la mayoría de los estudios sobre economía laboral y pobreza en México no consideran a los individuos que reportan trabajar, pero no informan sus ingresos. En algunos casos, estos registros no reportados se usan con $0 y en otros simplemente se eliminan de las bases de datos. Este vacío de información puede llevar a mediciones incorrectas de la pobreza y la desigualdad, afectando la precisión de las políticas públicas orientadas a la reducción de estas problemáticas.

Para abordar este desafío, la literatura recomienda el uso de diversos métodos de imputación para corregir los ingresos faltantes, tales como el pareamiento por puntajes de propensidad, el método de asignación de la mediana y el método de imputación en la mediana de un grupo[17]. Estos enfoques permiten estimar los ingresos de los trabajadores de manera más precisa, mejorando las mediciones de pobreza y desigualdad.

Sin embargo, si bien estos métodos clásicos tienen la ventaja de ser transparentes, en muchos casos se basan en supuestos lineales o en agrupamientos categóricos que pueden limitar su capacidad para capturar la complejidad real de los datos. Por ejemplo, variables como la edad, la escolaridad o la condición de ocupación no necesariamente se relacionan de forma lineal con el ingreso, y pueden interactuar entre sí de maneras que los enfoques tradicionales no consideran.

Este trabajo propone, por tanto, una alternativa complementaria a través del uso de técnicas modernas de aprendizaje automático (machine learning), que permiten modelar relaciones complejas y no lineales entre múltiples variables. El objetivo no es reemplazar los enfoques anteriores, sino aprovechar la capacidad predictiva de estos modelos para obtener estimaciones más precisas y consistentes del ingreso laboral no reportado.

Además, las técnicas de imputación no solo son importantes para corregir la falta de datos, sino también para generar comparaciones válidas entre diferentes períodos o subgrupos de la población. Por ejemplo, al realizar un análisis de la pobreza laboral en México, la imputación de ingresos no reportados puede contribuir a una visión más completa de la situación económica de los trabajadores y la evolución de la pobreza a lo largo de los años [18].

[16] Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320): 803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003

[17] Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320): 803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003

[18] Raymundo Campos. Movilidad social, empleo e ingresos laborales en méxico. Technical report, Centro de Estudios Económicos, El Colegio de México, 2021. URL https://movilidadesocial.colmex.mx/wp-content/uploads/2021/10/5.-Raymundo-Campos.pdf. ENOE & IMSS data used

3 Desarrollo de modelo para la predicción del ingreso no reportado

Contenidos

3.1 Reconocimiento de bases de datos

Para el desarrollo de esta investigación se hace uso de la Encuesta Nacional de Ocupación y Empleo (ENOE) dirigida por el INEGI, cuyo objetivo es recabar información acerca de las características de la población en términos de trabajo, ocupación e ingresos. Esta base de datos cuenta con información a nivel nacional de çuatro tamaños de localidad, de cada una de las 32 entidades federativas y para un total de 39 ciudades". Los datos presentados por dicha encuesta tienen una periodicidad trimestral; la muestra encuestada está dividida en cinco paneles y es rotativa çada uno [panel] permanece en la muestra

durante cinco trimestres, por lo que, pasado dicho tiempo, se sustituye por otro de características similares. La quinta parte de la muestra que ya cumplió con su ciclo de cinco visitas se reemplaza cada tres meses. Este esquema garantiza la confiabilidad de la información obtenida, ya que en cada trimestre se mantiene el 80 % de la muestra "[1].

A fin de medir el impacto del incremento salarial, en este trabajo, se analizará el último trimestre de 2018, considerándolo como el antes de la implementación de la política pública (el incremento salarial) y el segundo trimestre de 2024 como el último disponible al momento de comenzar con esta investigación, considerándolo como el después de la política pública.

Si bien la encuesta conserva su estructura general, es común que con el transcurso de los años se realicen modificaciones a la misma; el periodo trabajado no es la excepción. El número total de columnas incluidas en la base de datos del primer trimestre del año 2018 fue de 104, información que se mantuvo igual para los años 2019 y 2020. En el año 2021 la base de datos sufrió las siguientes modificaciones:

- Las variables **'EST_D' - Estrato de diseño**, **'T_LOC' - Tamaño de localidad** y **'FAC' - Factor de expansión** empezaron a reportarse de forma mensual y trimestral. Estas variables no corresponden a una pregunta de la encuesta, sino que sirven de apoyo para identificar las viviendas encuestadas. A raíz de esto, se añadieron 3 columnas a la base de datos quedando de la siguiente forma:
  - **'EST_D_TRI' - Estrato de diseño trimestral** y **'EST_D_MEN' - Estrato de diseño mensual** reemplazaron a la variable **'EST_D' - Estrato de diseño**
  - **'T_LOC_TRI' - Tamaño de localidad trimestral** y **'T_LOC_MEN' - Tamaño de localidad mensual** reemplazaron a la variable **'T_LOC' - Tamaño de localidad**
  - **'FAC_TRI' - Factor de expansión trimestral y 'FAC_MEN' - Factor de expansión mensual** reemplazaron a la variable **'FAC' - Factor de expansión**
- Se añadieron preguntas a la encuesta relacionadas a la migración y el lugar de trabajo. El total de preguntas añadidas fue de 5 y consta de las siguientes:
  - **'cs_p20a_1' - Hace un año, en (mes) de (año), ¿en qué estado de la República o en qué país vivía ...?**
  - **'cs_p20a_c' - Clave de entidad o país**
  - **'cs_p20b_1' - ¿En qué municipio (alcaldía) vivía ... hace un año?**
  - **'cs_p20b_c' - Clave de municipio**

[1] Instituto Nacional de Estadística y Geografía (INEGI). Encuesta nacional de ocupación y empleo (enoe): población de 15 años y más. https://www.inegi.org.mx/programas/enoe/15ymas/, 2024. Consultado el 16 de octubre de 2024

  - **'cs_p20c_1' - ¿Por qué ... dejó de vivir en ese (municipio o alcaldía o páis)?**

- Se añadió la variable **'TIPO' - Tipo de entrevista**, con valores 1=Entrevista cara a cara y 2=Entrevista telefónica. Esta variable se añade pues a raíz de la pandemia se empezaron a realizar entrevistas por teléfono.

- Se añadió la variable **'MES_CAL' - Mes calendario**, con valores 1=Enero, 2=Febrero, 3=Marzo y 96=Identificador de semana 1 del 1er trimestre. Esta variable está clasificada como de apoyo para procesamiento y no corresponde a una pregunta de la encuesta.

- Se añadió la variable **'CA' - Registro por Panel**, con valores 1=Registro de panel común y 2=Resto de paneles. Esta variable está clasificada como de apoyo para procesamiento y no corresponde a una pregunta de la encuesta.

La mayoría de estos cambios se han mantenido, con la excepción de que la variable **'CA' - Registro por Panel** fue eliminada de las bases de datos para años posteriores.

Ninguna de las modificaciones mencionadas tiene impacto que pueda ocasionar sesgos en la comparación de los resultados obtenidos entre los trimestres elegidos del 2018 y 2024.

3.2 *Tratamiento de datos en variables predictoras y variable a predecir*

Antes de construir los modelos de imputación, es necesario realizar un análisis detallado de las variables disponibles en la base de datos. El objetivo es identificar cuáles de ellas pueden ser utilizadas como variables predictoras del ingreso. Este proceso implica explorar el comportamiento de cada variable y evaluar su capacidad explicativa, de manera que se incluyan en los modelos de imputación únicamente aquellas que contribuyen significativamente a la predicción, pero también usando tantas como sea adecuado. De este modo, se busca aprovechar al máximo la riqueza informativa de la encuesta, sin introducir ruido innecesario que pueda afectar la calidad de las estimaciones.

3.2.1 *Variables Predictoras*

Para la manipulación de datos se utiliza Python 3.12. Posterior a la importación de librerías, la base de datos para el segundo trimestre de 2024 y el último trimestre de 2018 de la encuesta se carga en forma de

un DataFrame, que presenta 104 variables y un aproximado de 400,000 entradas por trimestre.

Esta encuesta considera la información de todos los miembros del hogar encuestado, es decir, genera un registro por miembro del hogar (incluyendo menores de edad), cuente con ingreso o no. Por este motivo, hay renglones en la encuesta que no son parte de este análisis y deben ser omitidos, pues no representan personas económicamente activas. Por ello, la primera modificación realizada a las bases de datos es la eliminación de los registros correspondientes a personas que no perciben ingresos. Esta información viene representada en las columnas siguientes:

- **'clase1' - Condición de actividad primera categoría (CLASE1)**: Clasificación de la población en población económicamente activa y población no económicamente activa. Cuenta con las siguientes opciones:

  - 0 - No Aplica
  - 1 - Población Económicamente Activa (PEA)
  - 2 - Población No Económicamente Activa (PNEA)

  Tomando en cuenta lo contenido en esta variable, se descartan todos los registros donde esta categoría No Aplica, es decir, todos los registros que cuentan con un valor "0", también fueron descartados los registros para la Población No Económicamente Activa.

- **'eda' - Edad**: Número de años transcurridos entre la fecha de nacimiento de las personas y la fecha de la entrevista. Únicamente se consideran los registros de personas mayores de 15 años.

- **'r_def' - Resultado definitivo (R_DEF)**: se refiere al resultado definitivo de la entrevista, diferenciando entre las categorías 0 para .Entrevista lograda7 15 para .Entrevista suspendida". Únicamente serán consideradas las entrevistas logradas.

- **'hrsocup' - Horas trabajadas a la semana (HRSOCUP)**: Número de horas trabajadas a la semana de la población ocupada. Para esta categoría, únicamente se consideran los registros diferentes de 0.

- **'ing7c' - Nivel de ingresos para ocupados (ING7C)**: Clasificación de la población ocupada por nivel de ingreso. Esta columna cuenta con 8 posibles entradas:

  - 0 - No aplica
  - 1 - Hasta un salario mínimo
  - 2 - Más de 1 hasta 2 salarios mínimos

- 3 - Más de 2 hasta 3 salarios mínimos
- 4 - Más de 3 hasta 5 salarios mínimos
- 5 - Más de 5 salarios mínimos
- 6 - No recibe ingresos
- 7 - No especificado

Para esta variable se descartaron todos los registros para la opción 6 - No recibe ingresos.

Los registros resultantes conforman la población económicamente activa, que trabaja al menos una hora a la semana y que recibe un pago por su trabajo.

Adicionalmente, la variable '**ing_x_hrs**' - Promedio de ingreso por hora trabajada, es eliminada de la base de datos. Esta variable tiene una correlación importante con la variable a predecir, sin embargo, cuando no se cuenta con un registro para la columna de Ingreso, tampoco se cuenta con información para la columna de Ingreso por hora. Por este motivo, si se toma en consideración para entrenar el modelo y generar una predicción, resultará en un impacto negativo en la predicción.

	clase1	eda	r_def	hrsocup	ing7c	ing_x_hrs
0	1	39	0	30	4	66.66667
1	2	38	0	0	0	0.00000
2	2	15	0	0	0	0.00000
3	2	12	0	0	0	0.00000
4	0	10	0	0	0	0.00000
5	0	8	0	0	0	0.00000
6	1	28	0	60	5	83.33333
7	1	27	0	30	2	33.33333
8	0	1	0	0	0	0.00000
9	1	52	0	48	3	31.25000

Figura 3.1: Variables a considerar para determinar a la población que percibe ingresos en 2018

Para el archivo de 2024 se realizan las modificaciones siguientes atendiendo a los cambios en la presentación de las bases de datos por parte INEGI:

- Se calcula la correlación entre las variables presentadas de forma mensual y trimestral. Se elimina una de las columnas, conservando los datos trimestrales, al coincidir con la descripción para los años anteriores.
- Se eliminan las columnas auxiliares que fueron agregadas posterior al 2020 **'mes_cal'** y **'tipo'**.

	clase1	eda	r_def	hrsocup	ing7c	ing_x_hrs
0	1	33	0	30	3	155.03876
1	1	52	0	0	7	0.00000
2	2	74	0	0	0	0.00000
3	1	24	0	66	1	18.18182
4	1	24	0	30	2	77.51938
5	1	50	0	25	1	0.00000
6	2	36	0	0	0	0.00000
7	2	39	0	0	0	0.00000
8	1	52	0	0	1	0.00000
9	1	28	0	0	0	0.00000

Figura 3.2: Variables a considerar para determinar a la población que percibe ingresos en 2024

Variable 1	Variable 2	Correlación
t_loc_men	t_loc_tri	1.0
est_d_men	est_d_tri	0.99
fac_men	fac_tri	0.86

Cuadro 3.1: Correlación entre variables mensuales y trimestrales

- Se eliminan las preguntas añadidas posterior al 2020 al no contar con datos en los años previos para incluir en el estudio **'cs_p20a_1', 'cs_p20a_c', 'cs_p20b_1', 'cs_p20b_c', 'cs_p20c_1'**.

Una vez que se cuenta con bases de datos equiparables para los dos años se inicia con la limpieza de datos al convertir los espacios en blanco o valores no numéricos a valores nulos. Previamente se realizó una inspección del archivo CSV para confirmar que todo valor distinto se trata de valores nulos. Una vez realizada está modificación, se realiza un concentrado de los porcentajes de valores nulos por columna, encontrando que un total de 12 variables poseen valores nulos en porcentajes que varían del 0.46 % al 100 %.

Para iniciar, se eliminan las columnas conteniendo más del 70 % de valores nulos, pues no aportarán información relevante al estar casi completamente vacías. Las columnas eliminadas en esta primera operación fueron las siguientes:

- **'loc' - Localidad** Es todo lugar ocupado por una o más viviendas que pueden estar habitadas o deshabitadas. Este lugar es reconocido por un nombre dado por la ley o la costumbre. Las localidades son de dos tipos: urbanas y rurales.
  - 100 % de valores nulos
- **'cs_p14_c' - Código de carrera** Código correspondiente al nombre de la carrera que estudiaron las personas.
  - 71.50 % de valores nulos

- **'cs_p15' - Antecedente escolar** Corresponde a la pregunta Antecedente escolar.... carrera del nivel normal, técnico o profesional, es decir, el nivel escolar que la persona tuvo que aprobar para poder cursar una carrera técnica o profesional.
  - 73.46 % de valores nulos
- **'cs_p16' - Egreso** Si terminaron o no sus estudios todas las personas que declararon estudiar o haber estudiado alguna carrera técnica, normal, profesional, maestría o doctorado.
  - 71.30 % de valores nulos
- **'cs_ad_mot' - Ausentes definitivos motivo** Motivo principal por el que se fue alguna persona del hogar.
  - 100 % de valores nulos
- **'cs_p20_des' - Ausentes definitivos descripción del motivo** Descripción de otro motivo por el que se fue la persona.
  - 100 % de valores nulos
- **'cs_ad_des' - Ausentes definitivos destino** ¿A qué estado de la república o país se fue . . . ?
  - 100 % de valores nulos
- **'cs_nr_mot' - Nuevos residentes motivo** Motivo principal por el que llegó alguna persona a formar parte del hogar.
  - 98.76 % de valores nulos
- **'cs_p22_des' - Nuevos residentes descripción del motivo** Nuevo residente, descripción de otro motivo por el cual llegó a formar parte del hogar.
  - 100 % de valores nulos
- **'cs_nr_ori' - Nuevos residentes origen** De qué lugar vinieron los nuevos residentes.
  - 99.02 % de valores nulos

Para las 2 variables restantes, se realizaron imputaciones a través de la creación de una nueva categoría para aquellas preguntas cuya respuesta era realmente un valor no especificado. En el caso de la variable **'n_hij'** el valor 99 representa 'Número de hijos no especificados', de acuerdo con el diccionario de variables de la ENOE. [2]

[2] Instituto Nacional de Estadística y Geografía (INEGI). Encuesta nacional de ocupación y empleo 2018, cuestionario ampliado, datos correspondientes al primer trimestre. diccionario de datos: Sdemt118. https://www.inegi.org.mx/rnm/index.php/catalog/448/data-dictionary/F10?file_name=SDEMT118, August 2022. Red Nacional de Metadatos, consultado el 17 de agosto de 2025

Variable imputada	Valor de imputación	Porcentaje de valores nulos
'n_hij'	99 Ñ Hijos no especificados	60.37 %
'mun'	0Ñ No especificado	0.46 %

Cuadro 3.2: Imputación de valores para variables predictoras

3.2.2 *Variable a predecir*

Finalmente, se dirigió el enfoque a la variable de interés 'ingocup', que representa el ingreso mensual declarado por cada encuestado. Como se describía en apartados anteriores, el valor de este rubro es ingresado como 0 cuando el encuestado decide no compartir el monto de su ingreso mensual. Por ello es importante detectar en qué casos este valor realmente es 0 y en qué casos deberá identificarse como un valor nulo. Para el caso del presente trabajo, la información ya se encontraba filtrada y únicamente se contaba con registros de personas pertenecientes a la PEA, con horas trabajadas, que perciben ingresos por su ocupación.

Es por ello, que la designación de NaN o valores nulos se realizó tomando en cuenta lo siguiente:

- Ingreso mensual de los ocupados igual a 0

Una vez terminado este proceso, se identificó que aproximadamente el 30 % de esta variable corresponde a valores nulos (NaN).

	ingocup
0	8600.0
6	21500.0
7	4300.0
9	6450.0
11	1419.0
12	6000.0
13	15000.0
15	12000.0
25	5590.0
26	6020.0
27	NaN
29	NaN
30	4200.0
31	NaN
35	NaN
36	6000.0
40	5590.0
41	5160.0
48	NaN
51	5600.0

Figura 3.3: Ejemplo de primeros 20 registros resultantes de la columna Ingocup para 2018

3.3 *Modelado Predictivo*

	ingocup
0	20000.0
3	5160.0
4	10000.0
5	NaN
11	NaN
13	NaN
16	NaN
17	NaN
19	NaN
25	NaN
27	24000.0
28	NaN
29	9000.0
30	3870.0
31	12000.0
33	8600.0
35	NaN
42	12000.0
43	NaN
44	9460.0

Figura 3.4: Ejemplo de primeros 20 registros resultantes de la columna Ingocup para 2024

3.3.1 *División del conjunto de datos*

Para iniciar se separó el conjunto de datos en dos grupos, aquellos que contenían valores nulos en la variable 'ingocup' , los cuales son los valores a predecir y el resto de los registros. A su vez, el conjunto de datos que no posee valores nulos, fue separado en valores de entrenamiento y valores de prueba, en preparación para su uso en el modelado predictivo. La separación se realizó con un 30 % de valores de prueba y el 70 % restante como valores de entrenamiento, respondiendo a la heurística común en la Ciencia de Datos que busca balancear adecuadamente la cantidad de datos para el aprendizaje del modelo y su evaluación en datos no vistos. También se especificó una semilla con el parámetro *random_state=42* para asegurar la repetibilidad de los resultados.

3.3.2 *Escalamiento, Valores atípicos (Outliers) y Reducción de Caractarísticas (Feature Reduction)*

El primer objetivo particular de este trabajo es contar con una base de datos que servirá de insumo, para el análisis posterior. El objetivo final de este trabajo, es analizar los efectos del incremento del salario mínimo en la distribución del ingreso por trabajo (salario), prestando un particular interés en la estructura de la población que percibe ingresos. Es por este motivo, que conservar la integridad de la información es vital.

Técnicas de preprocesamiento de datos como el escalamiento, eliminación o modificación de valores atípicos (outliers) y técnicas avanzadas de reducción de variables son ampliamente recomendadas cuando se trabaja con modelado predictivo, pues el modificar nuestras variables para que se ajusten mejor a los algoritmos predictivos coadyuva a obtener un desempeño superior. No obstante, considerando que es vital mantener la integridad de los valores para las características de la población, se ha decidido asumir las implicaciones de evitar el uso de estas técnicas. En su lugar, se buscará que los modelos propuestos sean evaluados con distintas métricas para analizar si las predicciones obtenidas son fiables.

3.3.3 *Selección de Características (Feature Selection)*

Como se mencionó en puntos anteriores, la base de datos cuenta con un alto número de columnas o variables predictivas. Este elevado número de variables podría dificultar que el modelo genere buenas predicciones, por ello se consideró importante identificar aquellas que mejor ayudan a definir la variable a predecir. Este procedimiento de selección de características no solo busca optimizar el desempeño del modelo, sino que también cumple una función crucial al evitar la inclusión de ruido innecesario (es decir, variables irrelevantes que pueden distorsionar las predicciones) o la exclusión de predictoras útiles que aportan información valiosa al análisis.

Como métodos de Selección de Características (Feature Selection) se compararán los resultados obtenidos a través de 2 metodologías distintas:

- **Random Forest:** Es un modelo de aprendizaje automático basado en estructuras más simples conocidas como árboles de decisión, combinado con la técnica de Bagging o Bootstrap Aggregation. Los árboles de decisión son estructuras jerárquicas en las que la información se particiona dependiendo de si cumple o no con ciertas condiciones, generando así una serie de ramificaciones que separan la información en subconjuntos más homogéneos. La técnica de Bagging reduce la varianza de los resultados al entrenar múltiples árboles de decisión en distintas muestras Bootstrap (conjuntos de datos creados al seleccionar aleatoriamente observaciones del conjunto de entrenamiento original). Cada árbol genera una predicción y, en el contexto de la selección de características o Feature Selection, el modelo evalúa la importancia de cada variable. Las variables más relevantes son aquellas que con mayor frecuencia participan en las particiones iniciales y generan una mayor ganancia de información. [3]
- **Gradient Boosting:** También se trata de un método de aprendizaje

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009

automático. En primer lugar, el Boosting es una técnica en la que a través de numerosas iteraciones, un modelo se entrena para ir corrigiendo los errores cometidos en las iteraciones previas. Para ello, se va realizando un ajuste en cada secuencia a partir de los errores entre los valores reales y las predicciones de la iteración anterior. Para realizar este ajuste, se utiliza el Gradiente Descendiente o Gradient Descent (derivando así el nombre de Gradient Boosting) donde se minimiza una función de pérdida. Con esta técnica, las variables más relevantes son aquellas que más contribuyen a reducir la función de pérdida [4].

