Sesgo de exposición en modelos predictivos aplicados en el Marketing Digital

Martina Gimenez

Trabajo Final para acreditar el título de Licenciatura en Gestión de Tecnología Informática

2025

Resumen

La presente investigación aborda el análisis del sesgo de exposición en modelos predictivos aplicados al marketing digital mediante un enfoque metodológico empírico-cuantitativo. El estudio se fundamenta en el análisis de un dataset, constituido por 10,000 impresiones publicitarias distribuidas entre 4,000 usuarios únicos, con 11 variables que capturan información demográfica y contextual.

La metodología se estructura en cinco fases: exploración y caracterización del dataset mediante cálculo de métricas de desigualdad, análisis de diversidad de opciones de consumo, validación experimental de técnicas de mitigación de sesgo, análisis de efectividad de campañas publicitarias y, por último, desarrollo de una taxonomía sistemática que caracteriza cinco tipos de sesgo de exposición.

Los resultados obtenidos evidencian la presencia de sesgo de exposición significativo. La validación de técnicas de mitigación demuestra que es posible mejorar métricas de equidad algorítmica sin degradar significativamente el rendimiento predictivo.

Las contribuciones de este trabajo comprenden: evidencia empírica cuantificable del sesgo de exposición en contextos de marketing digital, desarrollo de una taxonomía sistemática para la caracterización de tipos de sesgo, validación empírica comparativa de técnicas de mitigación y análisis de la mejora simultanea de equidad y efectividad de las campañas publicitarias.

Palabras clave: *análisis predictivo, marketing digital, modelos predictivos, personalización de campañas, sesgo de exposición*

Índice general

Índice de figuras

Capítulo 1

Introducción

En la era digital contemporánea, la tecnología de la información se ha convertido en un componente esencial y cada vez más complejo en diversas industrias, siendo la gestión de la experiencia del cliente y la toma de decisiones empresariales áreas críticas de aplicación. Dentro del ámbito del marketing digital, esta integración tecnológica se manifiesta en el auge de los Sistemas de Recomendación (SR), los cuales son reconocidos como herramientas altamente efectivas para mitigar la sobrecarga de información y ofrecer sugerencias personalizadas a cada usuario. Estos sistemas se encuentran en el corazón del e-commerce, las redes sociales y las plataformas de streaming, desempeñando un papel fundamental al impulsar la satisfacción del cliente, las ventas y el compromiso del usuario. Los modelos predictivos basados en Inteligencia Artificial (IA) y Machine Learning (ML) son el impulso de esta personalización, ya que son capaces de mejorar continuamente al ser expuestos a nuevos datos, asegurando así su relevancia y adaptabilidad en entornos digitales dinámicos. Sin embargo, el éxito de estos modelos predictivos se basa fundamentalmente en el análisis de los datos de comportamiento del usuario, los cuales, por naturaleza, son observacionales y no experimentales. Esta dependencia genera diversos problemas inherentes conocidos como sesgos de datos, entre los cuales se destaca el sesgo de exposición (exposure bias).

El sesgo de exposición se define como el fenómeno en el que ciertos elementos están sobrerrepresentados en los resultados de la recomendación, mientras que otros no están representados adecuadamente. Esto ocurre porque la interacción observada con un elemento está inherentemente condicionada por el mecanismo que lo expuso inicialmente al usuario. La incapacidad para distinguir si una interacción no observada se debe a una falta de interés genuino o simplemente a que el usuario desconocía el elemento (es decir, no fue expuesto) resulta en sesgos severos durante el entrena-

miento del modelo.
La naturaleza interactiva de los sistemas de recomendacion agrava significativamente este desafío técnico, ya que se establece un bucle de retroalimentación perjudicial, donde la secuencia es:

$$\text{Modelo} \rightarrow \text{Usuario} \rightarrow \text{Datos} \rightarrow \text{Modelo}$$

Este bucle no solo crea sesgos, sino que los intensifica con el tiempo. El resultado técnico directo de este fenómeno es la disminución de la diversidad de las recomendaciones y la intensificación de la homogeneización de los usuarios. La bibliografía subraya la gravedad de este ciclo:
One nature of RS is the feedback loop - the exposure mechanism of the RS determines user behaviors, which are circled back as the training data for the RS. Such feedback loop not only creates biases but also intensifies biases over time, resulting in the rich get richer Matthew effect (J. Chen y cols., 2021). Que se traduce en: Una naturaleza de los SR es el bucle de retroalimentación: el mecanismo de exposición del SR determina los comportamientos del usuario, los cuales vuelven al sistema como datos de entrenamiento para el SR. Dicho bucle de retroalimentación no solo crea sesgos, sino que también intensifica los sesgos con el tiempo, lo que resulta en el efecto Matthew de "el rico se hace más rico".
Además de los problemas de rendimiento, este sesgo representa un problema ético relevante, especialmente en plataformas multisectoriales (donde participan usuarios, proveedores y la propia plataforma). La equidad en la exposición es un objetivo crítico, ya que la sobrerrepresentación de ciertos artículos o proveedores plantea un problema de discriminación o injusticia. Los sesgos pueden socavar la satisfacción y la confianza del usuario. La necesidad de desarrollar estrategias de mitigación de sesgos que mantengan la imparcialidad sin sacrificar la precisión es un desafío importante para la gestión de tecnología.
Por lo tanto, la investigación sobre el sesgo de exposición y su mitigación es de vital importancia en el campo de la Gestión de Tecnología Informática.
Desde una perspectiva académica, este estudio contribuye a un cuerpo de conocimiento fragmentado sobre los sesgos en los sistemas de recomendación, buscando soluciones robustas ante la mezcla de diferentes tipos de sesgos que coexisten en el mundo real.
Desde una perspectiva práctica, la gestión de la tecnología requiere desarrollar e implementar soluciones que permitan a las organizaciones ofrecer recomendaciones justas y transparentes, lo que se traduce en el desarrollo de estrategias técnicas como enfoques de preprocesamiento, de procesamiento interno o de post-procesamiento,

con el fin de garantizar la equidad para todos los actores involucrados. Abordar este tema es urgente para transformar los modelos de investigación en mejoras prácticas tangibles para los sistemas comerciales.