[4] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. DOI: 10.1214/aos/1013203451

Durante la aplicación inicial del modelo de Gradient Boosting sobre la base completa con 95 variables, se observó que el modelo seleccionaba únicamente dos predictores: **ing7c** (nivel de ingresos para ocupados) y **cs_p14_c** (código de carrera). Si bien estas variables presentaban alta capacidad predictiva y generaban buenos resultados en términos de métricas globales como el coeficiente de determinación R^2, se identificó un problema importante: el modelo tendía a sobreestimar ingresos en diversos casos, con errores absolutos grandes a nivel individual. Esto reveló una falta de robustez en el modelo, atribuida a que ambas variables, aunque potentes, no incluyen información demográfica esencial como edad, sexo o nivel educativo.

Esta situación era especialmente problemática considerando que una de las metas principales del estudio es la implementación posterior de un análisis de regresión distributiva, el cual requiere información que refleje la composición demográfica de la población. Por esta razón, se decidió realizar una corrección metodológica: se eliminó deliberadamente la variable **ing7c** antes de ejecutar nuevamente el modelo de Gradient Boosting para seleccionar características. Esto obligó al algoritmo a identificar otras variables relevantes para la predicción. Posteriormente, una vez identificadas estas nuevas variables, se reincorporó **ing7c** al conjunto de predictores finales, ya que su inclusión mejora el desempeño del modelo, siempre y cuando no monopolice la información utilizada.

Este enfoque balanceado permitió construir un modelo más robusto, con un mejor desempeño a nivel individual, al tiempo que se conservaba la diversidad de variables necesarias para los análisis posteriores sobre la estructura salarial y composición de la fuerza laboral.

Para la selección de variables en este estudio, es fundamental conservar las variables originales sin alteraciones, dado que el análisis posterior de Regresión Distributiva requiere mantener la interpretación de cada variable socioeconómica. Por esta razón, no se consideraron métodos como el Análisis de Componentes Principales (PCA Principal Components Analysis), ya que este método genera nuevas variables

(componentes) a partir de combinaciones lineales de las originales, enfocándose en la reducción de dimensionalidad más que en la selección de variables.

Se optó por utilizar Random Forest y Gradient Boosting como técnicas complementarias de selección de variables por varias razones: ambos métodos permiten trabajar con un alto número de variables, ofrecen información sobre la importancia relativa de cada variable y son robustos frente a outliers, lo que garantiza que la selección sea confiable incluso en presencia de observaciones atípicas, como sucede con el ingreso. Random Forest proporciona una visión general de la relevancia de las variables a través del promedio de múltiples árboles, mientras que Gradient Boosting, al optimizar los errores de manera secuencial mediante gradiente descendente, permite confirmar y refinar la selección obtenida, identificando patrones sutiles que podrían pasar desapercibidos.

Otros métodos, como Lasso, basados en relaciones lineales, podrían haberse considerado, pero en este contexto no es apropiado asumir que las relaciones entre las variables y el resultado son principalmente lineales, lo que limita su aplicabilidad para la selección de características en este estudio.

También se realizaron pruebas utilizando el total de las variables predictivas para cada modelo de predicción, con la intención de conocer cuál es el que genera un mejor desempeño.

Es relevante subrayar que, a diferencia de las técnicas de reducción de variables, al realizar selección de variables no se modifica la estructura de la base de datos, únicamente identificamos aquellas que explican en mayor medida la variable a predecir.

Las variables seleccionadas a través de los diferentes métodos son las siguientes (se incluye la definición de cada variable conforme al Diccionario de Datos del INEGI):

Random Forest:

- **'mun' - Municipio:** Es la división político administrativa más pequeña en que se divide el territorio nacional, que es reconocido por un nombre dado por la ley o la costumbre y que está conformada por una o varias localidades. En México cada uno de los 31 estados está dividido en municipios, mientras que el Distrito Federal está dividido en delegaciones. El número de municipios varía en cada estado, en total existen 2 456 municipios y 16 delegaciones.
- **'con' - Control:** Es un consecutivo de cinco dígitos que al unirse con el número de entidad, permite identificar las áreas de listado en el ámbito nacional.
- **'upm' - Unidad primaria de muestreo:** Unidades Primarias de Muestreo: Éstas son unidades de área con límites identificables en el

terreno, que agrupan un conjunto de viviendas, cuyos criterios para su conformación dependen de cada proyecto.

- **'n_pro_viv' - Número progresivo de la vivienda:** Número consecutivo que se asigna a cada una de las viviendas existentes en un control o área de listado. Corresponde al anotado en el croquis de manzana o localidad rural para facilitar su identificación física sobre el terreno.
- **'nac_dia' - Día de nacimiento:** Día que nació la persona entrevistada.
- **'nac_mes' - Mes de nacimiento:** Mes en que nació la persona entrevistada.
- **'cs_p14_c' - Código de carrera:** Código correspondiente al nombre de la carrera que estudiaron las personas.
- **'fac' - Factor de expansión:** Valor numérico de seis dígitos que indica a cuántas personas representa el entrevistado en población.
- **'ing7c' - Nivel de ingresos para ocupados:** Clasificación de la población ocupada por nivel de ingreso
- **'hrsocup' - Horas trabajadas a la semana:** Número de horas trabajadas a la semana de la población ocupada.

Gradient Boosting:

- **'est' - Estrato:** Clasificación de las personas y hogares de acuerdo con las características sociodemográficas de los habitantes de las viviendas y las características físicas y equipamiento de las mismas. Se clasifica en: Alto, medio alto, medio bajo y bajo.
- **'n_ren' - Número de renglón:** Es el número consecutivo que se asigna a cada uno de los integrantes del hogar. Sirve para identificar el registro de cada persona de manera individual dentro de la vivienda y del hogar.
- **'eda' - Edad:** Número de años transcurridos entre la fecha de nacimiento de las personas y la fecha de la entrevista.
- **'nac_anio' - Año de nacimiento:** Año en que nació la persona entrevistada.
- **'cs_p14_c' - Código de carrera:** Código correspondiente al nombre de la carrera que estudiaron las personas.
- **'c_ocu11c' - Condición de ocupación:** Clasificación de la población ocupada por condición de ocupación. Se clasifican como:
  - 1 - Profesionales, técnicos y trabajadores del arte

- 2 - Trabajadores de la educación
- 3 - Funcionarios y directivos
- 4 - Oficinistas
- 5 - Trabajadores industriales artesanos y ayudantes
- 6 - Comerciantes
- 7 - Operadores de transporte
- 8 - Trabajadores en servicios personales
- 9 - Trabajadores en protección y vigilancia
- 10 - Trabajadores agropecuarios
- 11 - No especificado

- **'ambito2' - Tamaño de la unidad económica segunda categoría:** Clasificación del tamaño de la unidad económica segunda categoría para la población ocupada.
  - 1 - Micronegocios
  - 2 - Sin establecimiento
  - 3 - Con establecimiento
  - 4 - Pequeños establecimientos
  - 5 - Medianos establecimientos
  - 6 - Grandes establecimientos
  - 7 - Gobierno
  - 8 - Otros

- **'anios_esc' - Años de escolaridad:** Clasificación de número de años de escolaridad según nivel de instrucción para la población de 12 años y más.

- **'hrsocup' - Horas trabajadas a la semana:** Número de horas trabajadas a la semana de la población ocupada.

- **'tcco' - Dato para el cálculo de la tasa TCCO:** Clasificación de los datos para el cálculo de la tasa de condiciones críticas de ocupación, que trabajan menos de 35 hrs., más de 35 hrs. y más de 48 hrs.
  - 1 - Oh35rm (ocupados que trabajan menos de 35 hrs por razones de mercado)
  - 2 - Oh35sm (ocupados que trabajan de 35 hrs ó más y ganan hasta 1 salario mínimo)
  - 3 - Oh48sm (ocupados que trabajan más de 48 horas con ingresos de más de 1 hasta 2 s.M.)

- **'emp_ppal':** Clasificación de empleos formales e informales de la primera actividad
- **'mh_col':** Columnas de la matriz Hussmanns del trabajo principal
  - 1 - Trabajadores subordinados y remunerados - Asalariados INFORMALES
  - 2 - Trabajadores subordinados y remunerados - Asalariados FORMALES
  - 3 - Trabajadores subordinados y remunerados - Con percepciones no salariales INFORMALES
  - 4 - Trabajadores subordinados y remunerados - Con percepciones no salariales FORMALES
  - 5 - Empleadores INFORMAL
  - 6 - Empleadores FORMAL
  - 7 - Trabajadores por cuenta propia INFORMAL
  - 8 - Trabajadores por cuenta propia FORMAL
  - 9 - Trabajadores no Remunerados INFORMAL
  - 10 - Trabajadores no Remunerados FORMAL
- **'ing7c' - Nivel de ingresos para ocupados:** Clasificación de la población ocupada por nivel de ingreso.
  - 0 - No aplica
  - 1 - Hasta un salario mínimo
  - 2 - Más de 1 hasta 2 salarios mínimos
  - 3 - Más de 2 hasta 3 salarios mínimos
  - 4 - Más de 3 hasta 5 salarios mínimos
  - 5 - Más de 5 salarios mínimos
  - 6 - No recibe ingresos
  - 7 - No especificado

3.3.4 *Modelos de Predicción*

Para este análisis se entrenaron diversos modelos de predicción con el objetivo de comparar su desempeño y seleccionar aquel que ofreciera mejores resultados para la imputación del ingreso laboral. Como es práctica común en este tipo de tareas, se incluye la Regresión Lineal como modelo de referencia (benchmark), ya que su simplicidad permite establecer una línea base sobre la cual evaluar el desempeño de modelos más complejos.

Es importante tener en cuenta el tipo de mecanismo de datos faltantes al momento de seleccionar un enfoque de imputación. De acuerdo con la clasificación propuesta por Schafer y Graham [5], existen tres mecanismos fundamentales:

- **Completamente al azar (MCAR)**: la probabilidad de que un dato esté ausente es independiente tanto de los valores observados como de los no observados.
- **Al azar (MAR)**: la probabilidad de que un dato esté ausente puede depender de los valores observados, pero no de los valores faltantes en sí.
- **No al azar (MNAR)**: la ausencia de datos depende directamente de los valores faltantes, lo que representa el caso más complejo de manejar.

[5] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. DOI: 10.1037/1082-989X.7.2.147. URL https://doi.org/10.1037/1082-989X.7.2.147

De acuerdo con el análisis de la Nota Metodológica sobre la Imputación de Ingresos Laborales no Reportados en la ENOE, publicada en 2023 por la CONASAMI (Comisión Nacional de los Salarios Mínimos)[6], los datos faltantes en la variable de ingreso laboral presentan un comportamiento MAR, es decir, faltantes no completamente al azar. Esto se refleja en que "aquellos que no declaran su ingreso laboral tienden a estar en edades entre 25 y 55 años, presentan una menor proporción de mujeres, tienen mayores niveles de escolaridad y están ocupados en empleos formales." En otras palabras, la probabilidad de no declarar el ingreso está relacionada con el propio valor faltante del ingreso, así como con otras características sociodemográficas, lo cual implica que no se trata de una omisión aleatoria.

[6] CONEVAL. Imputación de ingresos no reportados en la enoe. https://www.gob.mx/cms/uploads/attachment/file/806698/Imputaci_n_de_ingresos_no_reportados_en_la_ENOE.pdf, 2022. Accedido en julio de 2025

Este tipo de faltantes impone un desafío particular, ya que descarta enfoques de imputación que asumen aleatoriedad, como la imputación por mediana o la asignación aleatoria de valores, los cuales no capturan la estructura real de los datos. Por ello, en este análisis se recurre a modelos capaces de identificar patrones complejos y no lineales, y que aprovechan la información observable para predecir ingresos de manera más precisa.

Teniendo esto en mente, los modelos seleccionados fueron los siguientes:

- **Linear Regression (Regresión Lineal):** Es uno de los métodos más simples y ampliamente utilizados. Este modelo asume que existe una relación lineal entre las variables predictoras (independientes) y la variable a predecir (dependiente). Su popularidad radica en su simplicidad y en su capacidad para interpretar cómo cada predictor afecta la predicción y en qué magnitud lo hace [7]. Debido a su

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009

naturaleza básica, a menudo se utiliza como modelo de referencia (benchmark), ya que los modelos más complejos deben superar su desempeño para justificar su uso.

- **K-Nearest Neighbors Regression (KNN):** Es un modelo bastante conocido y recurrido ya que permite realizar predicciones basándose en la similitud de las variables de cada registro. Su funcionamiento se basa en distancias (por lo general utiliza la distancia euclidiana) para realizar una predicción, al utilizar los valores de los "K"vecinos más cercanos en el espacio de variables. [8]
- **XGBoost Regression (Extreme Gradient Boost Regression):** Se trata de una implementación optimizada del anteriormente descrito Gradient Boosting. Realiza predicciones basándose en árboles de decisión y utiliza la técnica de Boosting para mejorar la precisión de los resultados. De forma adicional, aplica regularización para prevenir el sobreajuste (*overfitting*) de los datos. Asimismo, a diferencia del Gradient Boosting tradicional, la función de pérdida (Gradiente Descendiente) incorpora una aproximación de la segunda derivada para aumentar la rapidez de convergencia.[9]

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. ACM, 2016. DOI: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785

Dado que la evidencia sugiere un mecanismo MAR en los faltantes de ingreso, se incorporaron dos enfoques complementarios de imputación. En primer lugar, K-Nearest Neighbors (KNN), que estima el ingreso a partir de "vecindarios" de observaciones con perfiles similares (edad, escolaridad, ocupación, ámbito, etc.), sin imponer supuestos de linealidad y alineado con la idea de que los no-declarantes comparten características observables. En el ámbito económico, el algoritmo K-Nearest Neighbors (KNN) ha sido utilizado bajo el supuesto de que un registro pertenece al grupo al que se asemeja más, con base en la cercanía respecto a otras observaciones [10].

[10] Neha Thakur. Knn for classifying income, 2019. URL https://medium.com/@nehathakur912/knn-for-classifying-income-2a2b39d5984a. Medium. Accessed: 2025-08-17

En segundo lugar, XGBoost Regressor se emplea para capturar relaciones complejas e interacciones que trascienden la mera proximidad geométrica (uso de distancias como en KNN): al construir árboles de decisión de forma secuencial y regularizada, el modelo aprende patrones no lineales a partir de todo el conjunto de datos, es escalable a muestras grandes y robusto frente a sobreajuste, es por ello que se presenta como una propuesta frente a KNN. Esta combinación permite contrastar un método local (KNN, basado en similitud) con un método global (XGBoost, basado en boosting de árboles), fortaleciendo la calidad de las imputaciones individuales sin abandonar la interpretabilidad de las variables originales requerida por el análisis de Regresión Distributiva.

Con las diferentes opciones de modelos y selección de variables se estimaron nueve modelos diferentes, los cuales se muestran en la siguiente tabla:

Modelo	Método de Regresión	Selección de Variables	Número de variables
Modelo 1	Linear Regression	NA	95
Modelo 2	Linear Regression	Random Forests	10
Modelo 3	Linear Regression	Gradient Boosting Regressor	13
Modelo 4	K-Nearest Neighbors	NA	95
Modelo 5	K-Nearest Neighbors	Random Forests	10
Modelo 6	K-Nearest Neighbors	Gradient Boosting Regressor	13
Modelo 7	XGBoost Regressor	NA	95
Modelo 8	XGBoost Regressor	Random Forests	10
Modelo 9	XGBoost Regressor	Gradient Boosting Regressor	13

Cuadro 3.3: Características de los Modelos desarrollados

3.3.5 *Métricas de evaluación*

La finalidad de probar diferentes métodos de selección de variables y modelos predictivos es poder compararlos y decidir qué combinación resultó ser la más eficiente. Nuestros datos de prueba (testing) nos sirven para generar estas comparativas. Para ello, es necesario definir con qué métricas se va a generar dicha comparación. Al trabajar con modelos de regresión es común basarse en métricas que midan el error, es decir, qué tan distinta fue la predicción en relación con el valor verdadero. A continuación, se enuncian las métricas más utilizadas:

1. RMSE - Raíz del Error Cuadrático Medio: La Raíz del Error Cuadrático Medio (Root Mean Squared Error) mide la magnitud promedio de los errores al cuadrado entre los ingresos pronosticados y los reales. Este indicador penaliza más fuertemente los errores grandes y se define como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

donde:

- n: Número total de observaciones en el conjunto de testeo.
- y_i: Ingreso observado (real).
- $\hat{y}_i$: Ingreso pronosticado por el modelo.

2. MSE - Error Cuadrático Medio: El Error Cuadrático Medio (Mean Squared Error) mide el promedio de los errores al cuadrado. Es una métrica que refleja cuán alejados están, en promedio, los ingresos pronosticados de los ingresos reales. Su fórmula es:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3. MAE - Error Medio Absoluto: El Error Medio Absoluto (Mean Absolute Error) mide la magnitud promedio de los errores absolutos entre los ingresos pronosticados y los reales. Es más robusto frente a

valores atípicos en comparación con el RMSE y se calcula como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4. R^2 - Coeficiente de Determinación El Coeficiente de Determinación (R^2) evalúa qué tan bien el modelo explica la variabilidad en los ingresos reales en comparación con un modelo base (por ejemplo, el promedio de los ingresos reales). Su expresión es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde:

- $\bar{y}$: Media de los ingresos observados.

3.3.6 *Desempeño de modelos*

La tabla que se muestra a continuación contiene los resultados del año 2018 para cada combinación de modelo predictivo y selección de variables.

Modelo	**RMSE**	**MSE**	**MAE**	**R2**
Modelo 1	3,513.60	12,345,351.61	1,562.29	0.62
Modelo 2	3,603.37	12,984,289.03	1,575.06	0.60
Modelo 3	3,573.89	12,772,686.05	1,571.85	0.61
Modelo 4	5,129.02	26,306,807.74	2,950.49	0.19
Modelo 5	5,187.22	26,907,243.35	2,983.93	0.17
Modelo 6	4,526.56	20,489,704.77	2,240.53	0.37
Modelo 7	3,126.51	9,775,045.41	1,128.52	0.70
Modelo 8	3,332.21	11,103,618.60	1,170.04	0.66
Modelo 9	3,183.62	10,135,438.41	1,134.74	0.69

Cuadro 3.4: Tabla de métricas para último trimestre de 2018

Los resultados de la tabla anterior corresponden a las métricas de evaluación de cada uno de los 9 modelos estimados usando la base de datos correspondiente al cuarto trimestre del 2018.

Al aplicar los mismos modelos a los datos para el 2024, se obtienen los siguientes resultados:

Como se puede observar, los mejores resultados fueron obtenidos a través del modelo 7, que utiliza XGBoost Regressor con la totalidad de las variables disponibles. En ambos años analizados (2018 y 2024), esta configuración presenta uno de los Errores Medios Absolutos (MAE) más bajos y uno de los coeficientes de determinación R^2 más altos. Esto indica que las imputaciones generadas por este modelo presentan diferencias pequeñas respecto a los valores reales en los datos de prueba, y que el conjunto completo de variables utilizadas permite explicar en gran medida el ingreso.

Modelo	**RMSE**	**MSE**	**MAE**	**R2**
Modelo 1	4,118.36	16,960,883.39	2,319.94	0.75
Modelo 2	4,263.35	18,176,166.33	2,433.50	0.73
Modelo 3	4,219.64	17,805,326.81	2,389.22	0.74
Modelo 4	7,676.90	58,934,735.80	4,481.48	0.14
Modelo 5	5,842.16	34,130,823.65	2,923.09	0.50
Modelo 6	6,598.96	43,546,290.48	3,407.46	0.36
Modelo 7	3,274.25	10,720,693.26	1,651.40	0.84
Modelo 8	3,519.34	12,385,740.50	1,830.19	0.82
Modelo 9	3,350.17	11,223,659.29	1,691.73	0.84

Cuadro 3.5: Tabla de métricas para segundo trimestre de 2024

No obstante, también es importante destacar el buen desempeño del modelo 9, el cual considera únicamente 13 variables seleccionadas mediante Gradient Boosting. Este modelo logra capturar adecuadamente la estructura de los datos y genera predicciones comparables en calidad a las del modelo 7. La principal ventaja del modelo 9 radica en su eficiencia computacional: al requerir un menor número de variables, su ejecución es más rápida. Sin embargo, el modelo 7 también resulta lo suficientemente eficiente en términos de tiempo de cómputo, por lo que ambos modelos son opciones viables dependiendo de las prioridades específicas del análisis (precisión vs. eficiencia).

Como dato adicional el método de selección de variables arrojó que la siguiente fue identificada como la más significativa:

- ing7c: esta variable mide el nivel de ingresos, contando con los siguientes valores

0	No Aplica
1	Hasta un salario mínimo
2	Más de 1 hasta 2 salarios mínimos
3	Más de 2 hasta 3 salarios mínimos
4	Más de 3 hasta 5 salarios mínimos
5	Más de 5 salarios mínimos
6	No recibe ingresos
7	No especificado

Cuadro 3.6: Valores de la variable 'ing7c'

Aportando a esta comparación, se incluyen los siguientes gráficos de puntos. En ellos se realiza una comparación de los datos reales vs los datos de testeo, en donde si todas las predicciones coincidieran al 100 % con los datos reales, todos los puntos se encontrarían sobre la línea ideal, marcada en rojo. Se incluye una primera gráfica conteniendo todos los datos de testeo, así como, una segunda gráfica en donde se eliminan los *outliers* de los datos de testeo para apreciar con más detalle las predicciones. Estos *outliers* están identificados y filtrados

haciendo uso del Rango Intercuartílico (IQR), para ello se divide la información de las diferencias entre las observaciones reales (`y_test`) y las predicciones (`y_hat`) en 4 cuartiles. El valor de IQR será determinado por la diferencia del Q3 menos el Q1. Así, se definen los límites inferior Q1-1.5×IQR y superior Q3+1.5×IQR. Todos los valores que caen fuera de estos límites serán considerados como outliers.

Es importante aclarar que los outliers solo se eliminan para la visualización gráfica, pero están presentes en el análisis efectuado.

Para el modelo con peor desempeño en las métricas de evaluación, el gráfico muestra poca consistencia, existe una dispersión significativa, en particular peor para los valores más altos del ingreso. Se observa también que un considerable número de predicciones se encuentran muy alejadas del valor real.

En cuanto a los modelos con mejor desempeño para las métricas de evaluación, observamos que en los modelos 7, 8 y 9 las predicciones muestran una estratificación correspondiente con la variable '**ing7c**'. Ambos modelos demuestran apegarse más a la línea ideal, confirmando que sus predicciones son más apegadas a los datos reales. Se puede identificar que ambos modelos tienden a sobrestimar los ingresos a partir de cierto rango, siendo el modelo 7 más robusto cuando no hay presencia de outliers; sin embargo, los tres adecúan bastante bien los diferentes segmentos del ingreso.

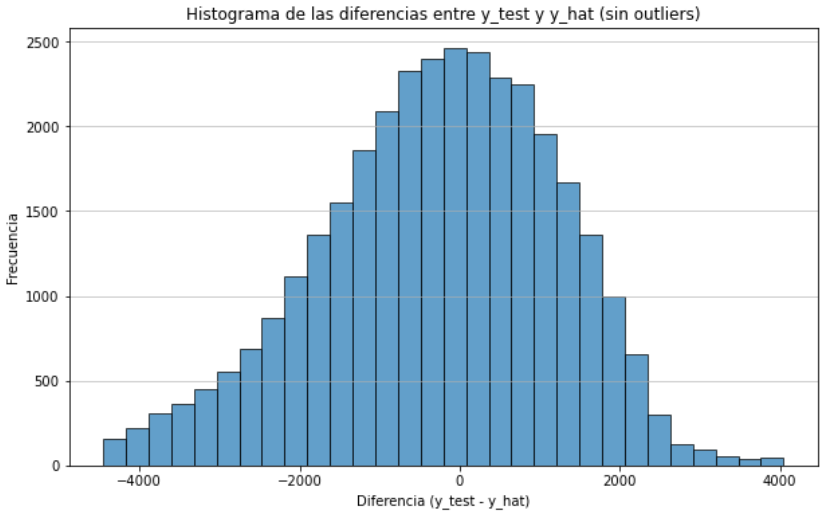

Figura 3.5: Modelo 1 - Histograma del error (sin outliers)

3.4 *Resultados de la imputación*

Existe un total de 157,210 registros en la base para el último trimestre del 2018, de los cuales 42,802 no cuentan con información del ingreso.

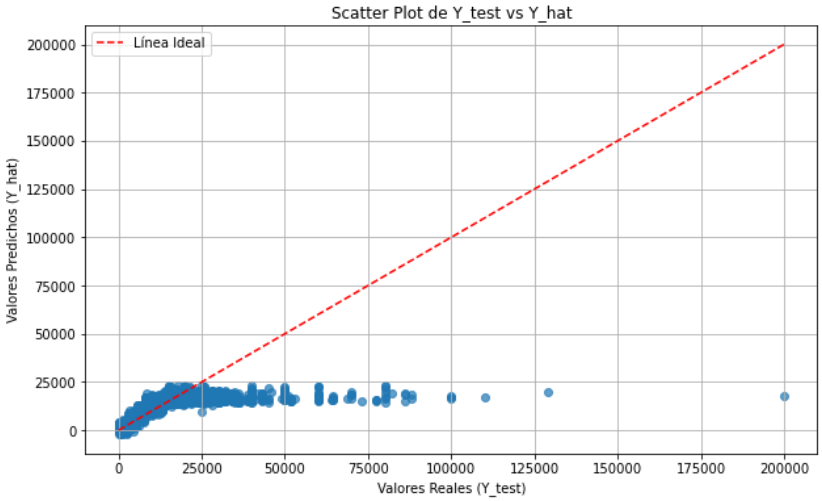

Figura 3.6: Modelo 1 - Datos reales vs Datos predicción

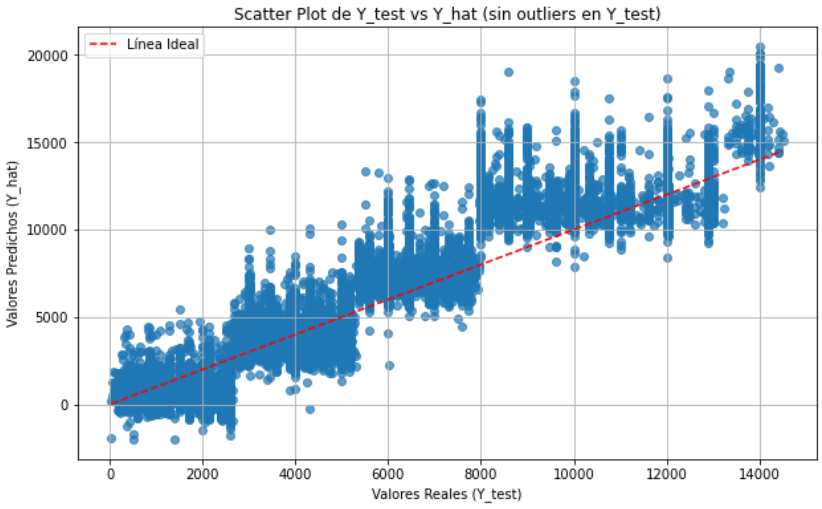

Figura 3.7: Modelo 1 - Datos reales vs Datos predicción (sin outliers)

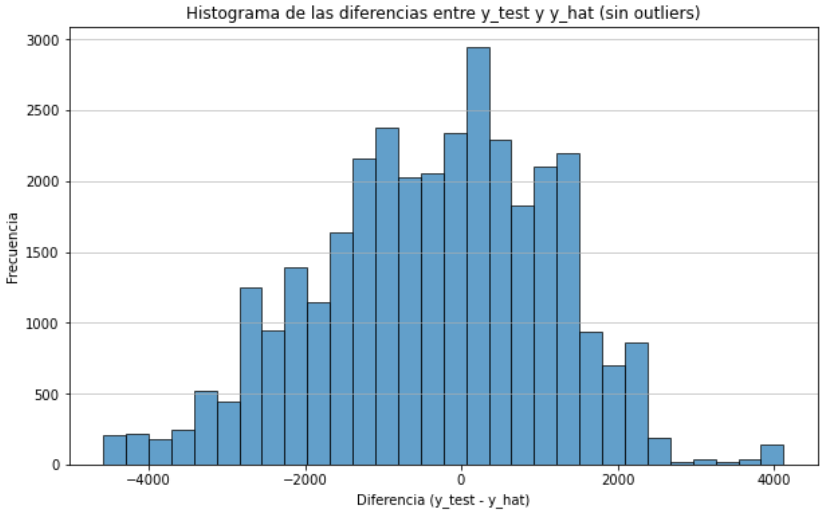

Figura 3.8: Modelo 2 - Histograma del error (sin outliers)

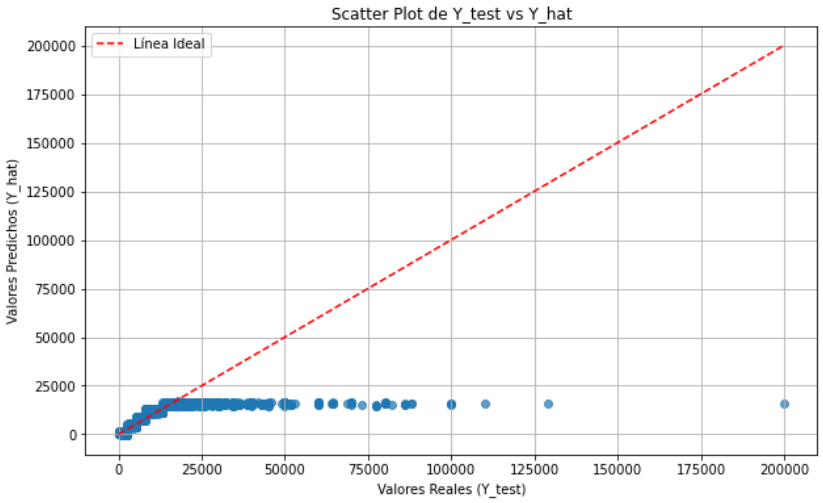

Figura 3.9: Modelo 2 - Datos reales vs Datos predicción

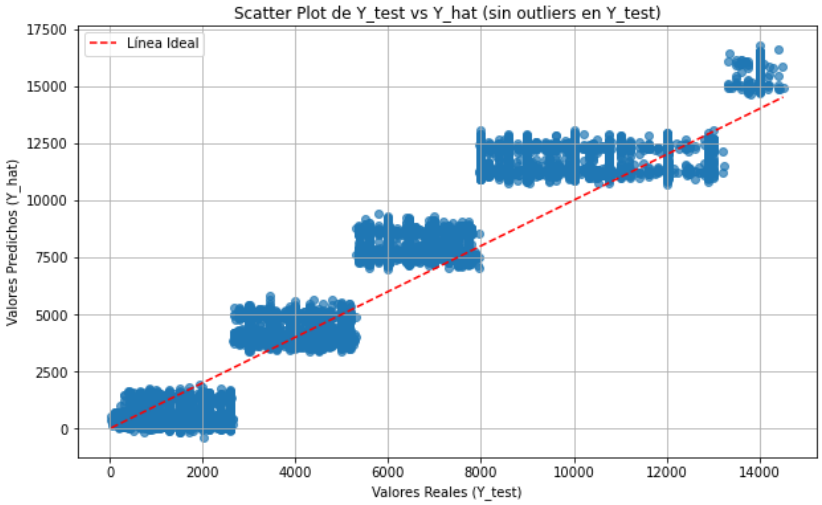

Figura 3.10: Modelo 2 - Datos reales vs Datos predicción (sin outliers)

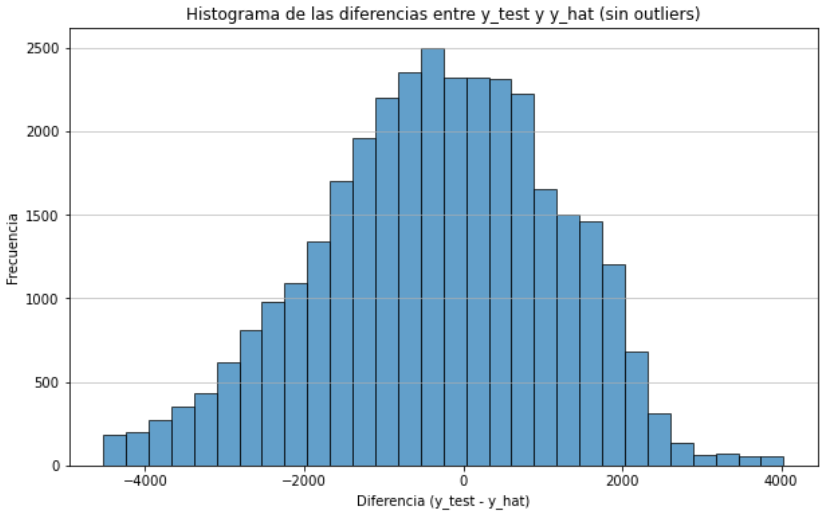

Figura 3.11: Modelo 3 - Histograma del error (sin outliers)

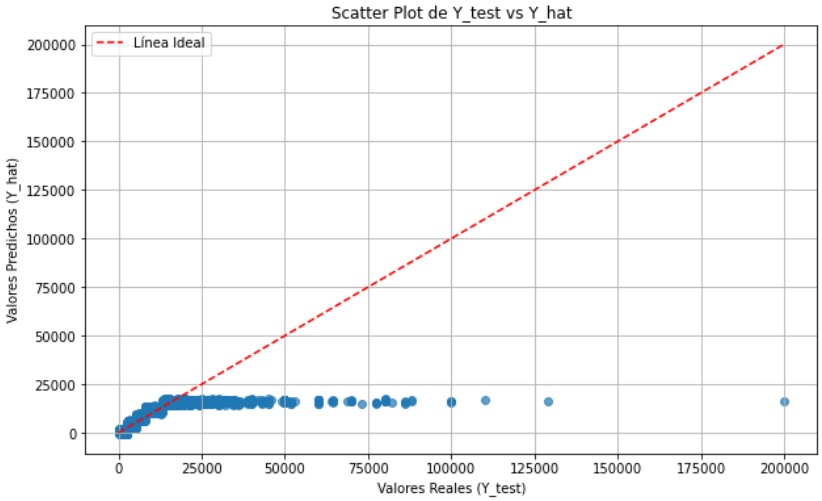

Figura 3.12: Modelo 3 - Datos reales vs Datos predicción

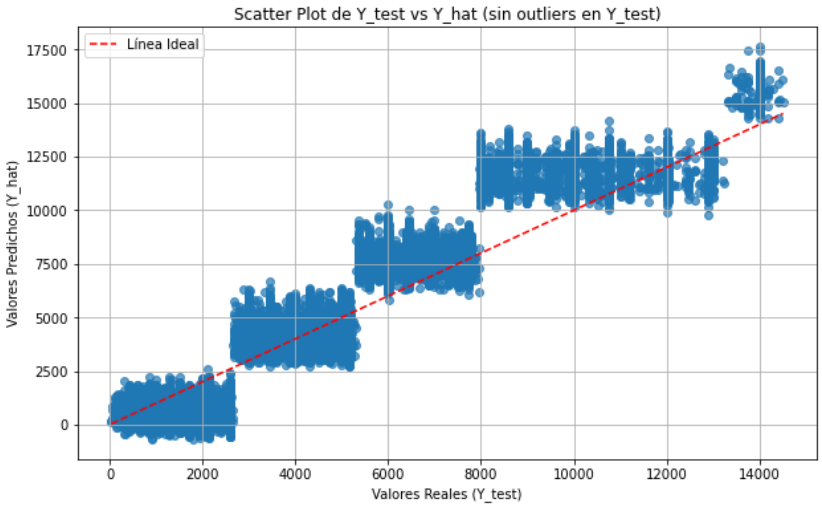

Figura 3.13: Modelo 3 - Datos reales vs Datos predicción (sin outliers)

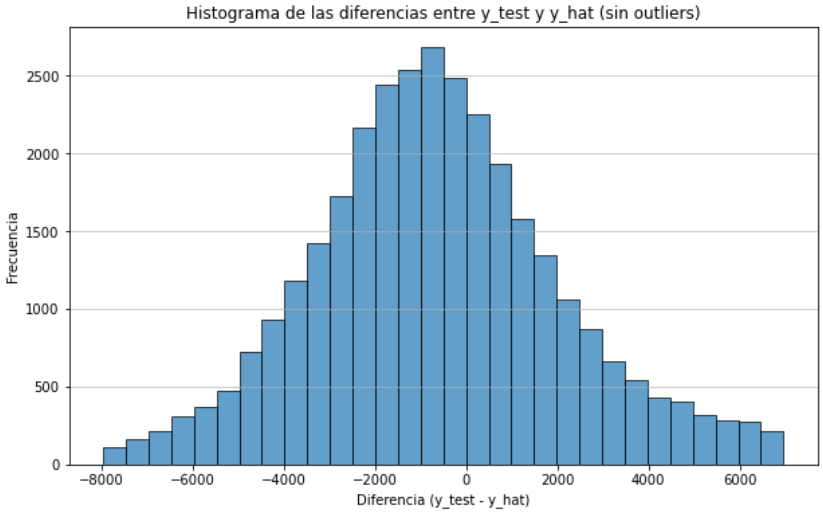

Figura 3.14: Modelo 4 - Histograma del error (sin outliers)

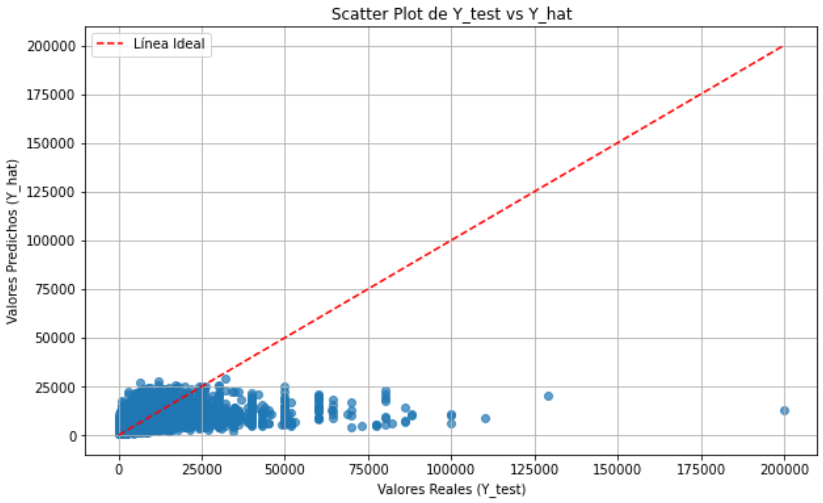

Figura 3.15: Modelo 4 - Datos reales vs Datos predicción

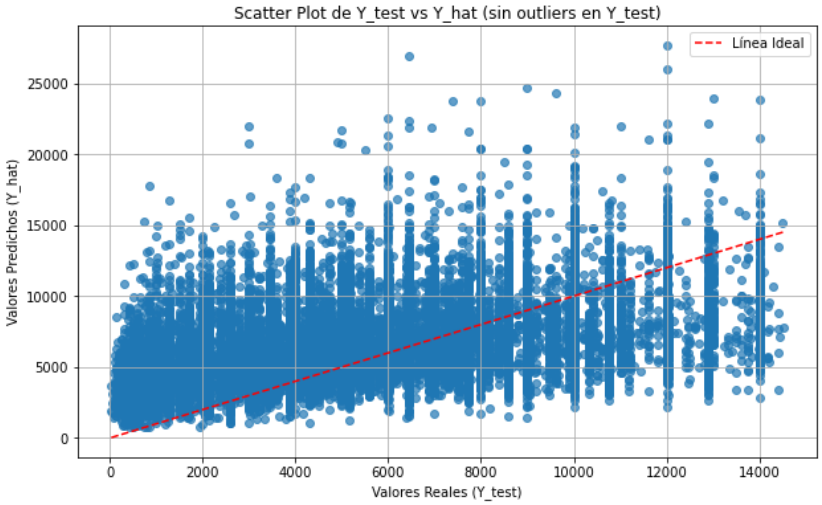

Figura 3.16: Modelo 4 - Datos reales vs Datos predicción (sin outliers)

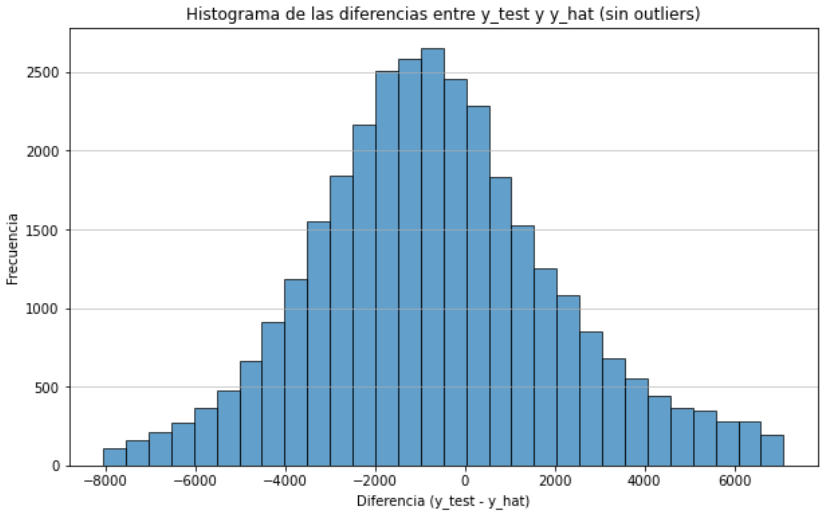

Figura 3.17: Modelo 5 - Histograma del error (sin outliers)

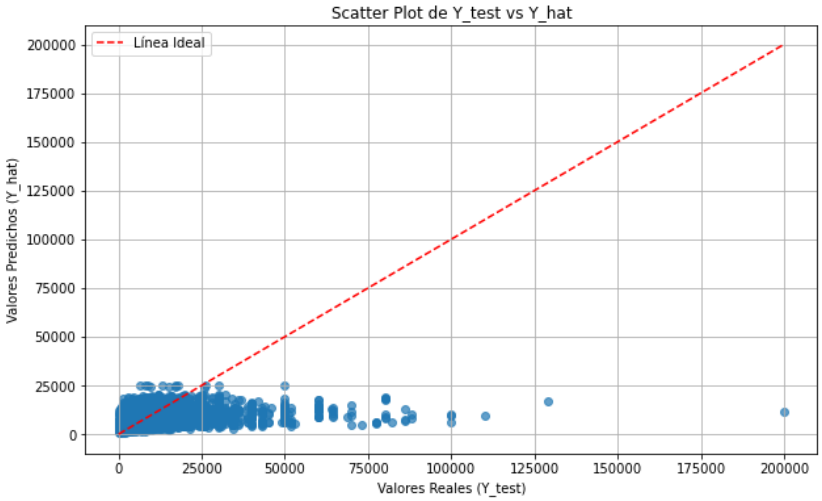

Figura 3.18: Modelo 5 - Datos reales vs Datos predicción

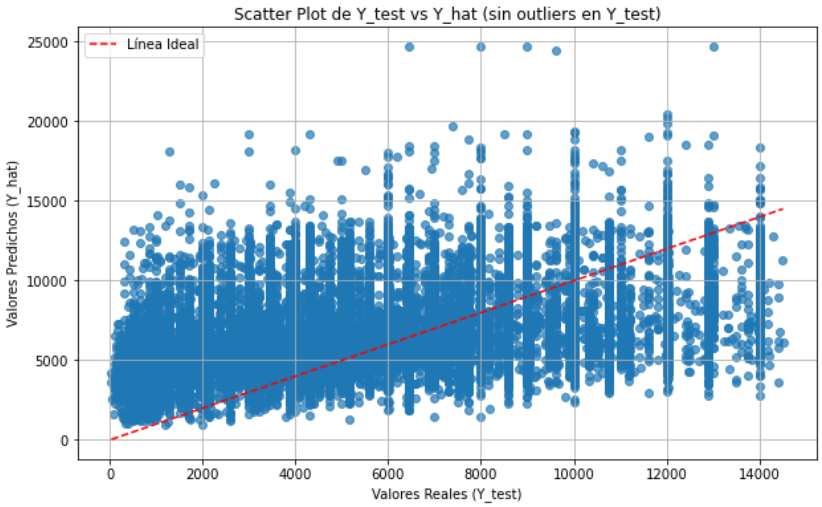

Figura 3.19: Modelo 5 - Datos reales vs Datos predicción (sin outliers para datos de testeo)