1.1. Problemática

El panorama del marketing digital moderno se sustenta en la eficacia de los sistemas de recomendación impulsados por el aprendizaje automático para personalizar el contenido y optimizar las interacciones con los consumidores. Sin embargo, la dependencia de los modelos predictivos en los datos de comportamiento observacionales ha generado una problemática crucial: el sesgo de exposición. Este sesgo compromete la precisión de los resultados, socava la equidad para los diversos participantes del mercado y plantea serios dilemas éticos y empresariales.

1.1.1. Manifestación y Naturaleza del Sesgo de Exposición en Modelos Predictivos

El sesgo de exposición se define como una forma de sesgo de datos que surge en sistemas de recomendación donde la retroalimentación del usuario es típicamente implícita, con clics y visualizaciones. El sesgo ocurre porque los usuarios solo están expuestos a una parte de elementos específicos, por lo que las interacciones no observadas no siempre representan preferencias negativas, como lo expresaron Chen y otros autores (J. Chen y cols., 2021).
En esencia, el modelo predictivo tiene ambigüedad al interpretar las interacciones no observadas, ya que no puede distinguir si la falta de interacción se debe a un desinterés real o a la simple falta de exposición al elemento. Por lo tanto, el sesgo de exposición es un fenómeno en el que los ítems y proveedores no están representados equitativamente en los resultados de la recomendación, como expresaron Mansoury y otros autores (Mansoury y Masoud, 2022) en "User-centered Evaluation of Popularity Bias in Recommender Systems".

Analogía del Mesero y la Sopa de Pollo: Un Bucle de Retroalimentación

A continuación se expone una analogía planteada por Krause, Deriyeva, Beinke, Bartels, y Thomas (2024) para comprender el sesgo de exposición.
Imagine un restaurante donde un mesero recomienda platos a los comensales.

- El Origen del Sesgo: El mesero menciona su plato favorito, la sopa de pollo, con particular frecuencia (sobreexposición).
- La Decisión del Usuario (Comensal): Esta recomendación influye en las elecciones de los comensales, y algunos la ordenan debido a la sugerencia del mesero.
- La Interpretación del Sistema (Chefs): Los chefs, al ver un aumento en la demanda de sopa de pollo, asumen que los clientes realmente disfrutan de este plato y que es un éxito genuino.
- La Amplificación del Sesgo: Basándose en esta retroalimentación observada, los chefs refinan la sopa de pollo, instruyen a otros meseros para que también la recomienden frecuentemente e incluso añaden platos similares al menú, creyendo que eso es lo que el público desea.

El problema radica en que los chefs no son conscientes de que algunos invitados solo pidieron la sopa debido al respaldo (exposición) del mesero, y no porque fuera su preferencia inicial real. Así, el restaurante ha caído víctima del sesgo de exposición, donde el mecanismo de recomendación (el mesero) influye sistemáticamente en el comportamiento de elección observado, y el sistema (los chefs) amplifica ese sesgo en futuras recomendaciones.

1.1.2. Causas Fundamentales del Sesgo de Exposición

El problema del sesgo de exposición surge por múltiples factores, originado principalmente por la naturaleza de los datos de entrenamiento y la dinámica de la retroalimentación algorítmica.

Datos Sesgados y Popularidad

La base de los modelos predictivos son los datos de comportamiento del usuario. Según J. Chen y cols. (2021) los datos son observacionales en lugar de experimentales. Esta naturaleza observacional implica que la información se confunde por el mecanismo de exposición del sistema y la autoselección del usuario.

Una de las manifestaciones más comunes de este problema es el sesgo de popularidad, que es considerado una forma de sesgo de exposición. Esto ocurre porque los ítems no se presentan de manera uniforme: "some items are more popular than others and thus receive more user behaviors"(J. Chen y cols., 2021) (Traducción: "algunos elementos son más populares que otros y, por lo tanto, reciben más comportamientos de los usuarios").

Esta disparidad hace que los algoritmos, segun Mansoury y Masoud (2022), favorezcan los elementos populares, lo que conduce a una exposición injusta de otros elementos.

Retroalimentación Algorítmica y Amplificación del Sesgo
La causa más compleja del sesgo de exposición es el circuito de retroalimentación propio de los sistemas de recomendación. Estos sistemas operan en un "feedback loop"donde el mecanismo de exposición del sistema de recomendación determina los comportamientos del usuario, los cuales vuelven al sistema como datos de entrenamiento, como lo expresaron J. Chen y cols. (2021) en "Bias and Debias in Recommender System: A Survey and Future Directions ".
El sesgo puede ser amplificado con el tiempo a medida que los usuarios interactúan con los elementos recomendados en cada momento y sus interacciones se utilizarían como entrada para el algoritmo de recomendación en los momentos posteriores, como lo relata Mansoury y Masoud (2022).