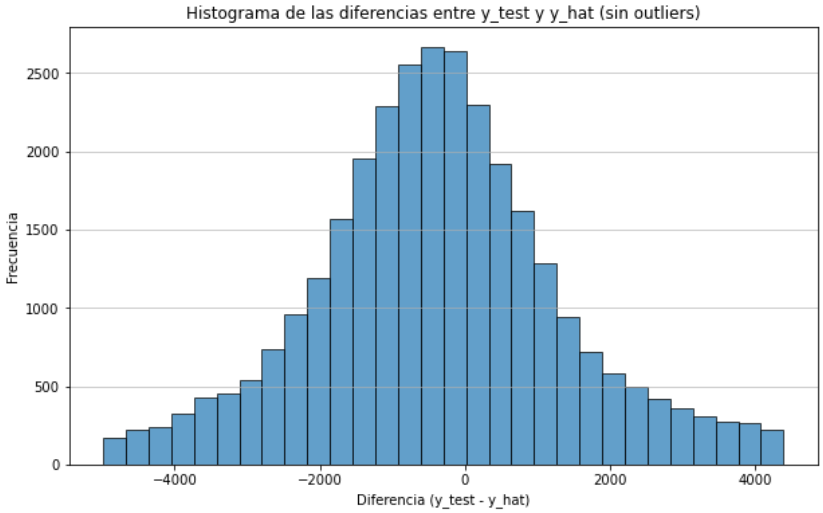

Figura 3.20: Modelo 6 - Histograma del error (sin outliers)

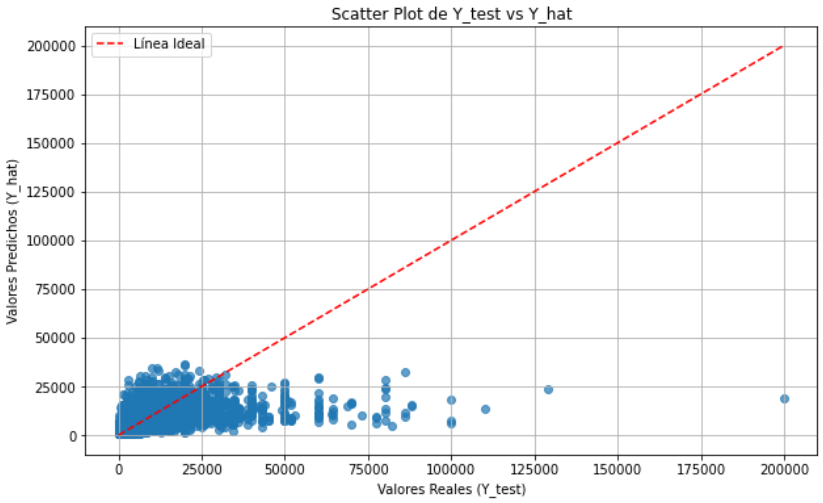

Figura 3.21: Modelo 6 - Datos reales vs Datos predicción

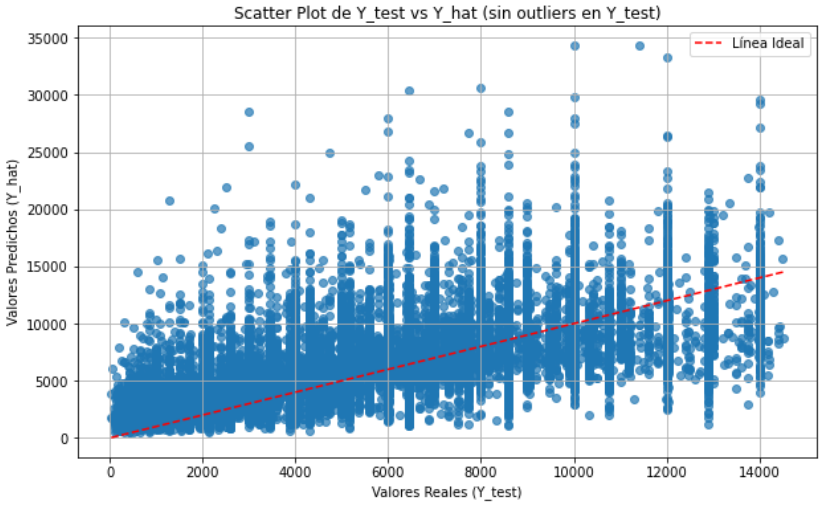

Figura 3.22: Modelo 6 - Datos reales vs Datos predicción (sin outliers)

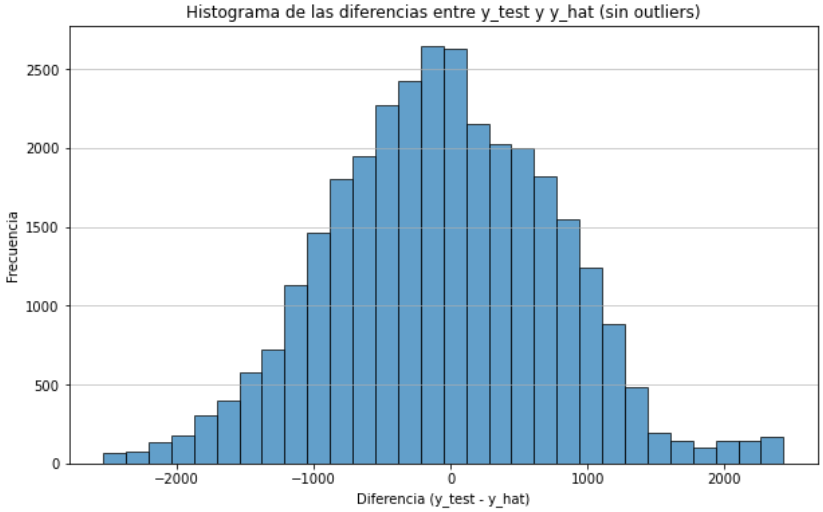

Figura 3.23: Modelo 7 - Histograma de error (sin outliers)

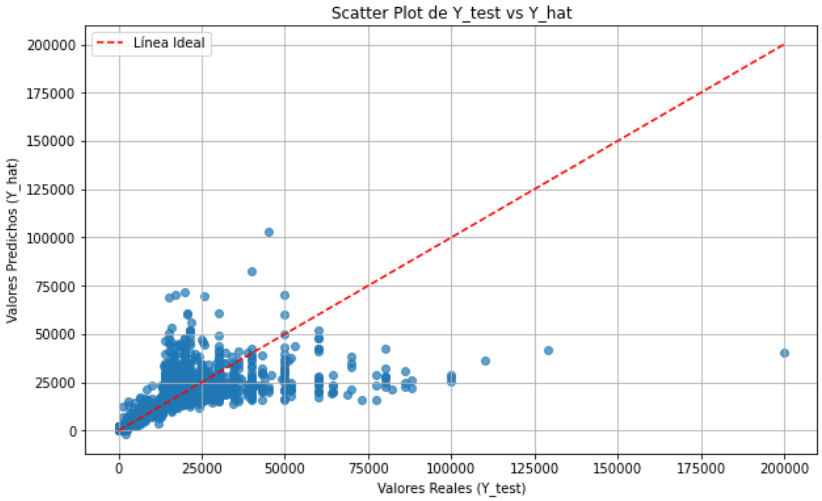

Figura 3.24: Modelo 7 - Datos reales vs Datos predicción

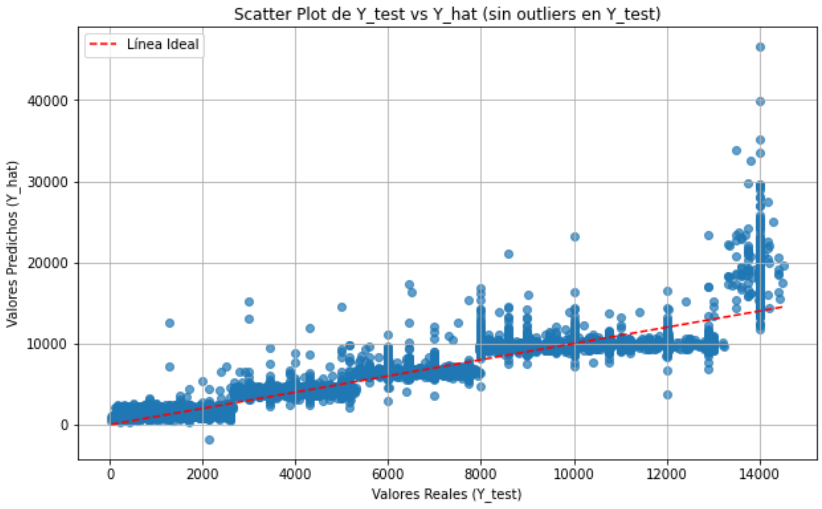

Figura 3.25: Modelo 7 - Datos reales vs Datos predicción (sin outliers)

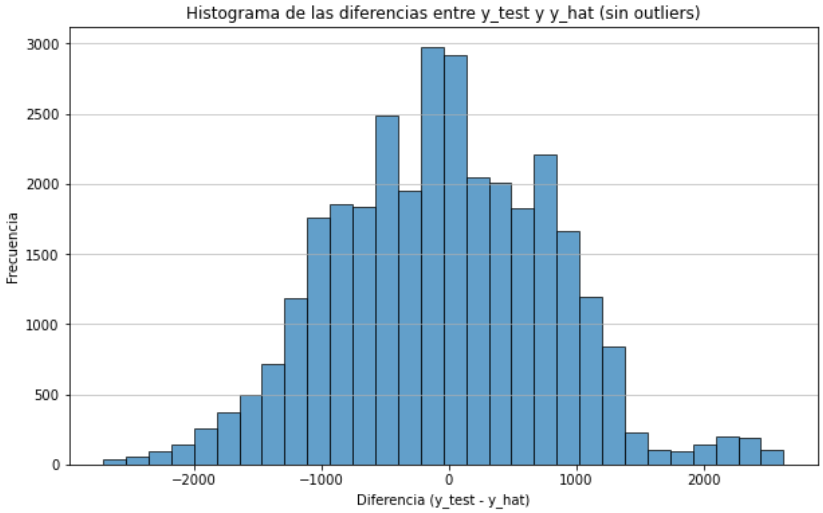

Figura 3.26: Modelo 8 - Histograma del error (sin outliers)

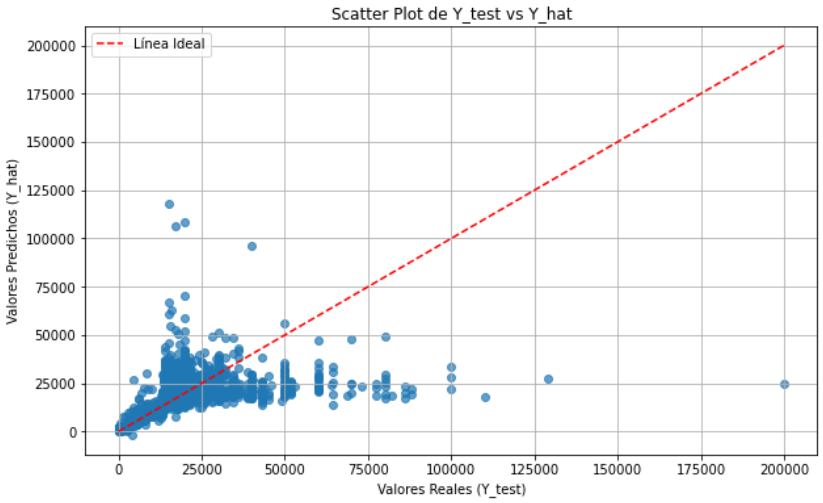

Figura 3.27: Modelo 8 - Datos reales vs Datos predicción

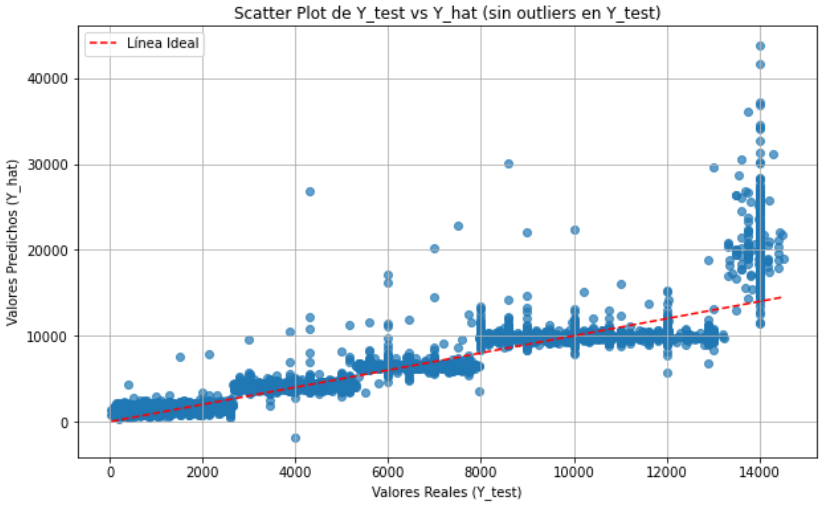

Figura 3.28: Modelo 8 - Datos reales vs Datos predicción (sin outliers)

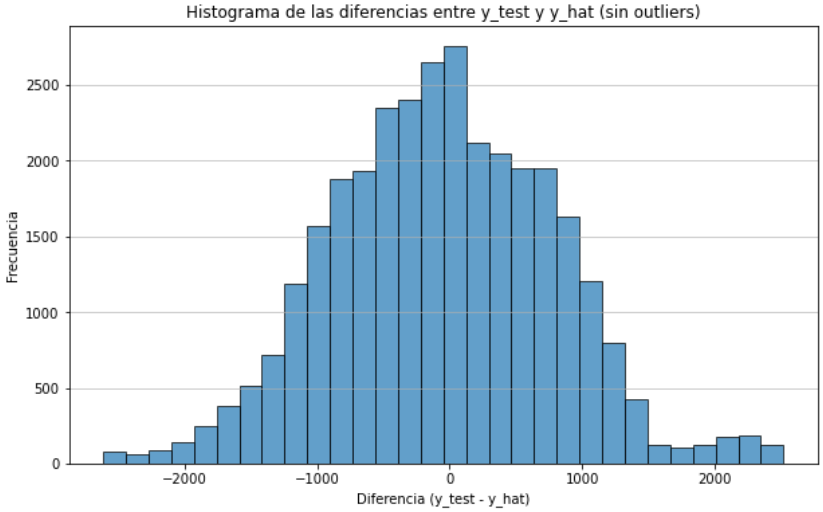

Figura 3.29: Modelo 9 - Histograma de error (sin outliers)

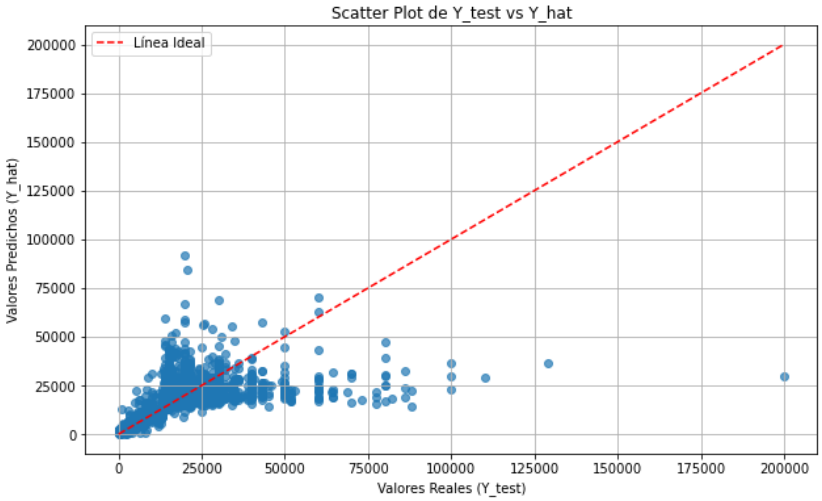

Figura 3.30: Modelo 9 - Datos reales vs Datos predicción

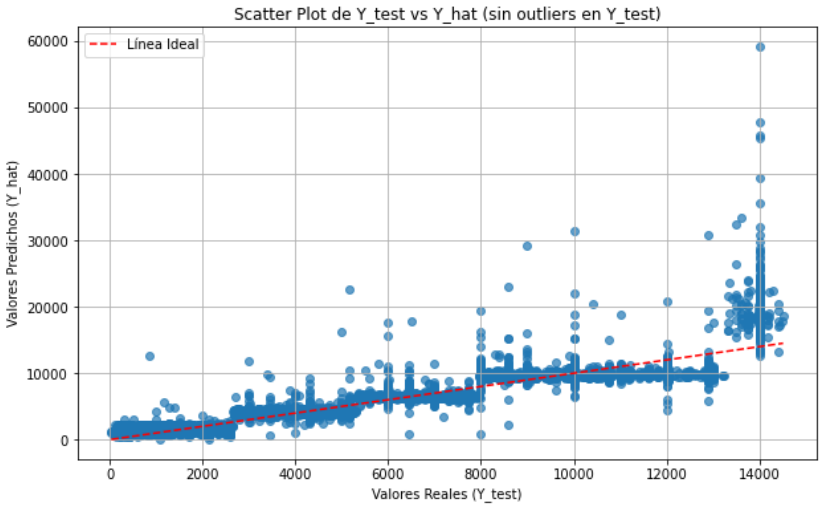

Figura 3.31: Modelo 9 - Datos reales vs Datos predicción (sin outliers)

De este total, existen 24,913 registros que no contestaron la pregunta referente al rango del salario (variable ing7c), los restantes sí declararon ubicarse en algún rango salarial. Estos registros que sí declararon ubicarse en algún rango salarial representan el 41.79 % del total de datos a predecir.

Para probar la viabilidad del modelo en la predicción del ingreso se utilizaron los 17,889 que no declararon ingreso, pero sí rango salarial y se validó que el ingreso predicho por el modelo cayera dentro del rango salarial declarado. Para cada uno de los diferentes modelos se llevó a cabo este procedimiento, obteniendo los siguientes resultados:

Modelo	Aciertos	Errores	Totales	% Aciertos
1	14,916	2,973	17,889	83.38
2	15,835	2,054	17,889	88.51
3	15,401	2,488	17,889	86.09
4	5,538	12,351	17,889	30.95
5	5,538	12,351	17,889	30.95
6	7,396	10,493	17,889	41.34
7	17,580	309	17,889	98.27
8	17,701	188	17,889	98.94
9	17,603	286	17,889	98.4

Cuadro 3.7: Ingresos predichos 2018 - Aciertos en los rangos salariales de la predicción

Al aplicar este mismo análisis a los datos del 2024, se tuvo que eliminar los registros correspondientes a la zona fronteriza, pues es importante recordar que a partir del 2019 se introdujo un salario mínimo distinto para dicha región. Con esta distinción, obtenemos la información siguiente:

Modelo	Aciertos	Errores	Totales	% Aciertos
1	23,533	291	23,824	98.77
2	23,687	137	23,824	99.42
3	23,745	79	23,824	99.66
4	9,753	14,071	23,824	40.93
5	12,011	11,813	23,824	50.4
6	14,129	9,695	23,824	59.3
7	23,761	63	23,824	99.73
8	23,823	1	23,824	99.99
9	23,786	38	23,824	99.84

Cuadro 3.8: Ingresos predichos 2024 - Aciertos en los rangos salariales de la predicción

En resumen, al evaluar el porcentaje de aciertos respecto a los rangos salariales, las métricas obtenidas y las gráficas generadas para cada modelo, tres de los nueve modelos se destacan por mostrar un desempeño satisfactorio: el modelo 7, el modelo 8 y el modelo 9. Todos logran capturar adecuadamente la estratificación del ingreso en rangos salariales, y los resultados del coeficiente de determinación R^2 indican que las variables utilizadas explican en gran medida el ingreso predicho.

Es importante señalar que estos tres modelos comparten una característica en común: todos utilizan el algoritmo XGBoost Regressor, lo cual sugiere que este tipo de regresión es especialmente eficaz para capturar las relaciones no lineales y complejas entre las variables predictoras y el ingreso.

XGBoost es un algoritmo ampliamente utilizado en la ciencia de datos aplicada debido a su capacidad para modelar relaciones no lineales y de alta complejidad entre múltiples variables. A diferencia de modelos más rígidos como la regresión lineal, XGBoost se basa en árboles de decisión ensamblados secuencialmente mediante técnicas de *boosting*, lo que le permite corregir los errores de modelos anteriores y mejorar progresivamente la precisión. Además, XGBoost cuenta con una arquitectura altamente optimizada que lo hace muy eficiente computacionalmente, permitiendo su escalabilidad incluso en conjuntos de datos grandes como los utilizados en este análisis. En escenarios de alta dimensionalidad, donde hay muchas variables predictoras, es capaz de explorar espacios complejos (*high-dimensional feature spaces*) sin perder precisión. También incorpora técnicas integradas para combatir el sobreajuste (*overfitting*), como la regularización (penalización del modelo para evitar que se ajuste demasiado a los datos de entrenamiento) y el *pruning* (poda de ramas irrelevantes en los árboles de decisión para evitar estructuras excesivamente profundas). Otra ventaja operativa es que puede tolerar datos faltantes en las variables predictoras, lo que lo hace particularmente robusto en contextos donde no siempre es posible contar con información completa. [11]

En suma, el buen desempeño observado en los modelos que implementan XGBoost no solo obedece a la calidad de los datos, sino también a la arquitectura del algoritmo, que resulta especialmente adecuada para tareas de imputación en encuestas complejas y con múltiples dimensiones de análisis.

Por el contrario, los modelos que implementan el método de los K Vecinos Más Cercanos (KNN) presentaron el peor desempeño. Incluso los modelos basados en Regresión lineal lograron resultados superiores, lo que indica que las técnicas de regresión, en general, son más adecuadas para este problema de imputación.

Destaca particularmente el desempeño del modelo 7, que mostró resultados consistentemente altos en ambos períodos evaluados. No obstante, cabe destacar que el modelo 9, que emplea solo un subconjunto de variables seleccionadas, también ofrece un desempeño muy competitivo, con métricas apenas inferiores a las del modelo 7. Su principal ventaja radica en su mayor eficiencia computacional, al requerir menos variables para realizar predicciones. Por ello, ambos modelos podrían considerarse apropiados, dependiendo de

[11] Stefano Bilotta, Luciano Alessandro Ipsaro Palesi, and Paolo Nesi. Exploiting open data for co estimation via artificial intelligence and explainable ai. *Expert Systems with Applications*, 291:128598, 2025. DOI: 10.1016/j.eswa.2024.128598

las necesidades específicas de análisis.

Dado su rendimiento consistente en ambos períodos evaluados y su capacidad para incorporar el total de la información disponible sin comprometer la robustez, se eligió el modelo 7 como el principal insumo para la imputación del ingreso en los análisis posteriores de esta investigación.

3.5 *Términos Reales*

Como último paso en nuestro preprocesamiento de datos fue necesario modificar el valor del ingreso para quitarle el efecto de la inflación y hacer esta variable comparable a través del tiempo. La encuesta proporciona los datos en términos nominales, no obstante, para efectos de este análisis es imperativo restar la inflación de un año a otro para poder cumplir con el objetivo de comparar el ingreso de la población en el año 2018 respecto al 2024.

En este sentido, se ha agregado una columna adicional 'ingocup_real' a nuestra base de datos, misma que contiene el valor real del ingreso tomando como base la segunda quincena de Julio 2018. Una vez contando con los valores deflactados, es posible proceder al análisis estadístico descriptivo.

3.6 *Discusión de resultados sobre la imputación*

Si bien en capítulos anteriores se justificó la necesidad de imputar datos, ahora se profundiza en la importancia de utilizar métodos que consideren la estructura de la información. Métodos como la imputación por mediana hacen suposiciones que no necesariamente reflejan la estructura salarial de la población. Como se evidencia en el análisis previo, quienes no reportan ingresos no son individuos aleatorios dentro de la encuesta, sino que siguen ciertos patrones demográficos. Por ello, es crucial emplear técnicas que consideren estas características. La literatura sobre imputación de ingresos no reportados, como la de Campos-Vázquez[12], presenta diversas metodologías. Entre ellas, el Pareamiento por Puntajes de Propensión y el método de Hot Deck, que son estrategias valiosas para corregir la falta de datos y reflejar con mayor precisión la distribución del ingreso. Sin embargo, su análisis no proporciona métricas claras sobre la exactitud de las imputaciones, como medidas de error o precisión. Por lo tanto, aunque constituyen soluciones prácticas, incluir métricas de validación es fundamental para garantizar la calidad de dichas imputaciones.

Además, Campos-Vázquez[13] también señala que las diferencias entre métodos de imputación simple y múltiple no resultan significativas en los resultados agregados, como el ingreso promedio o la medición

[12] Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320): 803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003

[13] Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320): 803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003

de pobreza. A continuación se muestra un pequeño resumen de los modelos descritos por Campos-Vázquez, así como las principales ventajas y desventajas que mencionan:

1. Pareamiento por Puntajes de Propensión

   a) Basado en métodos probabilísticos.

   b) Utiliza un modelo para determinar la pertenencia a un grupo y selecciona el vecino más cercano (o sea, k=1).

   c) Como ventajas mencionan su fácil interpretación y que el ingreso imputado toma un valor observado en la muestra.

   d) Como desventaja mencionan que es computacionalmente intensivo.

2. Método Hot Deck

   a) Se crean cuatro grupos (hombre/mujer y formal/informal) y se asignan ingresos válidos de manera aleatoria con reemplazo.

   b) Como ventajas mencionan nuevamente que los valores imputados corresponden a valores observados, así como, su facilidad de aplicación en imputación única o múltiple y su eficiencia computacional.

3. Imputación por Grupos con Aleatoriedad

   a) Calcula la mediana de cada grupo y le suma una desviación estándar ajustada.

   b) Dentro de las ventajas se menciona que su sencilla implementación y rápida ejecución.

   c) Se mencionan como desventajas que se asume que la mediana y la desviación estándar observadas son correctas, sin considerar incertidumbre o distribución real.

4. Pareamiento por Promedios Predictivos

   a) Basado en modelos predictivos entrenados con los ingresos observados.

   b) Se asigna el promedio de los valores predichos dentro de un grupo.

   c) Ventajas: valores imputados a partir de datos observados, fácil implementación en software estadístico y robustez estadística.

   d) Desventajas: computacionalmente intensivo.

Evaluando estos métodos y analizando los resultados que Campos presentan, se puede ver que compara los ingresos imputados con los observados, reportando diferencias mínimas entre métodos. Por

ejemplo, el ingreso promedio observado en 2005 T1 fue de $5,500, mientras que el imputado con Hot Deck fue de $5,600, lo que representa una diferencia del 1.82 %. Para 2012 T3, el ingreso observado fue de $4,800 y el imputado de $5,100, con una diferencia del 6.25 %. Aunque concluye que los cálculos de ingreso, pobreza y desigualdad son robustos ante estos 4 métodos, hace la recomendación del método Hot-Deck. No obstante, se basa más en su facilidad de aplicación y tiempo estimado de cómputo que en ventajas cuantitativas como métricas de desempeño. Por su parte, Durán[14] sugiere que el Pareamiento por Promedios Predictivos incluya un elemento iterativo para capturar la incertidumbre del proceso, aunque esto puede afectar la convergencia del modelo; finalmente, recomienda el uso de este método para imputar. Además, presenta una tabla con métricas de evaluación como R^2, MAE y RMSE. En este trabajo, se comparó la R^2 con la de Durán: el modelo que recomienda MICE pmm (Multivariate Imputation by Chained Equations usando el método de imputación de Pareamiento por Medias Predictivas) obtuvo un R^2 de 0.258, mientras que nuestro modelo 7, que utiliza XGBoost Regressor para la predicción, alcanzó un R^2 de 0.70, demostrando un mejor desempeño en la predicción. Es importante notar que estas métricas son tomadas del análisis de diferentes años; en el caso de este trabajo se analiza el desempeño de la predicción para el último trimestre del 2018, mientras que Durán analiza años entre el 2005 y hasta el 2017, tomando únicamente conjuntos de datos completos. Sin embargo, este coeficiente de determinación da explicación de qué tanto las variables utilizadas están sirviendo para explicar el modelo y generar una predicción, por lo que su comparación resulta interesante para los efectos de este análisis. En conclusión, la selección del método de imputación debe considerar no solo su facilidad de aplicación, sino también su capacidad para capturar la estructura de la información y su desempeño en términos de métricas cuantitativas. La evolución hacia modelos con validaciones más robustas es clave para mejorar los estudios sobre ingresos y desigualdad. Es importante señalar que, si bien en ocasiones se argumenta que valores elevados de R^2 pueden deberse simplemente al gran número de variables utilizadas, este análisis proporciona evidencia en contra de dicha suposición. En primer lugar, al comparar las métricas reportadas por Durán con los resultados obtenidos en este trabajo, se observa que, a pesar de contar también con una muestra amplia y un número considerable de variables, el desempeño de los modelos utilizados por Durán fue notablemente inferior. En segundo lugar, incluso dentro de los modelos desarrollados en este análisis, aquellos que emplearon el mismo conjunto de variables, como KNN o regresión lineal, obtuvieron valores de R^2 considerablemente más bajos. Esto indica que no es el número de variables lo que determina la calidad predictiva, sino la

[14] Benito Durán Romo. Comparación de metodologías de imputación aplicadas a ingresos laborales de la enoe. *Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía*, 10:4–27, 2019

capacidad del modelo para capturar relaciones complejas, posiblemente no lineales, entre las variables explicativas y el ingreso.

Finalmente, es importante señalar que, aunque métodos como KNN podrían parecer adecuados para tareas de imputación al basarse en la similitud entre observaciones, los resultados obtenidos muestran que los modelos de regresión, particularmente aquellos basados en técnicas como XGBoost Regressor, ofrecen un desempeño considerablemente superior. Esto sugiere que el problema no debe abordarse como uno de agrupamiento o clasificación basada en cercanía (clustering), sino más bien como un problema de modelado estructurado, donde es crucial capturar relaciones no lineales y complejas entre variables para predecir con precisión los ingresos no reportados. La elección de modelos de regresión permite no solo mejores métricas de desempeño, sino también una mejor incorporación de las múltiples dimensiones demográficas y laborales relevantes en el fenómeno de la no declaración de ingresos.

4 Estadística Descriptiva

Contenidos

4.1 Ingreso Promedio por modelo

Al contar con bases de datos completas y comparables a través del tiempo, es posible calcular estadísticas que permiten una mejor comprensión de la composición del ingreso. En este análisis se utilizó el factor de expansión proporcionado por la ENOE, lo cual permite obtener estimaciones representativas de la población.

Es importante señalar que los cálculos de ingreso promedio que se presentan a continuación consideran únicamente a personas ocupadas que reportaron haber trabajado al menos una hora durante la semana de referencia y haber percibido una remuneración (ingreso) por dicho trabajo. En otras palabras, se excluyen del análisis los registros pertenecientes a la población no económicamente activa (PNEA) o a personas sin remuneración, conforme al proceso de limpieza de datos previamente detallado.

Asimismo, cabe aclarar que al comparar la base de datos imputada con la no imputada, los registros con datos faltantes en esta última no son tratados como ceros, sino que son excluidos del análisis de ingreso promedio.

Al comparar el ingreso promedio obtenido al imputar con cada uno de los diferentes modelos, podemos observar que, de forma general, aquellos modelos con mejor desempeño (según las métricas analizadas en el capítulo anterior) son también los que producen promedios de ingreso más altos. En todos los casos, el ingreso promedio imputado supera al obtenido con la base de datos original (sin imputación). Esto refuerza el supuesto de que la población que no declara su ingreso no es aleatoria, sino que tiende a concentrarse en los deciles más altos de

la distribución. Asimismo, la diferencia entre el ingreso imputado y el no imputado es mayor en el segundo trimestre de 2024 en comparación con el último trimestre de 2018, lo que sugiere un incremento en la proporción de personas que omiten declarar su ingreso en años recientes.

Como se especificó anteriormente, para los análisis posteriores se utilizará el modelo 7, dado su buen desempeño en términos de ajuste y coherencia con los patrones observados.

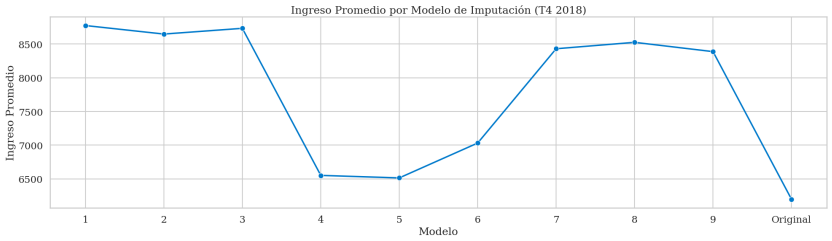

Figura 4.1: Ingreso Promedio por Modelo de Imputación (T4 2018)

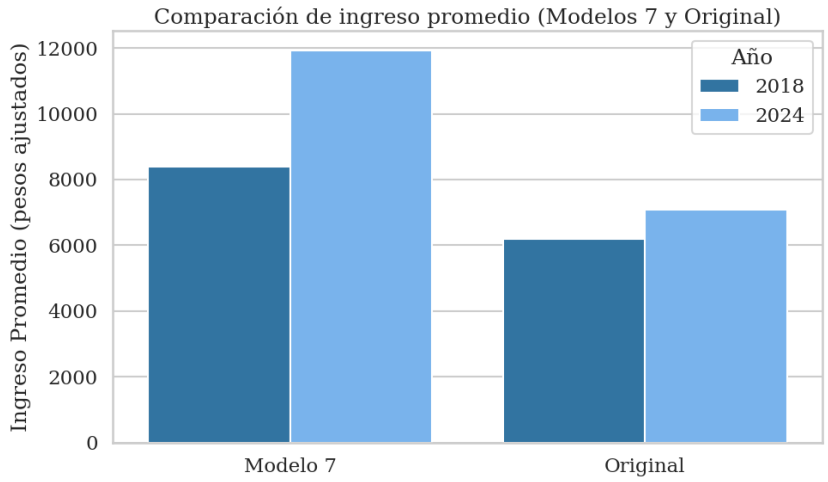

Figura 4.2: Comparación de ingreso promedio (Modelos 7 y Original - Precios base 2018

4.2 *Ingreso promedio 2018 vs 2024*

Una vez seleccionado el modelo de imputación, se procede al análisis del ingreso promedio para distintos subconjuntos de la población, tanto en el último trimestre de 2018 como en el segundo trimestre de 2024. Para ello, se utilizaron las bases de datos con ingresos imputados mediante el modelo 7, y se contrastaron con los valores originales (sin imputar). Es importante recalcar que en este análisis, ya se ha traído la información a precios reales del 2018 para hacer efectiva la comparación de las cifras.

Los resultados muestran que existe una diferencia consistente entre las medias del ingreso imputado y las del ingreso sin imputar. Esta brecha, además, se incrementa de forma notable en el segundo trimestre de 2024. Esto sugiere que en años recientes es mayor el número de personas que no declaran su ingreso, lo cual podría estar sesgando los análisis si no se realiza una imputación adecuada.

Respecto al impacto del incremento salarial implementado en 2019, los resultados permiten observar un aumento general del 41 % en el ingreso promedio de la población para el año 2024, en comparación con 2018. De forma general, esta es una cifra a tener en mente durante el siguiente capítulo de esta investigación.

Cabe destacar que todas las estadísticas se calcularon utilizando los factores de expansión proporcionados por la misma encuesta, con el fin de representar adecuadamente a la población nacional. Asimismo, los promedios del ingreso sin imputar se calcularon excluyendo los valores faltantes (NaN), sin asignarles un valor de cero.

Estadístico para la media del ingreso	2018 S/ Imputación	2018 con Imputación	Incremento
General	$ 6,198.98	$ 8,429.92	36 %
Hombres	$ 6,776.69	$ 8,975.36	32 %
Mujeres	$ 5,283.90	$ 7,548.86	43 %
Personas Alfabetizadas	$ 6,292.93	$ 8,532.29	36 %
Personas No Alfabetizadas	$ 3,216.85	$ 4,491.43	40 %
Egresados de programa profesionalizante	$ 10,701.15	$ 13,396.53	25 %
Mujeres Egresadas de programa profesionalizante	$ 9,180.91	$ 11,787.55	28 %
Hombres Egresados de programa profesionalizante	$ 12,270.15	$ 14,894.40	21 %

Cuadro 4.1: Media del ingreso para distintos grupos poblacionales en el último trimestre del 2018

Estadístico para la media del ingreso	2024 S/Imputación	2024 con Imputación	Incremento
General	$ 7,083.24	$ 11,922.18	68 %
Hombres	$ 7,849.13	$ 12,795.44	63 %
Mujeres	$ 5,988.18	$ 10,640.77	78 %
Personas Alfabetizadas	$ 7,180.08	$ 12,062.50	68 %
Personas No Alfabetizadas	$ 3,815.08	$ 6,099.40	60 %
Egresados de programa profesionalizante	$ 11,391.67	$ 17,988.65	58 %
Mujeres Egresadas de programa profesionalizante	$ 9,642.37	$ 16,068.64	67 %
Hombres Egresados de programa profesionalizante	$ 13,199.09	$ 19,845.70	50 %

Cuadro 4.2: Media del ingreso para distintos grupos poblacionales en el segundo trimestre del 2024

Estadístico para la media del ingreso	2018 con Imputación	2024 con Imputación	Incremento
General	$ 8,429.92	$ 11,922.18	41 %
Hombres	$ 8,975.36	$ 12,795.44	43 %
Mujeres	$ 7,548.86	$ 10,640.77	41 %
Personas Alfabetizadas	$ 8,532.29	$ 12,062.50	41 %
Personas No Alfabetizadas	$ 4,491.43	$ 6,099.40	36 %
Egresados de programa profesionalizante	$ 13,396.53	$ 17,988.65	34 %
Mujeres Egresadas de programa profesionalizante	$ 11,787.55	$ 16,068.64	36 %
Hombres Egresados de programa profesionalizante	$ 14,894.40	$ 19,845.70	33 %

Cuadro 4.3: Comparativa de la media del ingreso para el último trimestre del 2018 vs el segundo trimestre del 2024

4.2.1 *Estadísticos Generales para describir el ingreso*

Ahora, de manera adicional a la media, se presentan estadísticos descriptivos adicionales que permiten caracterizar mejor la distribución del ingreso.

Como primer análisis, se identifica que la mediana resulta sistemáticamente inferior a la media, lo que revela una distribución sesgada a la derecha: existe una gran concentración de personas con ingresos bajos o medios y un grupo reducido con ingresos muy altos que empujan la media hacia arriba.

La curtosis reportada en este capítulo corresponde al exceso de curtosis de Fisher (donde un valor de cero indica una distribución normal). Los valores observados son particularmente elevados en los datos sin imputación (119.77 en 2018 y 80.71 en 2024), lo que denota colas extremadamente pesadas y una alta influencia de valores atípicos. Con la imputación, la curtosis se reduce de manera sustancial (25.58 en 2018 y 5.15 en 2024), lo cual constituye evidencia de que los ingresos faltantes no eran aleatorios. Además, la caída de la curtosis entre 2018 y 2024 sugiere que la distribución se ha tornado menos extrema, posiblemente vinculada a la política de aumento del salario mínimo.

En cuanto a la dispersión, tanto la desviación estándar como el coeficiente de variación aumentan cuando se consideran los datos imputados. El coeficiente de variación pasa de 46.68 % a 53.15 % en 2018 y de 48.16 % a 68.82 % en 2024, indicando que los ingresos imputados amplían la heterogeneidad de la muestra. Asimismo, el incremento de la dispersión entre 2018 y 2024 muestra que los ingresos se han vuelto más desiguales en torno a la media. Estas observaciones serán relevantes para el análisis del siguiente capítulo.

Estadístico general del ingreso	2018 sin Imputación	2018 con Imputación	2024 sin Imputación	2024 con Imputación
Media	$ 6,198.98	$ 8,429.92	$ 7,083.24	$ 11,922.18
Desviación estándar	$ 5,728.25	$ 7,916.63	$ 6,356.78	$ 13,658.66
Mediana	$ 5,160.00	$ 6,000.00	$ 5,825.56	$ 6,699.47
Curtosis	119.77	25.58	80.71	5.15
Coeficiente de variación	46.68 %	53.15 %	48.16 %	68.82 %

Cuadro 4.4: Comparativa de estadísticos descriptivos para el último trimestre del 2018 vs el segundo trimestre del 2024

5 Metodología de la Regresión Distributiva

Contenidos

La revisión de la literatura ha permitido establecer un panorama amplio sobre las distintas metodologías utilizadas para el análisis de la desigualdad del ingreso, destacando tanto los índices sintéticos como el Índice de Gini y el Índice de Theil, como enfoques más estructurales; y contrafactuales, como el propuesto por DiNardo, Fortin y Lemieux[1]. Entre estos enfoques, la regresión distributiva ha emergido como una herramienta particularmente útil para examinar cómo diversas modificaciones afectan de manera diferencial a distintos sectores de la población, permitiendo trascender el análisis centrado en promedios y enfocarse en la distribución completa del ingreso.

En el presente capítulo se retoma el enfoque metodológico desarrollado por Redmond, Doorley y McGuinness[2] en 2020, adaptándolo al caso mexicano. En particular, se busca evaluar los cambios en la distribución salarial entre el cuarto trimestre del año 2018 y el segundo trimestre del 2024, en un contexto marcado por incrementos significativos en el salario mínimo. La pregunta que guía este análisis es: ¿en qué medida los cambios observados en la distribución del ingreso se deben a alteraciones en la estructura salarial (llamado efecto precio), y en qué medida obedecen a modificaciones en la composición de la población que percibe dichos ingresos (llamado efecto composición)?

El efecto precio representa el efecto puro del cambio en el salario mínimo, es decir, lo que habría ocurrido si las características de los

[1] John DiNardo, Nicole M. Fortin, and Thomas Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Technical Report w5093, National Bureau of Economic Research, 1995. URL https://doi.org/10.3386/w5093

[2] Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland

trabajadores se hubieran mantenido iguales, pero los salarios hubieran cambiado debido al aumento del salario mínimo. Mientras que el efecto composición, captura los cambios que se deben a diferencias en las características de la población entre los dos períodos.

Para responder esta pregunta, se construyen distribuciones salariales contrafactuales mediante modelos de regresión probit secuenciales. Este enfoque permite estimar las distribuciones marginales de ingreso a partir de las características observables de los trabajadores, manteniendo constante la estructura salarial del año base (2018). Posteriormente, se realiza una descomposición tipo Oaxaca-Blinder para identificar el peso relativo de los efectos precio y composición.

$$F_{W_{18}|X_{18}}(w \mid x)$$

Figura 5.1: Se define función para el ingreso (W) en 2018, dadas las características de la población (X)en 2018

$$F_{W_{24}|X_{24}}(w|x)$$

Figura 5.2: Se define función para el ingreso (W) en 2024, dadas las características de la población (X)en 2024

$$F_{W_{(18|24)}}(w) = \int_{\chi_{24}} F_{W_{18}|X_{18}}(w \mid x)\, dF_{X_{24}}(x)$$

Figura 5.3: Se construye una Contrafactual para encontrar el ingreso (w) dadas las características la población (x) en 2025 y la estructura del ingreso (w) en 2018 (sin aumentos salariales).

$$F_{W_{(24|24)}} - F_{W_{(18|18)}} = \left[F_{W_{(24|24)}} - F_{W_{(18|24)}}\right] + \left[F_{W_{(18|24)}} - F_{W_{(18|18)}}\right]$$

Figura 5.4: Se realiza una descomposición para encontrar el **Efecto Precio** y el **Efecto Composición**.

5.1 Aplicación de la Regresión Distributiva

5.1.1 Separación por Zonas de Salario Mínimo

La fuente de información utilizada en este análisis proviene de la Encuesta Nacional de Ocupación y Empleo (ENOE) del Instituto Nacional de Estadística y Geografía (INEGI). Los datos utilizados han sido preprocesados en el capítulo anterior, mediante técnicas de imputación para completar los ingresos faltantes, lo cual permite trabajar con distribuciones más robustas y representativas.