Priorización de Clics o Conversiones
En el marketing digital, la optimización se centra a menudo en métricas inmediatas como clics (CTR) o conversiones. Esta priorización exacerba el sesgo al reforzar la exposición de los artículos que históricamente han funcionado bien, independientemente de la diversidad o la relevancia a largo plazo.

1.1.3. Impacto en la Diversidad, la Equidad y la Precisión

El sesgo de exposición tiene consecuencias directas y perjudiciales en la calidad de los resultados de los modelos predictivos y en la distribución equitativa de las oportunidades.
Problemas en la Diversidad y Precisión
El sesgo amplificado hacia los elementos populares puede impulsar el sesgo de popularidad, haciendo que los elementos relevantes pero impopulares no se muestren. Esto se traduce en una disminución de la diversidad, según Gupta, Wang, Lipton, y Wang (2021).
En cuanto a la precisión, ajustar ciegamente los datos sin considerar los sesgos daña la satisfacción del usuario y la confianza en el servicio de recomendación, como plantea J. Chen y cols. (2021). La eficacia de un modelo predictivo, por lo tanto, no solo se mide en precisión, sino también en cómo mitiga este sesgo.

Impacto en la Equidad (Fairness)

La equidad es un objetivo crítico, especialmente en plataformas multisectoriales, es decir, plataformas que reunen distintos grupos con un fin en común. La equidad se ve comprometida porque el algoritmo, segun Krause y cols. (2024), distribuye los beneficios y las cargas de manera desigual entre diferentes individuos o grupos.

El sesgo de exposición es una de las principales preocupaciones de equidad porque conduce a una exposición injusta de otros elementos, dicho por Mansoury y Masoud (2022). Esto afecta negativamente a los elementos menos populares, a los elementos nuevos en el sistema e incluso al proveedor de estos elementos. En contextos donde los ítems son provistos por diferentes proveedores, como en plataformas multisectoriales, la falta de exposición justa es un problema de los diversos grupos participantes.

1.1.4. Implicaciones Éticas, Sociales y Empresariales

El sesgo de exposición en el marketing digital trasciende el rendimiento algorítmico, impactando en la sociedad, la ética y los resultados económicos.

Riesgos éticos y sociales para consumidores

A nivel social y ético, el sesgo de exposición contribuye a fenómenos conocidos como las burbujas de filtro y las cámaras de eco. Según Kidwai, Akhtar, y Nadeem (2023), el fenómeno de la burbuja de filtro ha generado preocupaciones sobre el impacto potencial en las experiencias del usuario y la exposición a la información y sobre la exposición de los usuarios a una variedad de ideas y contenido.

Además, J. Chen y cols. (2021) dice que el sesgo puede reforzar sesgos históricos hacia una mayoría y evitar que las minorías sean sociales con gran alcance. En la literatura se identifica la necesidad de asegurar que personas de diferentes grupos sensibles tengan la oportunidad de aparecer en los puestos más altos de las listas de recomendación.

Riesgos Empresariales

En el entorno empresarial, el sesgo de exposición tiene implicaciones críticas, especialmente en plataformas de comercio electrónico, donde, segun M. Mansoury y Mobasher (2023), se sesga la exposición del usuario a elementos populares, reduciendo de alguna manera, la diversidad de las recomendaciones. Para los proveedores y minoristas, la falta de exposición proporcional significa que sus elementos no reciben atención proporcional.

La investigación de este problema es fundamental, ya que, según Sharma y Sachin (2024), existe una creciente dependencia de la toma de decisiones algorítmica en

entornos de comercio electrónico y los posibles efectos que puede tener en el comportamiento del cliente.

1.1.5. Pertinencia para la Gestión de Tecnología Informática

La investigación sobre la mitigación del sesgo de exposición es relevante para la Gestión de Tecnología Informática, dado el papel integral de estos sistemas en las operaciones empresariales modernas.
Los sistemas de recomendación, basados en aprendizaje automático, son herramientas indispensables de la tecnología de la información. Como lo comentan Sun y Fu (2025), estos sistemas se han vuelto cada vez más complejos e integrales para el funcionamiento de industrias que van desde la atención médica hasta las finanzas, y desde la educación hasta el entretenimiento.
Luego, la gestión de estos sistemas implica no solo optimizar la eficiencia y la escalabilidad de la infraestructura de TI, sino también asegurar que los resultados de los modelos predictivos sean éticos y sostenibles. El desarrollo de estrategias de mitigación de sesgos son tareas que requieren una comprensión profunda de la ciencia de datos, la causalidad (como herramienta matemática efectiva para clarificar relaciones causales) y la ingeniería de software.

1.2. Objetivos

1.2.1. Objetivos generales

Analizar el impacto del sesgo de exposición en modelos predictivos aplicados al marketing digital para identificar cómo afecta la equidad algorítmica, la diversidad de opciones de consumo y la efectividad de las campañas de marketing digital.

1.2.2. Objetivos específicos

1. Cuantificar el sesgo de exposición mediante métricas de desigualdad, equidad algorítmica y diversidad.

2. Evaluar el impacto del sesgo de exposición en la diversidad de opciones de consumo disponibles para los usuarios mediante el análisis de restricciones

sistemáticas.

3. Validar técnicas de mitigación de sesgo evaluando su efectividad en términos de equidad algorítmica y rendimiento predictivo.