A partir de 2019, comenzaron a coexistir dos salarios mínimos en el país: uno correspondiente a la Zona Libre de la Frontera Norte (ZLFN), y otro aplicable al resto del territorio nacional. Esta diferenciación implica que los efectos del aumento del salario mínimo pueden haberse manifestado de forma distinta en cada zona, lo que complica el análisis si ambas se consideran en conjunto.

Por esta razón, se consideró metodológicamente más adecuado realizar un análisis separado para cada zona geográfica. Este enfoque, no obstante, presentó un reto importante: en el cuarto trimestre de 2018, la ENOE no identificaba explícitamente la ZLFN, mientras que en el segundo trimestre de 2024 ya se incorporaba esta distinción (Zona 1 = ZLFN; Zona 2 = resto del país). Para sortear esta dificultad, se procedió a identificar los registros correspondientes a la actual ZLFN en la base de 2018, utilizando los campos de entidad federativa y municipio. Con esta clasificación retroactiva, fue posible construir dos subconjuntos comparables entre 2018 y 2024: uno para la ZLFN y otro para el resto del país.

Asimismo, para fines del análisis distributivo, se calculó el ingreso por hora trabajada para cada registro. Esta variable se obtuvo dividiendo el ingreso mensual reportado en la columna `ingocup_real` (ingreso mensual proveniente de la ocupación principal, ajustado a precios constantes de 2018) entre las horas semanales trabajadas (`hrsocup`) multiplicadas por cuatro. Esta transformación permite un análisis más preciso al estandarizar los ingresos según el tiempo efectivamente laborado, facilitando así la estimación de los modelos de regresión probit secuenciales.

5.1.2 *Selección de variables*

La selección de variables explicativas en este análisis retoma el enfoque de Redmond, Doorley y McGuinness[3], quienes estudian el impacto del aumento del salario mínimo en la distribución del ingreso en Irlanda. En su modelo, los autores incluyen variables que describen características del jefe o jefa del hogar, tales como edad, edad al cuadrado, escolaridad, estado civil, nacionalidad irlandesa y situación en el mercado laboral. Estas variables permiten descomponer los cambios en la distribución del ingreso en dos componentes fundamentales: el efecto precio (cambios en la estructura de salarios) y el efecto composición (cambios en las características de la población). Este marco analítico se adopta y adapta en el presente trabajo para analizar la evolución de los ingresos laborales en México entre 2018 y 2024.

A diferencia del estudio irlandés, que se enfoca en los hogares, en esta investigación se analiza el ingreso individual, con base en la Población Económicamente Activa (PEA). Por lo tanto, no se incluye el estado civil, ya que su relevancia en el ingreso individual es más limitada que en el ingreso total por hogar. Asimismo, variables como nacionalidad o estatus migratorio fueron descartadas, tanto por la baja frecuencia de personas en estas categorías como por la falta de información comparable entre años. Por ejemplo, aunque en la ENOE de 2024 se incluye información sobre migración interestatal e internacional, en 2018 estas preguntas no estaban presentes, y en 2024 sólo el 0.5 % de los individuos reportan haber vivido en otro estado y el 0.1 % en otro país durante el año anterior, por lo que no se considera una dimensión de peso suficiente para el análisis.

Por otro lado, se decidió incluir el sexo como variable explicativa. Aunque no está presente en el modelo de Irlanda, en México la literatura ha documentado de manera consistente una brecha salarial por género. Por ejemplo, Campos-Vázquez[4] señala que "las mujeres que trabajan se enfrentan a menores ingresos en promedio que los hombres (13–15 % en 2017)", lo que justifica su inclusión como una dimensión clave de composición demográfica.

En resumen, las variables seleccionadas buscan capturar características observables de los trabajadores que reflejan la composición de la fuerza laboral, y no necesariamente aspectos del mercado de trabajo que están directamente correlacionados con el salario (como la industria o la ocupación). Esto es especialmente importante, ya que el objetivo metodológico central es estimar los efectos composición y precio de manera separada. En ejercicios previos de selección de variables, como los realizados para la imputación del ingreso, se consideraron variables fuertemente relacionadas con los

[3] Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland

[4] Raymundo Campos. Movilidad social, empleo e ingresos laborales en méxico. Technical report, Centro de Estudios Económicos, El Colegio de México, 2021. URL https://movilidadesocial.colmex.mx/wp-content/uploads/2021/10/5.-Raymundo-Campos.pdf. ENOE & IMSS data used

salarios (como el tipo de ocupación, el sector económico o la zona geográfica); sin embargo, para el análisis de regresión distributiva dichas variables podrían contaminar la estimación del efecto precio vs efecto composición.

Con base en estos criterios, las variables explicativas utilizadas son:

- **Edad ('eda')**: Variable continua que indica la edad del informante en años cumplidos. Es un proxy de la experiencia laboral.
- **Edad al cuadrado ('edad_c')**: Captura la no linealidad en la relación entre experiencia e ingreso, siguiendo la hipótesis de una relación cóncava.
- **Sexo ('sex')**: Variable binaria que identifica el sexo del informante (1 = Hombre, 0 = Mujer), incluida por su relevancia en la estructura salarial mexicana.
- **Años de escolaridad ('anios_esc')**: Número total de años de educación formal completados por el informante. Es una de las variables más relevantes para explicar diferencias en ingresos, al estar asociada a la productividad y al tipo de empleo al que se accede.

Estas variables fueron seleccionadas tanto por su relevancia teórica como por su disponibilidad en ambas bases de datos (2018 y 2024), lo cual permite la comparación entre períodos sin introducir sesgos por diferencias en el diseño de la encuesta. Adicionalmente, en ejercicios previos de selección de variables realizados durante la fase de imputación (modelos Gradient Boosting y Random Forest) se identificó que variables como edad, nivel de escolaridad, estrato sociodemográfico, carrera estudiada y condición de ocupación son algunas de las más relevantes para predecir el ingreso. Esto respalda la pertinencia de las variables finalmente seleccionadas para el análisis de regresión distributiva, en la medida en que capturan componentes clave de la composición demográfica de la fuerza laboral sin sobreajustar el modelo a características propias del mercado laboral que podrían interferir con la identificación del efecto precio.

5.1.3 *Cálculo de las Funciones de Distribución Acumulada*

Posteriormente, se calcularon los cuantiles del 1 % al 100 % de la variable `ingocup_hr` tanto para el año 2018 como para 2024. A partir de esta información, se estableció el percentil 99 como umbral superior de ingreso, con el objetivo de evitar la influencia de valores extremos o atípicos. Con un incremento de 1 unidad (Pesos Mexicanos), se generó una secuencia de puntos que va desde 1 hasta dicho percentil. Cada

uno de estos puntos representa un umbral de ingreso por hora a partir del cual se estimará la probabilidad acumulada de que una persona perciba un ingreso igual o menor, dado su conjunto de características observadas. Este procedimiento puede considerarse un modelo probit secuencial, ya que se estiman múltiples modelos binarios (uno por cada punto de corte), y al graficar los resultados obtenidos se reconstruye la forma de la función de distribución acumulada del ingreso.

Para efectos de este análisis, se construirán tres funciones de distribución acumulada o CDF's (por sus siglas en inglés Cumulative Distribution Function), tanto para la Zona 1 como para la Zona 2:

- CDF para 2018 (observada): refleja la distribución del ingreso por hora, con base en las características poblacionales y salariales correspondientes del año base.
- CDF para 2024 (observada): posterior a los aumentos en el salario mínimo, refleja la distribución del ingreso por hora, considerando las características de la población y los precios salariales del 2024.
- CDF contrafactual (construida): nos muestra el panorama hipotético que existiría si se hubiesen conservado los salarios del 2018 pero la estructura poblacional del año 2024, es decir, lo que hubiera ocurrido si el salario mínimo no hubiera sufrido modificaciones.

5.2 *Descomposición de la distribución*

Retomando el enfoque metodológico propuesto por Redmond, Doorley y McGuinness[5], se procedió a descomponer las funciones de distribución acumulada (CDF) con el objetivo de identificar y cuantificar los factores que explican el cambio en la distribución del ingreso entre 2018 y 2024. En particular, se busca distinguir entre el efecto atribuible al incremento salarial y el efecto asociado a los cambios en la composición de la población ocupada, mediante una descomposición tipo Oaxaca-Blinder adaptada a distribuciones.

La descomposición se lleva a cabo en tres pasos:

1. **Cambio total**: Se calcula como la diferencia punto a punto entre la CDF observada en 2024 y la CDF observada en 2018. Esta resta captura el cambio total en la distribución del ingreso por hora entre ambos años.
2. **Efecto del salario**: Se obtiene restando la CDF contrafactual de 2024 a la CDF observada en 2024. Esta diferencia refleja el impacto del aumento en los ingresos (por ejemplo, por modificaciones en el salario mínimo), manteniendo constante la estructura poblacional de 2024.

[5] Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland

3. **Efecto de la estructura poblacional**: Se estima al restar la CDF de 2018 a la CDF contrafactual. Esta diferencia captura el cambio atribuible a la transformación en la composición de la población ocupada, bajo el supuesto de que los niveles salariales se hubieran mantenido como en 2018.

5.3 *Percentiles y Razones*

Finalmente, con el objetivo de complementar el análisis gráfico de las distribuciones, se realiza una comparación entre los percentiles clave del ingreso por hora en la distribución observada de 2024 y en la distribución contrafactual. En particular, se calcularon los percentiles P10, P25, P75 y P90, así como los ratios P90/P10 y P75/P25, los cuales son métricas comúnmente utilizadas en la literatura económica para evaluar la desigualdad en la distribución del ingreso.

Los percentiles dividen la distribución de los ingresos en cien partes iguales. Por ejemplo:

- El percentil 10 (P10) indica el ingreso por hora por debajo del cual se encuentra el 10 % de la población. Representa a los trabajadores con ingresos más bajos.
- El percentil 25 (P25) corresponde al primer cuartil, es decir, el ingreso que marca al 25 % de la población con menores ingresos.
- El percentil 75 (P75) señala el punto por debajo del cual se encuentra el 75 % de los trabajadores; representa un ingreso relativamente alto dentro de la parte media-alta de la distribución.
- Finalmente, el percentil 90 (P90) muestra el ingreso que delimita al 90 % de la población, permitiendo observar el comportamiento en la parte más alta de la distribución.

Estos puntos resumen el comportamiento de la distribución en distintos tramos de ingresos, permitiendo identificar quiénes ganan menos, quiénes se sitúan en la media-alta, y quiénes están en la cima de la distribución salarial.

Para evaluar de manera más explícita la desigualdad, se construyen ratios entre percentiles, que permiten medir la dispersión entre distintos tramos de la distribución:

- El ratio P90/P10 compara los ingresos del percentil 90 contra los del percentil 10. Este ratio da una medida de la desigualdad entre los extremos de la distribución. Un valor alto indica una gran brecha entre los que más y los que menos ganan.

- El ratio P75/P25 compara los ingresos del percentil 75 contra los del percentil 25. Este ratio se enfoca en la desigualdad en la parte central de la distribución, es decir, entre los trabajadores de ingresos medios-bajos y medios-altos.

Estos indicadores son ampliamente utilizados porque son intuitivos, fáciles de interpretar y permiten comparar desigualdad entre diferentes momentos del tiempo, grupos poblacionales o escenarios contrafactuales, como en este caso.

6 Discusión de Resultados

Contenidos

6.1 Análisis de las CDF 2018, 2024 y Contrafactual

La Zona 1, correspondiente a la franja fronteriza del norte del país, experimentó un aumento notable en el salario mínimo entre 2018 y 2024, pasando de $88.36 a $374.44 pesos diarios. Este cambio sustancial se refleja con claridad en las funciones de distribución acumulada (CDF) de los ingresos por hora reales, las cuales fueron calculadas con base en datos ajustados a precios constantes de 2018.

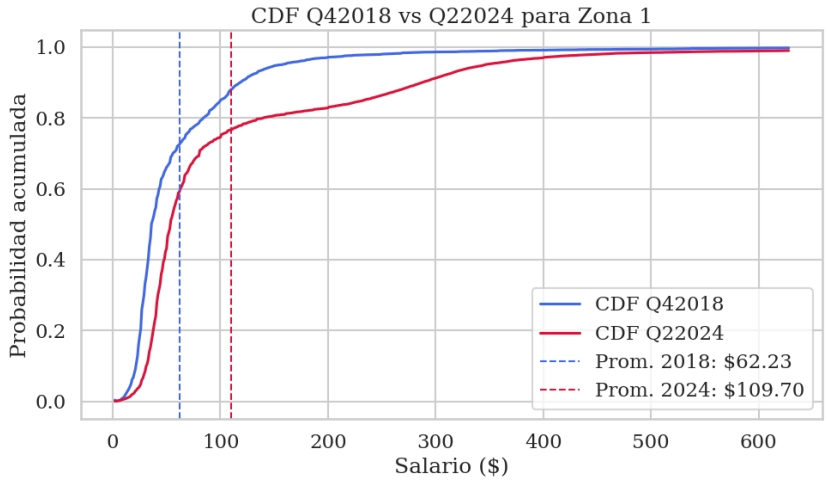

Figura 6.1: Funciones de Distribución Acumulada Para el Q4 del 2018 y Q2 del 2024 correspondientes a la Zona 1

La comparación entre las curvas de 2018 y 2024 evidencia un

desplazamiento generalizado hacia la derecha en 2024, lo cual indica que una proporción menor de trabajadores percibe salarios bajos en comparación con 2018. En particular, este desplazamiento es más pronunciado en la base de la distribución, lo que sugiere que los principales beneficiarios del aumento salarial fueron los trabajadores con menores ingresos.

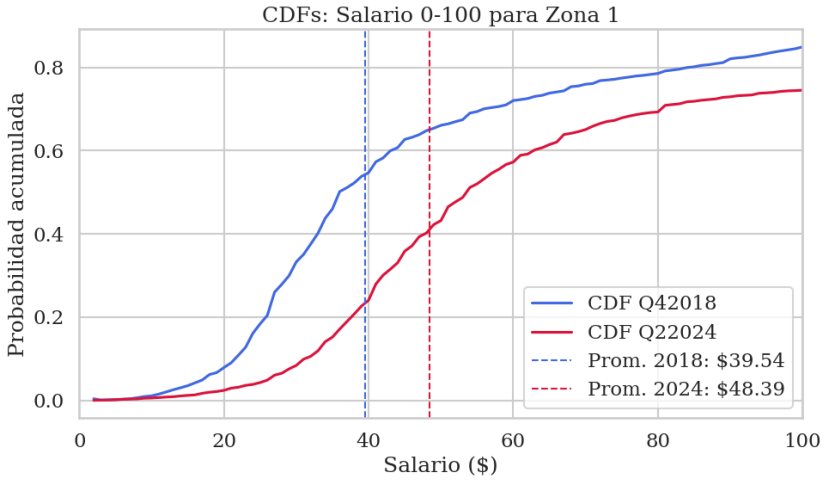

Figura 6.2: Funciones de Distribución Acumulada correspondientes para el rango salarial de $0-$100 para la Zona 1

El salario promedio por hora también refleja esta mejora, al pasar de $62.83 en 2018 a $109.70 en 2024, lo que representa un aumento real cercano al 75 %. Esta mejora se traduce en un movimiento general de toda la curva hacia niveles salariales más altos. Además, siguiendo el análisis de la CDF para 2024 (línea roja), se observa que la mayor distancia respecto a la distribución de 2018 se concentra entre los $100 y 300$ pesos por hora, lo cual podría estar reflejando los efectos acumulados de la política de incrementos salariales en los tramos medios de la distribución.

Para comprender mejor el efecto atribuible exclusivamente al aumento del salario mínimo, se incorpora la curva contrafactual. Esta representa la distribución salarial que habría prevalecido en 2024 si los salarios se hubieran generado con la estructura salarial de 2018, aplicada a las características sociodemográficas de los trabajadores de 2024. Al comparar esta curva con la observada en 2024, se observa un desplazamiento adicional hacia la derecha, lo que confirma que los salarios en 2024 superaron los niveles que se habrían registrado en ausencia del cambio en la política salarial.

La mayor diferencia entre la CDF observada y la contrafactual también se encuentra en los tramos bajos y medios de la distribución, particularmente alrededor de los $200 por hora, lo que refuerza la idea de que la política tuvo un efecto redistributivo positivo y focalizado

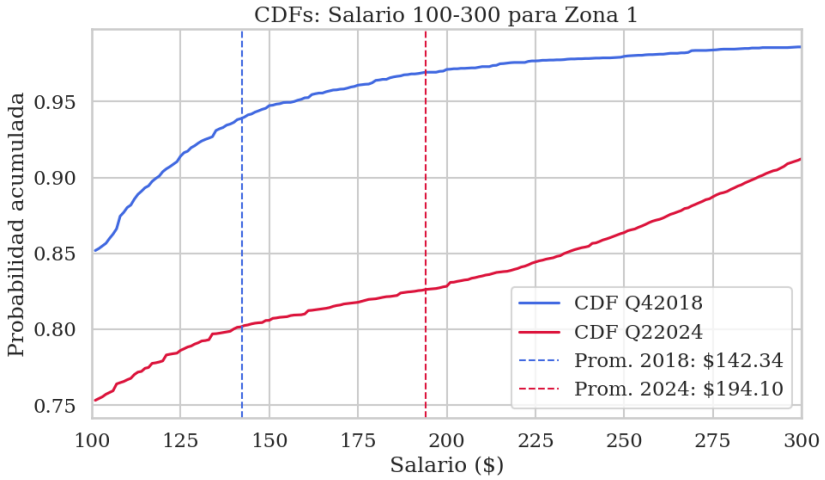

Figura 6.3: Funciones de Distribución Acumulada correspondientes para el rango salarial de $100-$300 para la Zona 1

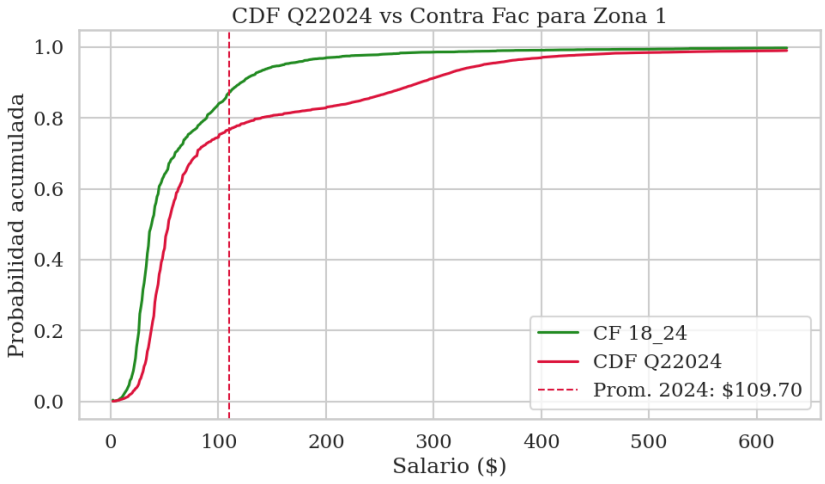

Figura 6.4: Función de Distribución Acumulada para 2024 y Contra Factual generada para Zona 1

en la parte baja de la escala salarial. En contraste, la parte alta de la distribución se mantuvo prácticamente sin cambios, lo que sugiere que la política no alteró significativamente los ingresos de los trabajadores mejor remunerados.

La Zona 2, que abarca al resto del país fuera de la franja fronteriza, también experimentó un aumento sostenido del salario mínimo durante el periodo 2018–2024, aunque no tan pronunciado como en la Zona 1. En términos nominales, el salario mínimo pasó de $88.36 a $241.14 pesos diarios. Ajustando esta última cifra a precios constantes de 2018, el equivalente es de aproximadamente **$165.59 pesos diarios**, lo cual sigue representando un aumento real considerable.

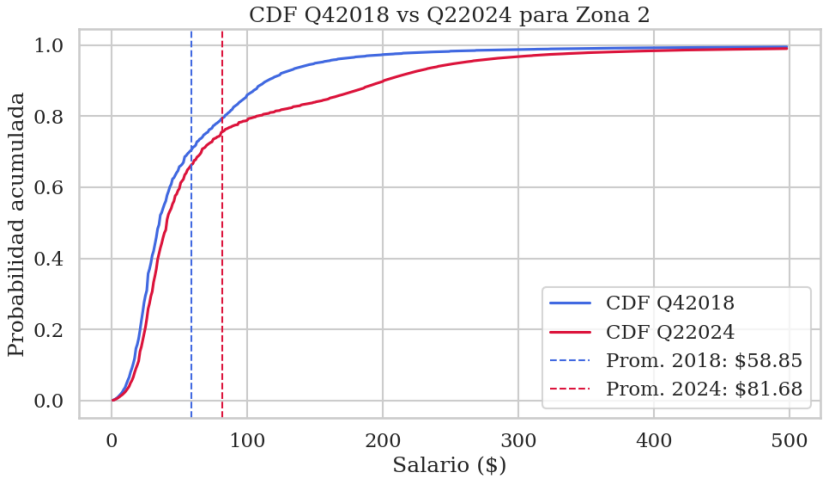

Figura 6.5: Funciones de Distribución Acumulada Para el Q4 del 2018 y Q2 del 2024 correspondientes a la Zona 2

Al observar las funciones de distribución acumulada (CDF) del ingreso por hora real en esta zona, se aprecia un patrón similar al de la Zona 1: la curva correspondiente al año 2024 se encuentra desplazada hacia la derecha respecto a la de 2018. Este desplazamiento indica que, en general, una menor proporción de trabajadores se ubica en los tramos salariales más bajos en 2024, lo cual sugiere una mejora en las condiciones salariales. Sin embargo, a diferencia de la Zona 1, este cambio es más moderado y menos abrupto, lo que es consistente con el menor incremento en el salario mínimo en esta región.

El desplazamiento de la curva es especialmente visible en la base de la distribución, particularmente en el rango de ingresos por hora entre $100 y $300, donde se observa la mayor distancia entre ambas curvas. Esto sugiere que los trabajadores que se encuentran más cerca del salario mínimo —aunque no necesariamente en el punto más bajo de la distribución— fueron quienes recibieron el mayor beneficio de los aumentos salariales. Si bien también existe un desplazamiento positivo para los ingresos más bajos (entre $0 y $100 por hora), este es menos

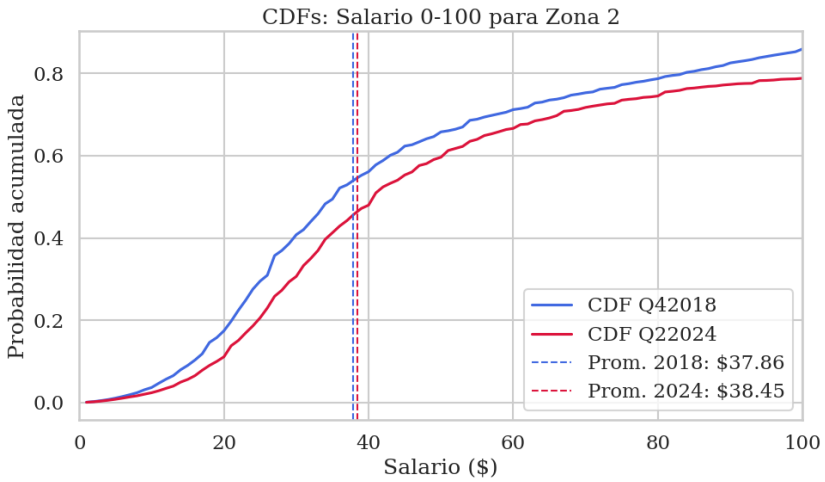

Figura 6.6: Funciones de Distribución Acumulada correspondientes para el rango salarial de $0-$100 para la Zona 2

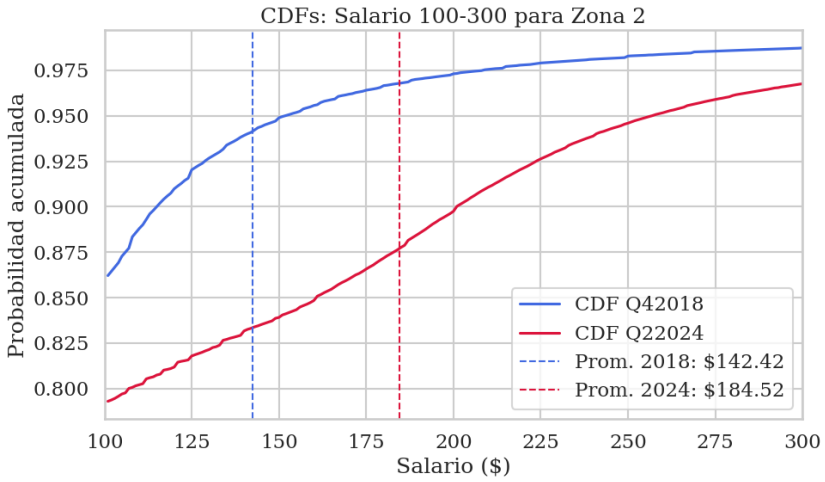

Figura 6.7: Funciones de Distribución Acumulada correspondientes para el rango salarial de $100-$300 para la Zona 2

pronunciado, lo que indica que el impacto de la política se intensificó a partir del umbral del salario mínimo.

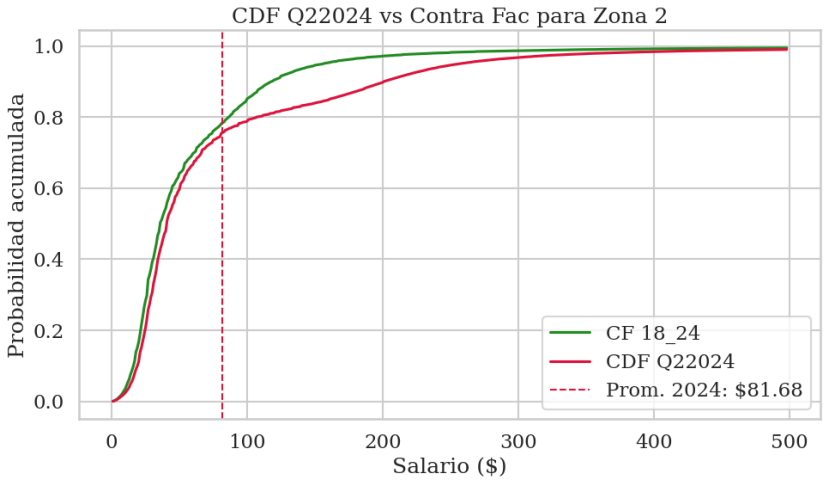

Figura 6.8: Función de Distribución Acumulada para 2024 y Contra Factual generada para Zona 2

Esta diferencia en la magnitud del cambio también se refleja al comparar la curva observada de 2024 con la curva contrafactual, construida aplicando la estructura salarial de 2018 a las características de la población ocupada en 2024. La curva contrafactual permite estimar cómo se habrían distribuido los ingresos si no se hubiera implementado la política de incrementos al salario mínimo. Al igual que en la Zona 1, la CDF observada de 2024 se encuentra consistentemente a la derecha de la contrafactual, lo que confirma que la política tuvo un efecto real en elevar los ingresos, especialmente entre los trabajadores de menores y medianos ingresos.

El hecho de que la distancia entre la observada y la contrafactual sea más marcada en el rango medio-bajo de la distribución (aproximadamente entre $100 y $300 por hora) sugiere que los efectos positivos de la política no se limitaron a un grupo extremadamente reducido, sino que alcanzaron a un sector amplio de la población trabajadora. No obstante, es importante resaltar que, en comparación con la Zona 1, el efecto fue menos intenso y más gradual, lo que concuerda con la diferencia en la magnitud de los aumentos salariales implementados en ambas zonas.

En resumen, aunque el desplazamiento de la distribución salarial en la Zona 2 no fue tan drástico como en la Zona 1, los resultados muestran una mejora real y significativa en los ingresos laborales, especialmente en la parte baja de la distribución. La comparación con la contrafactual refuerza la conclusión de que esta mejora es atribuible, al menos en parte, a la política de incrementos al salario mínimo implementada en el país durante este período.

6.1.1 *Comparativa del ingreso por hora promedio por deciles*

Para complementar el análisis, se examina la distribución del ingreso por hora en deciles, lo cual aporta a la evaluación de en qué segmentos de la población ocupada se concentraron los efectos del aumento al salario mínimo y cómo cambió dicha estructura entre la distribución observada y la contrafactual.

Los resultados para la Zona 1 muestran que los deciles más bajos (P10 a P40) presentan incrementos notables en el ingreso promedio observado para 2024 en comparación con la distribución contrafactual, con aumentos porcentuales superiores al 40 %. Esto sugiere que las mejoras salariales beneficiaron de manera significativa a los trabajadores con ingresos más bajos. El efecto se mantiene positivo, aunque de forma más moderada, en los deciles intermedios (P50 a P70), donde el crecimiento porcentual se reduce progresivamente hasta ubicarse por debajo del 35 %. En contraste, los deciles superiores (P80 y P90) muestran incrementos nuevamente más pronunciados, alcanzando aumentos del 57 % y hasta del 137 %.

Este patrón no lineal en el incremento, donde tanto la base como la cúspide de la distribución reciben mejoras importantes, mientras que los deciles medios presentan aumentos más contenidos, refuerza la posibilidad de que los efectos del alza al salario mínimo fueron heterogéneos. Si bien se logró una mejora sustancial para los trabajadores de menores ingresos, los beneficios más desproporcionados se concentran en el extremo superior.

En conjunto, el análisis de deciles para la Zona 1 confirma un impacto mixto: avances importantes en la base de la pirámide salarial y una creciente concentración del ingreso en los niveles más altos.

Métrica	Distribución Observada	Distribución Contrafactual	Diferencia (%)
P10	31.15439	21.7829	43.00 %
P20	37.59655	26.09148	44.10 %
P30	41.96293	29.47909	42.30 %
P40	47.7675	33.5125	42.50 %
P50	53.52543	37.93793	41.10 %
P60	62.80089	44.68643	40.50 %
P70	80.45081	59.76267	34.60 %
P80	139.9007	89.07432	57.10 %
P90	287.545	121.3186	137.00 %

Cuadro 6.1: Deciles para Distribuciones Observadas y Contrafactuales con Diferencias Percentuales para la Zona 1

Para la Zona 2, también se observa un incremento generalizado en el ingreso promedio por hora en 2024 frente a la distribución contrafactual, lo cual indica que el alza al salario mínimo tuvo efectos positivos a lo largo de toda la población ocupada.

El análisis por deciles muestra un patrón similar al de la Zona 1, aunque con menor intensidad. Los deciles más bajos (P10 a P40)

presentan aumentos porcentuales que van desde el 18.2 % hasta el 11.7 %, lo que refleja un impacto positivo pero más moderado en los trabajadores con ingresos bajos. Este efecto decreciente es especialmente notorio hasta el decil P70, donde el incremento se reduce al 8.7 %, el valor más bajo de todos los deciles.

A partir del decil P80, el incremento repunta de manera importante: el ingreso promedio crece un 23.7 % en el P80 y un 70.3 % en el P90. Este repunte en los deciles superiores refleja que, al igual que en la Zona 1, el mayor crecimiento salarial se concentra en los niveles más altos de la distribución.

Un punto interesante en esta zona es el comportamiento de los deciles medios: mientras que el P40 muestra un aumento de 11.7 %, el P50 lo supera con un 13.5 %, antes de volver a descender ligeramente en P60 (12.7 %) y alcanzar su mínimo en P70. Este patrón sugiere que el impacto en los ingresos medios no fue uniforme, con algunas oscilaciones en los niveles intermedios de la distribución.

En resumen, aunque la Zona 2 también experimentó un crecimiento salarial a lo largo de toda la distribución, los efectos fueron más atenuados en comparación con la Zona 1 y con una mayor variabilidad en los deciles intermedios.

Métrica	Distribución Observada	Distribución Contrafactual	Diferencia (%)
P10	19.02207	16.09111	18.20 %
P20	24.70547	21.42483	15.30 %
P30	29.51531	26.0689	13.20 %
P40	34.21039	30.62662	11.70 %
P50	40.68141	35.85104	13.50 %
P60	50.23312	44.57345	12.70 %
P70	66.25915	60.95255	8.70 %
P80	106.9671	86.45426	23.70 %
P90	200.9049	117.9387	70.30 %

Cuadro 6.2: Deciles para Distribuciones Observadas y Contrafactuales con Diferencias Percentuales para la Zona 2

6.2 *Cambio total en la desigualdad*

Para complementar el análisis de las distribuciones acumuladas, se realizó una evaluación del cambio en la desigualdad salarial a partir del uso de percentiles clave y ratios de dispersión. En particular, se comparan los percentiles 10 (P10), 25 (P25), 75 (P75) y 90 (P90) del ingreso por hora entre la distribución observada de 2024 y la contrafactual. Esta última representa una simulación de lo que habría ocurrido si no se hubieran aplicado los incrementos al salario mínimo, manteniendo la estructura salarial de 2018 pero aplicada a la población ocupada de 2024.

Además de los percentiles, se calcularon dos indicadores comúnmente utilizados para medir desigualdad:

- **P90/P10** - refleja la dispersión entre los extremos de la distribución.
- **P75/P25** - captura la desigualdad en la parte media de la distribución.

Estos indicadores permiten identificar no solo si hubo mejoras en los ingresos, sino también cómo se distribuyeron dichos beneficios entre los distintos segmentos de la población.

Los resultados para la Zona 1 muestran un incremento notable en todos los percentiles analizados al comparar la distribución observada con la contrafactual. En términos absolutos, los percentiles bajos y medios presentan un crecimiento significativo:

- El P10 pasó de $21.78 a $31.15 por hora, lo que representa un aumento del 43 %.
- El P25 creció 48.2 %, mientras que el P75 aumentó 40.7 %.

Estos datos reflejan una mejora generalizada en los ingresos, particularmente en la base y el centro de la distribución. La diferencia entre P10, P25 y P75 es bastante consistente con variaciones en torno al 40 %, lo que sugiere que el alza salarial benefició de forma bastante uniforme a estos segmentos de la población.

Sin embargo, el P90 muestra un comportamiento muy distinto: este percentil tuvo un crecimiento de 137 %, una cifra considerablemente superior a la de los demás percentiles. Esto indica que si bien la población en general recibió mejoras salariales, los mayores beneficios se concentraron en el extremo superior de la distribución. Es probable que este crecimiento tan acelerado esté influido por valores extremos o una baja densidad de observaciones en esta parte alta de la distribución.

Este patrón también se refleja en los indicadores de desigualdad:

- El P90/P10 pasó de 5.56 a 9.22, lo que representa un incremento del 65.7 %, lo cual sugiere un aumento de la desigualdad en los extremos.
- En contraste, el P75/P25 se redujo ligeramente en un 5.1 %, lo que implica una leve compresión en la parte media de la distribución salarial.

En conjunto, estos resultados apuntan a un efecto mixto del aumento al salario mínimo en la Zona 1:

- Por un lado, los trabajadores con menores y medianos ingresos experimentaron una mejora sustancial en sus niveles salariales, con aumentos consistentes y positivos.
- Por otro lado, el crecimiento acelerado en el P90 revela que los ingresos altos crecieron aún más que el resto, lo cual amplió la brecha entre los extremos de la distribución.

- La ligera reducción en la ratio P75/P25 indica que la parte central de la distribución se volvió más equitativa, aunque dicha mejora fue opacada por la creciente desigualdad en la cúspide salarial.

Métrica	Distribución Observada	Contrafactual	Diferencia (%)
P10	31.1543	21.7829	43 %
P25	40.2396	27.1521	48.2 %
P75	100.6066	71.5182	40.7 %
P90	287.5449	121.3186	137 %
P90/P10	9.2296	5.5694	65.7 %
P75/P25	2.5001	2.6339	-5.1 %

Cuadro 6.3: Percentiles y Ratios para Distribuciones Observadas y Contrafactuales con Diferencias Percentuales para la Zona 1

En el caso de la Zona 2, también se observa un incremento en los ingresos para todos los percentiles analizados, aunque de menor magnitud en comparación con la Zona 1. Este comportamiento es consistente con la naturaleza de los aumentos al salario mínimo en esta región, donde, si bien se implementaron alzas progresivas, estas fueron menos pronunciadas.

El crecimiento fue relativamente homogéneo en los percentiles bajos y medios de la distribución:

- El P10 pasó de $16.09 a $19.02, lo que representa un incremento del 18 %.
- En el P75, el ingreso por hora creció de $72.2 estimados en la contrafactual a $80.5 conforme a lo observado en el 2024, representando un aumento del 11.5 %.
- Para el P25, el aumento fue algo más elevado, con un crecimiento del 13.9 %.

Estos datos muestran que, aunque el crecimiento salarial fue moderado, se distribuyó de manera relativamente pareja en la base y parte media de la distribución, lo que sugiere cierta coherencia en los efectos del aumento al salario mínimo para los sectores cercanos a los ingresos bajos y medios.

Sin embargo, el mayor incremento se da, nuevamente, en el extremo superior de la distribución:

El P90 pasó de $117.93 (en la contrafactual) a $200.9 en la distribución observada, lo que representa un aumento del 70.3 %.

Este aumento, aunque menor al observado en la Zona 1 para el mismo percentil, sigue siendo considerablemente mayor que el de los otros percentiles, lo que contribuye a una mayor dispersión en los extremos de la distribución.

Este comportamiento se refleja también en los indicadores de desigualdad:

- El ratio P90/P10 aumentó un 44.1 %, lo cual indica que la brecha entre los extremos de la distribución se amplió, aunque en menor medida que en la Zona 1.
- Por el contrario, el ratio P75/P25 mostró una ligera disminución del 2.1 %, lo que implica una leve reducción en la desigualdad del centro de la distribución.

Estos resultados, si bien menos extremos que los observados en la Zona 1, mantienen una tendencia similar:

- Hubo una mejora generalizada en los ingresos, especialmente en los segmentos bajos y medios.
- El impacto más notorio se dio nuevamente en el percentil más alto, lo cual amplifica la desigualdad en los extremos.
- La parte media de la distribución muestra una leve compresión, lo cual podría interpretarse como un efecto positivo de la política salarial en términos de equidad relativa entre trabajadores de ingresos similares.

En síntesis, los resultados sugieren que la política de incrementos al salario mínimo también tuvo efectos positivos en la Zona 2, particularmente entre quienes perciben ingresos cercanos o moderadamente superiores al salario mínimo. Sin embargo, la persistencia y, en algunos casos, profundización de la desigualdad en los extremos de la distribución sugiere que los beneficios no fueron equitativos a lo largo de toda la población ocupada.

Métrica	Distribución Observada	Contrafactual	Diferencia (%)
P10	19.0220	16.0911	18.2 %
P25	26.7158	23.4586	13.9 %
P75	80.5058	72.2036	11.5 %
P90	200.9048	117.9387	70.3 %
P90/P10	10.5616	7.3294	44.1 %
P75/P25	3.0134	3.0779	-2.1 %

Cuadro 6.4: Percentiles y Ratios para Distribuciones Observadas y Contrafactuales con Diferencias Percentuales para la Zona 2

6.3 *Descomposición del efecto de la estructura y efecto de la política*

Para profundizar en el análisis de los efectos observados en la distribución del ingreso, se recurre a la descomposición del cambio total en la función de distribución acumulada (CDF) del ingreso por hora entre 2018 y 2024. Esta descomposición permite distinguir entre el impacto atribuible a cambios en la estructura demográfica de la

población trabajadora (como edad, escolaridad o sexo) y el efecto específico de la política pública, en este caso, el aumento al salario mínimo.

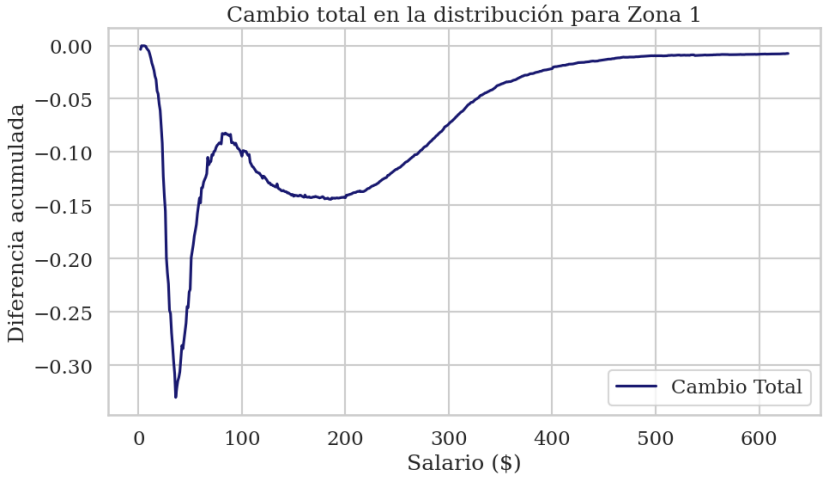

Figura 6.9: Cambio total en la distribucion para Zona 1

La gráfica del cambio total en la Zona 1 muestra una diferencia acumulada negativa pronunciada en los tramos bajos de la distribución salarial, alcanzando un pico que supera al 30 % en el rango de aproximadamente $30 a $60 pesos por hora. Esto indica que, en 2024, una proporción significativa de trabajadores dejó de percibir ingresos tan bajos como en 2018. La magnitud del cambio comienza a disminuir progresivamente conforme se incrementa el ingreso por hora, y tiende a desaparecer alrededor de los $400, lo que sugiere que los efectos más importantes del cambio se concentraron en la base de la distribución. Este patrón general refleja el desplazamiento hacia la derecha de la CDF de 2024, ya discutido en secciones anteriores.

Al descomponer este cambio total, se observa que el componente atribuible a cambios estructurales (esto es, diferencias demográficas en la población trabajadora entre ambos años) tiene un impacto acumulado modesto, alcanzando alrededor de 2 % en los tramos bajos y resultando prácticamente plano en el resto de la distribución. Esto sugiere que el cambio en características como edad, escolaridad y sexo explica solo una fracción muy pequeña del cambio total observado, y que la composición de la fuerza laboral en 2024 era apenas marginalmente más favorable que en 2018.

Por el contrario, la curva del efecto atribuible a la política pública, calculado como la diferencia entre la distribución observada de 2024 y la contrafactual, sigue casi exactamente la misma trayectoria que la curva del cambio total. En particular, reproduce el mismo pico en el rango de ingresos bajos (cercano a los $30 pesos por hora) y mantiene

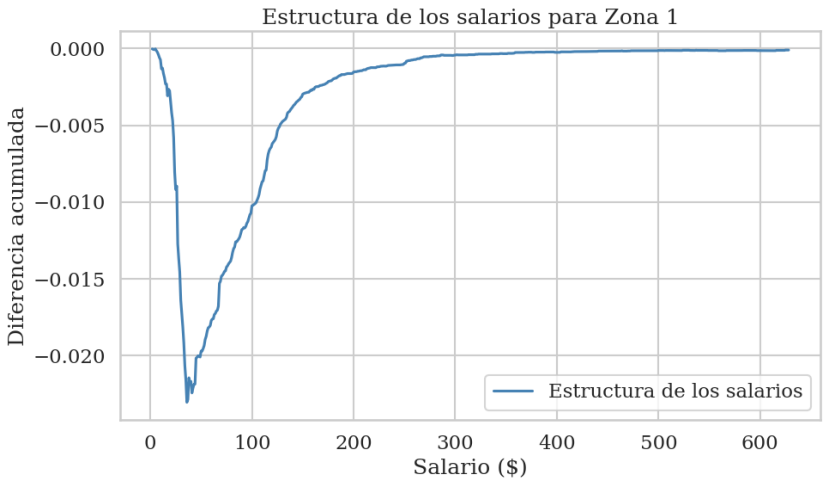

Figura 6.10: Cambio atribuible a la estructura de los salarios para Zona 1

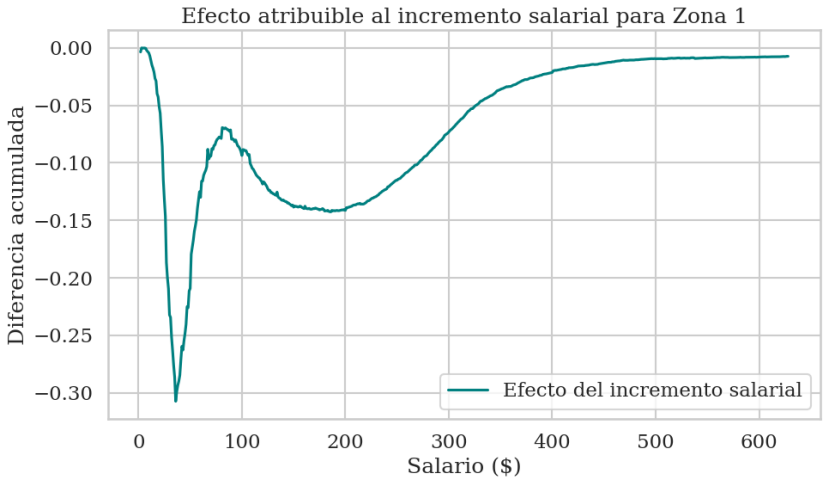

Figura 6.11: Cambio atribuible al incremento salarial para Zona 1

la forma descendente hasta estabilizarse alrededor de los $400. Esto permite concluir que la mayor parte del cambio observado puede ser directamente atribuida al incremento en el salario mínimo, más que a transformaciones estructurales en la fuerza laboral.