4. Analizar compensación entre equidad algorítmica y efectividad de campañas publicitarias para determinar si es posible mejorar simultáneamente ambos objetivos.

5. Caracterizar los diferentes tipos de sesgo de exposición en modelos predictivos aplicados a la segmentación y personalización en marketing digital mediante el desarrollo de una taxonomía sistemática.

1.3. Alcance

La presente investigación se enfoca en el análisis del sesgo de exposición en modelos predictivos aplicados específicamente al contexto del marketing digital, considerando sistemas de segmentación y personalización de campañas publicitarias. El trabajo aborda tres dimensiones principales del impacto del sesgo: equidad algorítmica, diversidad de opciones de consumo y efectividad operativa de las campañas.
Los aportes de esta investigación comprenden:

- Evidencia empírica cuantificable del sesgo de exposición en marketing digital
- Taxonomía sistemática de cinco tipos de sesgo con métricas específicas
- Validación empírica comparativa de técnicas de mitigación
- Marco metodológico para auditoría continua de sesgo
- Análisis de trade-offs entre equidad y efectividad

Capítulo 2

Marco Teórico

2.1. Modelos Predictivos en Marketing Digital

Los modelos predictivos han transformado fundamentalmente las estrategias de marketing digital, permitiendo la segmentación precisa de audiencias y la personalización de campañas publicitarias a escala. Estos modelos utilizan técnicas de machine learning para predecir comportamientos de usuarios, como la probabilidad de hacer click en un anuncio, completar una compra o interactuar con contenido específico (Narang y Shankar, 2019)

La adopción de modelos predictivos en marketing digital se ha acelerado debido a la disponibilidad de datos masivos sobre comportamiento de usuarios, avances en técnicas de aprendizaje automático, y la necesidad de optimizar el retorno de inversión en publicidad. Sin embargo, esta dependencia creciente de algoritmos para tomar decisiones sobre distribución de recursos y oportunidades de exposición ha dado lugar a preocupaciones sobre sesgos algorítmicos que pueden crear desigualdades sistemáticas, como lo mencionan J. Chen y cols. (2021)

Los modelos predictivos en marketing digital típicamente operan en un pipeline que incluye: recolección de datos de comportamiento de usuarios, extracción de features relevantes (demográficas, contextuales, históricas), entrenamiento de modelos predictivos (como redes neuronales), generación de predicciones (probabilidades de click, conversión, etc.), y utilización de estas predicciones para ranking y selección de anuncios a mostrar (Narang y Shankar, 2019)

2.2. Inteligencia Artificial Aplicada al Marketing

La inteligencia artificial ha revolucionado el marketing digital mediante la automatización de decisiones complejas sobre segmentación, personalización y optimización de campañas. Los sistemas de recomendación y selección utilizan algoritmos de machine learning para identificar qué contenido o anuncios mostrar a cada usuario, optimizando métricas como click-through rate (CTR), tasa de conversión o retorno de inversión (ROI) (J. Chen y cols., 2021)
El funcionamiento de estos modelos se basa en el aprendizaje automático, el cual se entrena utilizando los datos de comportamiento del usuario. Sin embargo, la naturaleza de estos datos presenta un desafío fundamental:
User behavior data, which lays the foundation for recommendation model training, is observational rather than experimental (J. Chen y cols., 2021). Traducción: "Los datos de comportamiento del usuario, que sientan las bases para el entrenamiento del modelo de recomendación, son observacionales en lugar de experimentales".

2.3. Sesgos algorítmicos: concepto general, clasificación y causas

La dependencia de los modelos predictivos respecto a los datos observacionales introduce inevitablemente el problema de los sesgos. El sesgo, en el contexto de los sistemas de recomendación, se entiende como un fenómeno de desviación de la distribución de datos.
"The sample selection or user decision would inevitably be affected by many undesirable factors, such as the exposure mechanism of RS or public opinions, making the training data distribution deviate from test distribution"(J. Chen y cols., 2021). Traducción: "La selección de la muestra o la decisión del usuario se verían inevitablemente afectadas por muchos factores indeseables, como el mecanismo de exposición de los SR o las opiniones públicas, lo que hace que la distribución de los datos de entrenamiento se desvíe de la distribución de prueba".
Existen múltiples tipos de sesgos que afectan a los sistemas de recomendación. J. Chen y cols. (2021) resumen siete tipos de sesgos en los sistemas de recomendación. Los más relevantes para el contexto de los datos son:

- Sesgo de Selección (Selection Bias): Ocurre cuando los datos observados no son una muestra representativa de todas las interacciones posibles.

- Sesgo de Posición (Position Bias): Los ítems mejor clasificados tienen más probabilidades de ser seleccionados, independientemente de su relevancia, ya que los usuarios exploran solo una lista truncada.
- Sesgo de Exposición (Exposure Bias): Es el foco principal de este trabajo y se detalla en la siguiente sección.
- Sesgo de Popularidad (Popularity Bias): Ítems populares son sobrerrecomendados, lo que afecta negativamente a la diversidad.

Una causa crucial de la amplificación de los sesgos es el bucle de retroalimentación (feedback loop).