En síntesis, la descomposición para la Zona 1 muestra que el aumento del salario mínimo tuvo un impacto sustantivo y focalizado en la parte baja de la distribución salarial, contribuyendo significativamente a reducir la proporción de trabajadores con ingresos laborales más bajos. Mientras tanto, los cambios en la composición de la población trabajadora jugaron un papel claramente marginal en la transformación de la distribución observada.

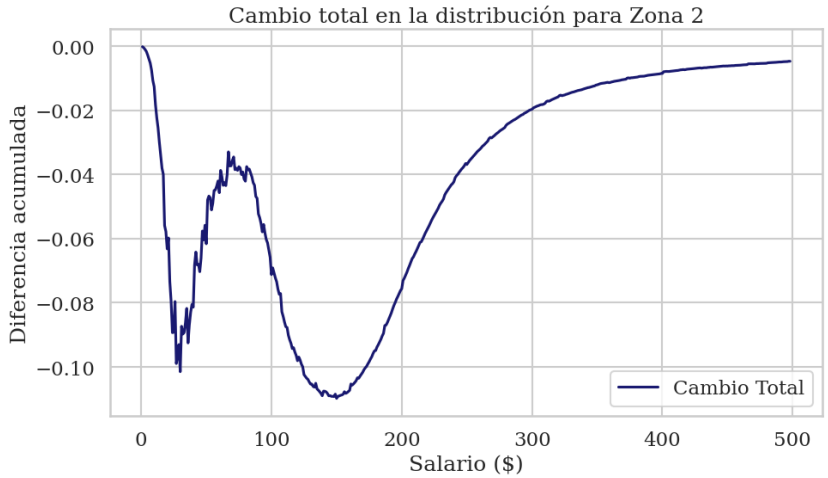

Figura 6.12: Cambio total en la distribución para Zona 2

En la Zona 2 también se observa una diferencia acumulada negativa en la comparación de la distribución de salarios por hora entre 2018 y 2024, aunque de menor magnitud que en la Zona 1. El cambio total presenta un pico negativo en el rango de $30 a $80 pesos por hora, alcanzando una magnitud cercana al 10 %. Sin embargo, a diferencia de la Zona 1, en la Zona 2 el cambio negativo más pronunciado se observa en los rangos bajos a medios, específicamente entre $100 y $400 por hora, donde la diferencia acumulada se aproxima hasta casi el 12 %. Este patrón también refleja un desplazamiento hacia la derecha de la CDF de 2024 para la Zona 2, es decir, una reducción en la proporción de trabajadores con ingresos dentro de esos tramos salariales bajos y medios.

Este comportamiento sugiere que, aunque también hubo beneficios para los trabajadores de menores ingresos, en la Zona 2 el impacto fue mayor en la base media de la distribución, particularmente en quienes percibían entre $100 y $300 por hora. La población con ingresos más bajos sí muestra una mejora, aunque más moderada que en la Zona 1.

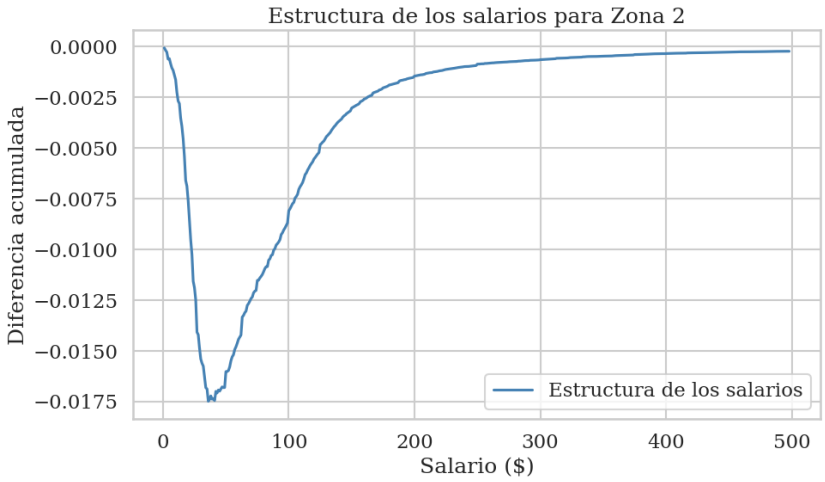

Figura 6.13: Cambio atribuible a la estructura de los salarios Zona 2

Cuando se descompone este cambio total, el efecto atribuible a cambios en la estructura demográfica de la población trabajadora resulta mínimo, con una reducción acumulada de apenas 1.75 % en los tramos bajos de ingreso. Este dato refuerza la conclusión de que los cambios en variables como edad, educación y sexo tienen un papel muy limitado en la transformación de la distribución observada en esta zona.

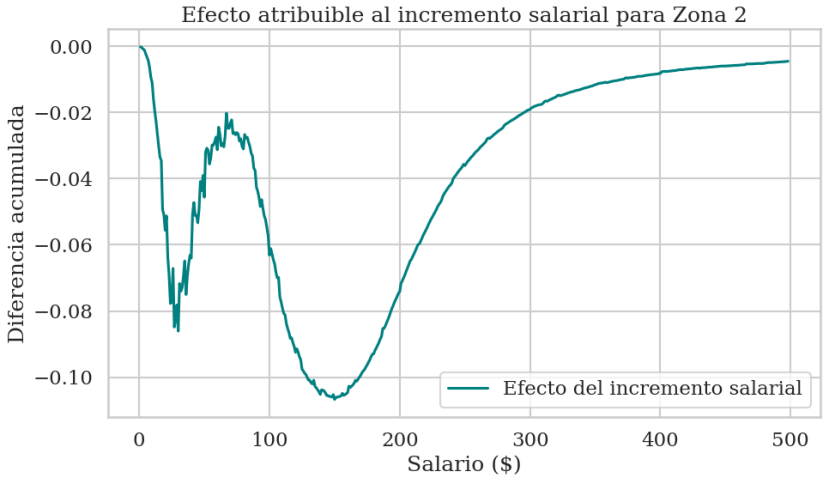

Figura 6.14: Cambio atribuible al incremento salarial Zona 2

En contraste, el efecto atribuible a la política pública de aumento al salario mínimo muestra una mayor coincidencia con la curva del cambio total. Se identifican impactos considerables en los tramos de ingreso entre $100 y $300 por hora, con diferencias acumuladas de más de 10 %, y una contribución importante también en los rangos más

bajos (alrededor de 8 %). Esto indica que, como en la Zona 1, el cambio observado se explica en gran medida por los incrementos salariales asociados a la política de salario mínimo, más que por modificaciones en la composición de la fuerza laboral.

Al comparar los resultados de ambas zonas, se identifica un patrón similar, aunque con algunas diferencias importantes en la magnitud y localización de los efectos. En la Zona 1, el impacto del aumento al salario mínimo es más profundo y focalizado en los tramos más bajos del ingreso (alrededor de $30 a $60 pesos por hora), con un cambio acumulado que llega hasta 30 %, mientras que en la Zona 2 el impacto es más moderado y distribuido entre los tramos bajos y medios de la distribución (principalmente entre $100 y $300 por hora).

En ambas zonas, el efecto de la estructura demográfica es marginal o muy limitado, lo cual refuerza que el cambio en la distribución de ingresos se debe casi exclusivamente a la política pública de incremento al salario mínimo. Sin embargo, la Zona 1 presenta una redistribución más intensa hacia los trabajadores de menores ingresos, mientras que en la Zona 2 el beneficio se extiende más hacia los rangos medios del ingreso

7 Conclusiones y trabajo futuro

Contenidos

7.1 Conclusiones respecto a los resultados del análisis con Regresión Distributiva

Uno de los principales objetivos de esta investigación fue analizar el impacto del aumento del salario mínimo en la distribución del ingreso salarial en México, a partir de las políticas implementadas durante el último sexenio. Para ello, se modelaron las distribuciones salariales de la población ocupada mediante la metodología de Regresión Distributiva, comparando el último trimestre de 2018 y el segundo trimestre de 2024. Además, se buscó descomponer dicho impacto entre los cambios atribuibles a la política salarial y aquellos derivados de modificaciones en la estructura de la población trabajadora.

Este enfoque, inspirado en el trabajo de Redmond, Doorley y McGuinness[1], resulta particularmente valioso al permitir una evaluación completa de la función de distribución acumulada (CDF, por sus siglas en inglés: Cumulative Distribution Function) del ingreso. A diferencia de otras metodologías que se enfocan únicamente en índices o en segmentos específicos de la población (por ejemplo, cuartiles o deciles), la Regresión Distributiva nos permite observar los efectos a lo largo de toda la distribución del ingreso, punto por punto, ofreciendo además un análisis visual valioso.

Gracias a esta metodología, se pudo no solo comparar la distribución salarial de dos períodos distintos, sino también construir una distribución contrafactual para 2024. Esta contrafactual simula cómo habría sido la distribución del ingreso en ese año si no se hubieran implementado los incrementos al salario mínimo y se hubieran

[1] Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland

mantenido los sueldos de 2018, solamente actualizando la información de la población que percibe ingresos. De esta manera, fue posible estimar de forma precisa y robusta el efecto aislado de la política de aumentos salariales.

Los resultados revelan un efecto positivo y generalizado del aumento al salario mínimo en la distribución del ingreso: todos los tramos de la curva presentan mejoras, aunque en magnitudes diferentes. Este hallazgo fue sorprendente, ya que gran parte del debate público anticipaba posibles efectos negativos o limitados al extremo inferior de la distribución. Sin embargo, los resultados muestran que el beneficio se extendió más allá de los primeros percentiles, alcanzando también a sectores medios y altos.

Adicionalmente, la descomposición entre efecto composición y efecto precio permitió corroborar que la mejora observada en los ingresos no se explica únicamente por cambios en la composición de la población ocupada, sino que, en mayor medida, es atribuible a los incrementos del salario mínimo. Si bien hubo una ligera contribución por parte del cambio en la estructura poblacional, es el componente de política salarial el que genera las diferencias más significativas.

Este análisis proporciona evidencia sólida sobre el impacto distributivo de la política salarial reciente en México y destaca el potencial de metodologías como la Regresión Distributiva para analizar de forma más granular y precisa los efectos de políticas públicas.

7.2 *Conclusiones de la imputación*

Uno de los objetivos particulares de esta investigación fue construir una base de datos completa y curada que pudiera ser utilizada como insumo para el análisis principal. Para ello, fue fundamental abordar la ausencia de información en la variable de ingreso, que afecta aproximadamente al 30 % de las personas que reportan haber trabajado al menos una hora a la semana y haber recibido una remuneración por su trabajo. Gracias a los modelos utilizados, este objetivo se ha cumplido: se logró imputar los ingresos faltantes mediante la aplicación de algoritmos de Machine Learning, cuidando tanto la estructura como la coherencia interna de los datos. Es importante precisar que la imputación no se realizó mediante técnicas tradicionales como la sustitución por media o mediana, sino a través de algoritmos de Machine Learning, específicamente entrenando modelos de regresión sobre los registros válidos. Este enfoque representó un reto adicional, ya que fue necesario preparar adecuadamente las variables predictoras sin alterar su integridad ni inducir distorsiones, asegurando así que los modelos capturaran patrones reales presentes en la información original, pero sobre todo que desde la perspectiva económica no hubiera

alteración de los registros.

A lo largo del trabajo, se utilizaron técnicas de Ciencia de Datos para abordar un problema económico relevante. El uso de algoritmos como XGBoost Regressor para la predicción y métodos de selección de variables como Gradient Boosting permitió obtener resultados superiores en términos de desempeño, al compararlos con estudios previos. Las métricas alcanzadas por nuestros modelos no solo fueron mejores, sino que también reflejan una mayor capacidad para capturar relaciones complejas y no lineales entre las variables que explican el ingreso laboral. Esto refuerza la utilidad del enfoque propuesto como una alternativa moderna y precisa al tratamiento de los datos faltantes.

Además, la diferencia observada en estadísticas descriptivas básicas, como el ingreso promedio, entre la base original (sin imputación) y la base imputada fue notable. En el caso del segundo trimestre de 2024, la media del ingreso aumentó hasta en un 68 % después de la imputación, lo que indica una subestimación considerable al utilizar únicamente los registros con datos válidos. Esto confirma que los ingresos faltantes no son aleatorios, y su omisión afecta de forma desigual a la distribución. Por lo tanto, esta etapa de imputación resultaba imprescindible para garantizar un análisis posterior más certero.

Aunque se utilizaron modelos más robustos que los tradicionalmente empleados, como el pareamiento por puntajes de propensión o el método Hot Deck, esto no implica desvalorizar los enfoques previos. Al contrario, esta investigación reconoce su aporte y busca construir sobre ellos, incorporando nuevas herramientas para mejorar su precisión y aplicabilidad. El uso de técnicas de Machine Learning representa un puente entre la Economía aplicada y la Ciencia de Datos, donde la complementariedad entre enfoques clásicos y modernos puede enriquecer el análisis.

7·3 *Trabajo a futuro*

Este trabajo abre múltiples líneas de investigación que pueden fortalecer y ampliar los hallazgos presentados. Una de las extensiones más inmediatas consiste en aplicar la base de datos imputada al análisis de indicadores laborales ampliamente estudiados, como la pobreza laboral. Esto permitirá evaluar el impacto que tiene una imputación más precisa del ingreso sobre estas métricas, considerando que, como se discutió previamente, los ingresos tienden a estar subestimados cuando se eliminan los registros faltantes. Al observar cómo cambian estos indicadores tras imputar los ingresos no reportados, se podrá valorar mejor la relevancia del procedimiento propuesto y ajustar los modelos para futuras aplicaciones; pero también, esto abrirá la puerta a analizar de forma más precisa estos indicadores y el efecto de las

políticas públicas en los mismos.

Asimismo, una siguiente etapa lógica será extender la imputación del ingreso a otros periodos de tiempo, de modo que se pueda construir una serie histórica más completa y confiable. En ese mismo sentido, se propone comparar de forma directa los resultados obtenidos por nuestros modelos de Machine Learning con los que se derivan de metodologías tradicionales de imputación, como el método Hot-Deck. Además de evaluar métricas de ajuste y precisión, también sería valioso correr el análisis de regresión distributiva utilizando los datos imputados con dicho enfoque tradicional, y contrastar los resultados frente a los obtenidos en esta investigación. Esto permitiría valorar no solo cuál metodología consigue mejores resultados, sino también cómo cada enfoque afecta la interpretación de políticas públicas clave como el aumento al salario mínimo.

Finalmente, si bien el desempeño del modelo seleccionado (XGBoost Regressor) fue adecuado, existen oportunidades claras de mejora en el proceso de modelado. En esta primera etapa, los hiperparámetros del modelo fueron definidos automáticamente por la función empleada, pero una optimización más precisa mediante técnicas como Grid Search o Bayesian Optimization podría aumentar su rendimiento y reducir aún más el error de imputación. También resulta prometedor explorar otros algoritmos, como Redes Neuronales o Redes Bayesianas, que tienen la capacidad de capturar interacciones tanto lineales como no lineales. Estas técnicas podrían ofrecer alternativas sólidas, especialmente en escenarios con estructuras de datos aún más complejas.

Estas propuestas delinean un camino claro para fortalecer la aplicación de herramientas de Ciencia de Datos al análisis económico, con el objetivo de mejorar la calidad de las estadísticas sociales y contribuir con evidencia más precisa para la toma de decisiones de política pública.

Bibliografía

Stefano Bilotta, Luciano Alessandro Ipsaro Palesi, and Paolo Nesi. Exploiting open data for co estimation via artificial intelligence and explainable ai. *Expert Systems with Applications*, 291:128598, 2025. DOI: 10.1016/j.eswa.2024.128598.

Raymundo Campos. Movilidad social, empleo e ingresos laborales en méxico. Technical report, Centro de Estudios Económicos, El Colegio de México, 2021. URL https://movilidadesocial.colmex.mx/wp-content/uploads/2021/10/5.-Raymundo-Campos.pdf. ENOE & IMSS data used.

Raymundo M. Campos-Vazquez and Gerardo Esquivel. The effect of the minimum wage on poverty: Evidence from a quasi-experiment in mexico. *The Journal of Development Studies*, 59(3):360–380, 2023. DOI: 10.1080/00220388.2022.2130056.

Raymundo M. Campos-Vázquez. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en méxico. *El Trimestre Económico*, 80(320):803–839, 2013. URL https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0041-30112013000400003.

Tadeo Campoy. México se ubica entre las economías con alta desigualdad, según informe del banco mundial. https://www.elimparcial.com/dinero/2024/10/15/mexico-se-ubica-entre-las-economias-con-alta-desigualdad-segun-informe-del-banco-mundial/, 2024. Publicado el 15 de octubre de 2024, accedido en junio de 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. ACM, 2016. DOI: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Comisión Nacional de los Salarios Mínimos (CONASAMI). Evolución del salario mínimo. https://www.gob.mx/conasami/documentos/evolucion-del-salario-minimo?idiom=es, n.d. Accedido en junio de 2025.

CONEVAL. Imputación de ingresos no reportados en la enoe. https://www.gob.mx/cms/uploads/attachment/file/806698/Imputaci_n_de_ingresos_no_reportados_en_la_ENOE.pdf, 2022. Accedido en julio de 2025.

Merari Cortés Sánchez, Adriana Zambrano-Reyes, and Tomás Gómez-Rodríguez. Evolución de la desigualdad salarial en méxico 2016-2020, un problema para el desarrollo económico. *Boletín Científico de las Ciencias Económico Administrativas del ICEA*, 12(23):6–13, 2023. ISSN 2007-4913. DOI: 10.29057/icea.v12i23.11572. URL https://doi.org/10.29057/icea.v12i23.11572. Recibido el 4 de septiembre de 2023; aceptado el 23 de octubre de 2023; publicado el 5 de diciembre de 2023.

Instituto Nacional de Estadística y Geografía (INEGI). Encuesta nacional de ocupación y empleo 2018, cuestionario ampliado, datos correspondientes al primer trimestre. diccionario de datos: Sdemt118. https://www.inegi.org.mx/rnm/index.php/catalog/448/data-dictionary/F10?file_name=SDEMT118, August 2022. Red Nacional de Metadatos, consultado el 17 de agosto de 2025.

John DiNardo, Nicole M. Fortin, and Thomas Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Technical Report w5093, National Bureau of Economic Research, 1995. URL https://doi.org/10.3386/w5093.

Benito Durán Romo. Comparación de metodologías de imputación aplicadas a ingresos laborales de la enoe. *Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía*, 10:4–27, 2019.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. DOI: 10.1214/aos/1013203451.

Joe Hasell. Measuring inequality: what is the gini coefficient? https://ourworldindata.org/what-is-the-gini-coefficient, 2023. Online resource, accessed June 2025.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

Susana Herrero Olarte. The minimum wage in ecuador. *Revista de Economía y Trabajo*, 2023. Based on ENEMDUM and INEC data, Panel Data Regression methodology.

Instituto Nacional de Estadística y Geografía (INEGI). Encuesta nacional de ocupación y empleo (enoe): población de 15 años y

más. https://www.inegi.org.mx/programas/enoe/15ymas/, 2024. Consultado el 16 de octubre de 2024.

David S. Kaplan and Francisco Pérez Arce Novaro. El efecto de los salarios mínimos en los ingresos laborales de méxico. *El Trimestre Económico*, 73(289):139–173, 2006. URL https://doi.org/10.20430/ete.v73i289.556. Accedido en julio de 2025.

Carlo Lombardo, Lucía Ramírez-Leira, and Leonardo Gasparini. Does the minimum wage affect wage inequality? a study for the six largest latin american economies. *Latin American Economic Review*, 2024. URL https://laer-journal.springeropen.com/articles/10.1186/s40503-024-00103-7. EPH, PNAD, CASEN, GEIH, ENIGH, ENAHO; comparative cross-country analysis.

Stefano Marchetti and Nikos Tzavidis. Robust estimation of the theil index and the gini coeffient for small areas. *Journal of Official Statistics*, 37(4):955–979, 2021. DOI: 10.2478/jos-2021-0041. URL https://doi.org/10.2478/jos-2021-0041.

Index Mundi. Mexico - indice de gini. https://www.indexmundi.com/es/datos/mexico/indicador/SI.POV.GINI, 2025. Accedido en junio de 2025.

Paul Redmond, Karina Doorley, and Seamus McGuinness. The impact of a minimum wage change on the distribution of wages and household income. *Labour Economics*, 65:101845, 2020. DOI: 10.1016/j.labeco.2020.101845. Based on data from Central Statistics Office in Ireland.

Ricardo E. Rodríguez Pérez, Deyanira Castro Lugo, and Martín Mendoza López. Desigualdad salarial y trabajo informal en regiones de méxico. *Región y Sociedad*, 31:1–23, 2019. DOI: 10.22198/rys2019/31/1062. URL https://doi.org/10.22198/rys2019/31/1062.

Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. DOI: 10.1037/1082-989X.7.2.147. URL https://doi.org/10.1037/1082-989X.7.2.147.

Neha Thakur. Knn for classifying income, 2019. URL https://medium.com/@nehathakur912/knn-for-classifying-income-2a2b39d5984a. Medium. Accessed: 2025-08-17.