2.3.1. Definiciones Formales de Fairness Metrics

Para caracterizar formalmente los sesgos algorítmicos, es necesario recurrir a las definiciones estándar de métricas de equidad (fairness) establecidas en la literatura. Barocas y Selbst (2016) identifican las siguientes categorías principales de definiciones de equidad:

Demographic Parity (Paridad Demográfica): Un algoritmo satisface paridad demográfica si la probabilidad de recibir un resultado positivo es la misma para todos los grupos demográficos. Formalmente, para grupos A y B, se requiere que:

$$P(\hat{Y} = 1|A) = P(\hat{Y} = 1|B)$$

donde $\hat{Y}$ es la predicción del modelo. La diferencia de paridad demográfica (Demographic Parity Difference) mide la violación de esta condición:

$$DPD = |P(\hat{Y} = 1|A) - P(\hat{Y} = 1|B)|$$

Equal Opportunity (Igualdad de Oportunidad): Un algoritmo satisface igualdad de oportunidad si la tasa de verdadero positivo es la misma para todos los grupos. Formalmente:

$$P(\hat{Y} = 1|Y = 1, A) = P(\hat{Y} = 1|Y = 1, B)$$

donde Y es la etiqueta verdadera. La diferencia de igualdad de oportunidad (Equal Opportunity Difference) mide:

$$EOD = |P(\hat{Y} = 1|Y = 1, A) - P(\hat{Y} = 1|Y = 1, B)|$$

Además de estas definiciones de equidad, es importante caracterizar los tipos de sesgo. Según la taxonomía de Barocas y Selbst (2016), los sesgos pueden originarse en:

- Sesgo de muestra: Datos de entrenamiento no representativos
- Sesgo de representación: Features que codifican estereotipos o proxies de características protegidas
- Sesgo de evaluación: Métricas que no capturan adecuadamente la equidad

2.4. Sesgo de exposición: definición, origen, mecanismos y ejemplos en contextos digitales

El sesgo de exposición es un tipo de sesgo de datos que surge de la naturaleza de la retroalimentación implícita, donde las interacciones no observadas son equívocas. "Exposure bias happens as users are only exposed to a part of specific items so that unobserved interactions do not always represent negative preference "(J. Chen y cols., 2021). Traducción: "El sesgo de exposición ocurre ya que los usuarios solo están expuestos a una parte de ítems específicos, de modo que las interacciones no observadas no siempre representan una preferencia negativa".
Esta ambigüedad es fundamental: una interacción no observada puede significar que el ítem no interesa al usuario, o bien, que el usuario simplemente no fue consciente de su existencia. La incapacidad de los modelos para distinguir entre interacciones negativas reales (expuesto pero desinteresado) y las potencialmente positivas (no expuesto) provoca sesgos graves.

Origen y Mecanismos
Se define el sesgo de exposición en los sistemas de recomendación como "the fact that some items and suppliers are over-represented in recommendation results, while other items and suppliers are not adequately represented"(Mansoury y Masoud, 2022). Traducción: "el hecho de que algunos ítems y proveedores están sobrerrepresentados en los resultados de recomendación, mientras que otros ítems y proveedores no están representados adecuadamente".
Las investigaciones sugieren varias dimensiones de la exposición de datos:

- Política del sistema previo: La exposición está influenciada por la política de SR anteriores, que controla qué ítems mostrar, llevando a que algunos trabajos lo denominen "sesgo del modelo previo".
- Selección del usuario: Si los usuarios buscan activamente, los ítems muy relevantes tienen más probabilidades de ser expuestos, lo que a veces se llama "sesgo de selección del usuario".
- Popularidad: Los ítems populares tienen más probabilidades de ser vistos por los usuarios.

Este sesgo se amplifica con el tiempo debido al ciclo de retroalimentación. En un entorno dinámico, el sistema recomienda repetidamente ciertos ítems, y las interacciones de los usuarios con esos ítems amplificarán el sesgo hacia esos ítems con el tiempo, como lo menciona Mansoury y Masoud (2022).

Ejemplos El sesgo de exposición puede observarse en múltiples plataformas digitales contemporáneas. Los siguientes ejemplos ilustran cómo este sesgo se manifiesta en contextos reales:

- YouTube: El algoritmo de recomendación de YouTube puede crear "burbujas de filtro"donde usuarios reciben recomendaciones principalmente de contenido similar a lo que previamente han consumido. Por ejemplo, si un usuario muestra interés inicial en contenido educativo sobre ciencia, el algoritmo puede concentrar futuras recomendaciones en este tipo de contenido, limitando la exposición a otros géneros o perspectivas. Esto crea sesgo de confirmación, donde la diversidad de opciones se restringe progresivamente.
- Instagram Ads: Los algoritmos de segmentación en Instagram Ads pueden crear disparidades demográficas en exposición a oportunidades. Por ejemplo, anuncios de oportunidades laborales de alto nivel pueden mostrarse predominantemente a usuarios con ciertas características demográficas (género, edad, ubicación), mientras que usuarios con perfiles similares pero de grupos diferentes reciben sistemáticamente menos exposición.

Estos ejemplos ilustran que el sesgo de exposición no es un problema teórico, sino una realidad observable en sistemas de marketing y recomendación ampliamente utilizados, con implicaciones significativas para equidad y diversidad de acceso a oportunidades y contenidos.

2.5. Implicaciones éticas: impacto sobre la equidad, la transparencia y la toma de decisiones

El sesgo de exposición conlleva importantes implicaciones éticas y desafíos para la gestión tecnológica, especialmente en lo que respecta a la equidad y la transparencia en las plataformas de múltiples partes interesadas.

Equidad y Multi-Sided Fairness: La equidad se ha convertido en un objetivo crítico a nivel de sistema en los sistemas de recomendación. Según Mansoury y Masoud (2022), en plataformas multisectoriales, es crucial optimizar las utilidades no solo para el usuario final, sino también para otros actores, como los vendedores o productores de ítems, que desean una representación justa de sus ítems.

El sesgo de exposición, al estar ligado a la popularidad, conduce a una exposición desigual, lo que se traduce en un problema de discriminación o injusticia, tanto para proveedores como para usuarios finales. Esta última cuestión es abordada por Mansoury y Masoud (2022), quienes han demostrado que el sesgo de popularidad puede causar injusticia desde la perspectiva de los usuarios, ya que no todos son tratados de igual manera según sus intereses hacia los ítems populares.

Transparencia y Gestión de la Toma de Decisiones: La gestión de los sistemas de recomendación requiere un enfoque proactivo en la mitigación de sesgos. El reinforcement learning se ha propuesto como un enfoque para intervenir en el sistema con estrategias más inteligentes para beneficios a largo plazo, buscando un equilibrio adaptativo entre la exploración y la explotación.

Para abordar la transparencia y la comprensión de los sesgos, el uso de gráficos causales (Causal Graphs) se está convirtiendo en una herramienta poderosa.

"Cause graph is an effective mathematical tool for elucidating potentially causal relationships from data, deriving causal relationships from combinations of knowledge and data, predicting the effects of actions, and evaluating explanations for observed events and scenarios"(J. Chen y cols., 2021). Traducción: "El gráfico causal es una herramienta matemática eficaz para esclarecer posibles relaciones causales a partir de datos, derivar relaciones causales a partir de combinaciones de conocimientos y datos, predecir los efectos de las acciones y evaluar explicaciones para eventos y escenarios observados".

Los gráficos causales permiten modelar las relaciones complejas y los efectos de confusión (confounding bias), donde factores como la exposición previa influyen tanto en las características del ítem/usuario como en las calificaciones finales. El counterfactual learning (aprendizaje contrafactual) se utiliza para inferir cómo habría sido la interacción si la exposición no hubiera ocurrido, ayudando a obtener un modelo

imparcial (unbiased model) que refleje los verdaderos intereses de los usuarios. En última instancia, la gestión tecnológica debe asegurar que los modelos predictivos en el marketing digital no solo sean precisos, sino también equitativos y transparentes, promoviendo un mercado digitalmente inclusivo, como lo nombra Irena, Wahyudi, Wairooy, y Andro (2024).

Capítulo 3

Estado del arte

La presente investigación se releva a partir de una revisión sistemática de la literatura, enfocada en sesgos que se producen en modelos predictivos que se implementan en campañas de marketing digital.
Las fuentes proporcionadas ofrecen una revisión profunda sobre los sistemas de recomendación y los desafíos que presentan los sesgos en ellos. La primera revisión arrojó aproximadamente 50 fuentes, de las cuales, luego de ejecutar un filtro que ajustaba los escritos a la relevancia de la investigación, han quedado 18.
En el Cuadro 3.1 se exponen los trabajos que se utilizaron para construir el presente Estado del Arte y que resultan como base para la presente investigación. El cuadro identifica los títulos de cada trabajo y cada columna representa las temáticas que aborda. La última columna representa el tipo de documento.

Cuadro 3.1: Trabajos seleccionados del estado del arte

Título del Trabajo	**Marketing**	**Sesgos**	**Sesgo Exp.**	**Modelos Pred.**	**Tipo**
Bias and Debiasing in Recommender System: A Survey and Future Directions	No	Sí	Sí	Parcial	Survey/Revisión
Correcting Exposure Bias for Link Recommendation	No	Sí	Sí	Parcial	Empírico
Fairness and Bias in E-commerce Recommendation Systems: A Literature Review	Sí	Sí	Parcial	Parcial	Survey/Revisión
Fairness of Exposure in Dynamic Recommendation	No	Sí	Sí	Parcial	Empírico
From Data to Decisions: The Power of Machine Learning in Business Recommendations	Sí	No	No	Sí	Aplicado
The Optimized Probabilistic Recommendation Model in Exposure	No	Sí	Sí	Sí	Empírico
Multi-sided Exposure Bias in Recommendation	No	Sí	Sí	Parcial	Empírico
Unbiased Cascade Bandits: Mitigating Exposure Bias in Online Learning to Rank Recommendation	No	Sí	Sí	Sí	Empírico
Understanding and Mitigating Multi-sided Exposure Bias in Recommender Systems	No	Sí	Sí	Parcial	Empírico
Mitigating Exposure Bias in Online Learning to Rank Recommendation: A Novel Reward Model for Cascading Bandits	No	Sí	Sí	Sí	Empírico
The Relevance of Item-Co-Exposure For Exposure Bias Mitigation	No	Sí	Sí	Parcial	Empírico

A continuación se detalla en que situación se encuentran los sistemas de recomendación implementados en el marketing digital y que relación tienen actualmente con el sesgo de exposición.

3.1. Introducción y Contexto Actual de la Investigación

Los sistemas de recomendación, según Gangadharan y cols. (2025), se han consolidado como herramientas esenciales en la era digital para filtrar y sugerir contenido en línea de manera efectiva. La función de estos sistemas es crucial para mitigar la sobrecarga de información, ofreciendo a los usuarios ítems basados en sus intereses personalizados. La tecnología de la información se ha vuelto sumamente compleja, abarcando infraestructuras de computación en la nube y análisis de big data, las cuales son fundamentales para mejorar la eficiencia y la toma de decisiones, como lo expresan Sun y Fu (2025).
Sun y Fu (2025) dicen que el aprendizaje automático ha transformado la operación de los sistemas de recomendación, permitiendo que mejoren con la experiencia y se ajusten a nuevos datos de manera automática. Esta adaptabilidad es vital en entornos digitales que cambian rápidamente, donde las preferencias de los usuarios evolucionan constantemente y el volumen de datos no deja de crecer. Los motores de recomendación, impulsados por algoritmos avanzados y análisis de datos, contribuyen significativamente a aumentar las ventas, los ingresos y la ventaja competitiva de las empresas.

3.2. Sesgo de Exposición en Sistemas Predictivos: Definición y Manifestaciones

A pesar de los logros de los sistemas de recomendación, estos enfrentan serios problemas de sesgo que pueden deteriorar la efectividad de las recomendaciones, como lo resume J. Chen y cols. (2021) y Y. Chen, Li, Liu, y Wu (2023). El sesgo de exposición es una de las limitaciones persistentes en los algoritmos de recomendación.
La investigación sobre sesgo de exposición en sistemas de recomendación ha documentado sistemáticamente cómo los algoritmos pueden crear distribuciones desiguales de exposición. (J. Chen y cols., 2021) en su trabajo sobre "Debiasing Recommendation by Learning Identifiable Latent Confounders"presentado en conferencias de

sistemas de recomendación, demuestran empíricamente cómo el sesgo de exposición surge de la confusión entre preferencias reales del usuario y factores de exposición, mostrando que los modelos tradicionales que no consideran este sesgo pueden generar recomendaciones que perpetúan desigualdades en la distribución de visibilidad.

3.3. Implicaciones Tecnológicas y Desafíos Abiertos

El sesgo de exposición puede provocar consecuencias indeseadas, como la formación de burbujas de filtro y cámaras de eco, tal como lo mencionan Krause y cols. (2024). Un desafío inherente a los sistemas de recomendación es el bucle de retroalimentación (feedback loop).
La mayoría de los estudios sobre el sesgo de exposición se han llevado a cabo en entornos estáticos, analizando una sola ronda de resultados de recomendación. Sin embargo, el impacto a largo plazo de este sesgo en modelos dinámicos de aprendizaje para clasificar es un área que requiere una exploración significativa.
Además, persisten limitaciones relacionadas con la calidad y disponibilidad de los datos, ya que muchos algoritmos requieren conjuntos de datos extensos que a menudo resultan incompletos, sesgados o de baja calidad, como lo expresan Sun y Fu (2025).

3.4. Avances Tecnológicos Recientes en la Mitigación del Sesgo

Los avances más recientes en el desarrollo tecnológico se centran en la implementación de técnicas robustas y modelos basados en la causalidad para contrarrestar el sesgo de exposición en diferentes etapas del proceso de recomendación.

- **Enfoques Basados en Inferencia Causal**
  La inferencia causal ha ganado atención como una estrategia prometedora, ya que permite a los sistemas de recomendación basar sus recomendaciones en relaciones causales estables e invariantes, en lugar de depender únicamente de correlaciones observacionales, como lo expresan Zhu, Ma, y Li (2023).

  1. Modelos Causalidad y Aprendizaje Contrario a los Hechos:

- El Grafo Causal es una herramienta matemática efectiva que se considera muy prometedora para las tareas de eliminación de sesgos en los sistemas de recomendación, ya que facilita el razonamiento sobre la ocurrencia, la causa y el efecto del sesgo.
- Segun J. Chen y cols. (2021), la clave para el éxito en la eliminación de sesgos es la capacidad de razonar sobre la ocurrencia, la causa y el efecto en los modelos o datos de recomendación.
- Se han propuesto estimadores que mitigan el bucle de retroalimentación aprovechando las probabilidades de exposición conocidas, lo que permite que se sigan recomendando ítems relevantes a pesar de su baja propensión (Gupta y cols., 2021)

2. **Sistemas Dinámicos y Aprendizaje por Refuerzo**
   Los sistemas que pueden integrarse en tiempo real y adaptarse a las preferencias dinámicas son cruciales para el futuro de los sistemas de recomendación, com lo expresaron Sun y Fu (2025). Estos autores señalan que la investigación futura debe enfocarse en algoritmos que aprendan y se adapten en tiempo real en función de las interacciones del usuario y la retroalimentación, posiblemente empleando técnicas de aprendizaje por refuerzo.
3. **Nuevos Modelos y Enfoques de Datos**

   - Modelos de Elección Discreta: Krause y cols. (2024) señalan que se ha encontrado que incorporar modelos de elección discreta en los sistemas de recomendación que utilizan feedback implícito reduce de manera efectiva el sesgo de exposición. Estos modelos logran la mitigación al registrar y considerar los ítems observados pero no elegidos durante el entrenamiento. Además, han revelado que la composición de los conjuntos de elección es otra fuente de sesgo de exposición que estudios previos habían descuidado.
   - Transformación de Ratings (Preprocesamiento): Se ha propuesto una técnica de preprocesamiento que transforma las calificaciones de ítems en valores percentiles antes de la generación de la recomendación, lo cual ayuda a aliviar el sesgo de popularidad inherente en los datos de entrada. En la metodología de el presente trabajo es una de las técnicas que se aborda para intentar mitigar el sesgo de exposición.

3.5. Conclusión y Prospectiva

Un desafío clave para la investigación es la creación de un marco general de eliminación de sesgos que pueda abordar la combinación de múltiples sesgos que coexisten en el mundo real. Es imperativo que las plataformas de comercio electrónico, y por extensión, el marketing digital, prioricen el desarrollo de estrategias que impacten en la equidad sin comprometer la efectividad de los sistemas. El desafío del sesgo de exposición en el marketing digital está directamente asociado con el desafío que presentan los modelos predictivos que solo pueden aprender de lo que se les muestra (la exposición sesgada), lo que perpetúa la sobrerrepresentación de los ítems populares y oculta el verdadero interés del usuario en lo que no fue expuesto, como lo señalan J. Chen y cols. (2021). Los avances tecnológicos actuales buscan introducir mecanismos para mirar más allá de los datos presentados y determinar lo que el usuario realmente elegiría bajo una exposición equitativa.

Capítulo 4

Metodología

La investigación adopta un enfoque empírico-cuantitativo que combina:

- Análisis exploratorio de datos (EDA): Identificación de patrones de sesgo mediante estadística descriptiva y visualizaciones
- Métricas cuantitativas de equidad: Implementación de métricas estándar en la literatura de equidad algorítmica
- Validación experimental: Comparación controlada de técnicas de mitigación mediante diseño experimental

Para comprender la estructura metodológica de esta investigación, la Figura 4.1 presenta la arquitectura general del sistema experimental, mostrando el pipeline completo desde los datos crudos hasta el análisis final de resultados. Esta arquitectura ilustra cómo los componentes del sistema se integran para permitir la caracterización del sesgo, la evaluación de técnicas de mitigación, y el análisis de efectividad. La figura proporciona una visión sistémica del proceso de investigación.

4.1. Obtención de datos

Para el desarrollo de la presente investigación se evaluaron múltiples repositorios de datos vinculados al desempeño de campañas publicitarias digitales y a modelos predictivos utilizados en marketing. El propósito central de esta revisión fue identificar un conjunto de datos que permita analizar empíricamente el sesgo de exposición en procesos de segmentación y personalización algorítmica, considerando tanto variables de comportamiento del usuario como métricas de rendimiento publicitario.

Figura 4.1: Pipeline de datos a resultados

Tras el análisis comparativo, se seleccionó el Ad Click Prediction Dataset (.1) por ser el que ofrece las características técnicas y estructurales más adecuadas para abordar los objetivos de la investigación.

En primer lugar, a diferencia de otros datasets centrados exclusivamente en métricas agregadas de campaña (impresiones, clics, costo o posición del anuncio), este conjunto de datos incluye atributos asociados a los usuarios, tales como edad, género, país y tipo de dispositivo. La disponibilidad de variables vinculadas a características demográficas o contextuales del usuario resulta indispensable para examinar diferencias en la distribución de impresiones entre segmentos, lo que permite operacionalizar y

medir el sesgo de exposición de manera rigurosa.
En segundo lugar, el dataset contiene tanto información de impresiones como de resultados de interacción (clics), lo que facilita la construcción de modelos predictivos representativos del proceso real de decisión algorítmica empleado en plataformas de publicidad digital. Esto habilita no solo el análisis del comportamiento del modelo, sino también la evaluación del impacto del sesgo de exposición sobre la efectividad de la segmentación y la personalización.
En tercer lugar, su estructura a nivel de registro individual por impresión permite aplicar métricas de auditoría algorítmica tales como disparidad de exposición, paridad demográfica, igualdad de oportunidades o índices de concentración. Tales métricas son esenciales para evaluar en qué medida la asignación desigual de oportunidades de visualización puede afectar la equidad, la diversidad de opciones de consumo y la efectividad empresarial.
Finalmente, el dataset presenta un volumen adecuado y una variedad de variables suficiente para implementar técnicas de mitigación de sesgo y evaluar su impacto en modelos de predicción del clic, lo que responde directamente al segundo objetivo específico del estudio, orientado a proponer y validar estrategias metodológicas de corrección.
En función de todo lo expuesto, el Ad Click Prediction Dataset constituye la opción más robusta y pertinente para el análisis del sesgo de exposición en modelos predictivos aplicados al marketing digital, permitiendo abordar de manera empírica y demostrable las dimensiones de equidad, comportamiento algorítmico y efectividad de las estrategias de segmentación que se buscan estudiar.

4.1.1. Limitaciones del Dataset

El análisis del Ad Click Prediction Dataset requiere consideraciones éticas importantes relacionadas con el manejo de datos personales y la protección de privacidad. Aunque el dataset utilizado en este trabajo ha sido anonimizado y no contiene información identificable personalmente, es importante reconocer que los datos de comportamiento digital pueden reflejar sesgos estructurales inherentes a la sociedad y a los sistemas que los generan.
El dataset puede contener sesgos inherentes originados en desigualdades históricas en acceso a tecnología o servicios, sesgos previos en algoritmos utilizados para recolección o selección de datos, representación desbalanceada de grupos demográficos en los datos originales.
Estos sesgos inherentes pueden afectar los análisis realizados y limitar la generalizabilidad de los resultados. Por lo tanto, los hallazgos deben interpretarse en el

contexto específico del dataset y reconocerse que pueden no ser generalizables a otros contextos sin validación adicional.
Por otra parte, el dataset presenta las siguientes limitaciones que deben considerarse al interpretar los resultados:
Limitación temporal: El dataset representa un snapshot temporal y no captura evolución temporal del sesgo o efectos de feedback loops a largo plazo.
Limitación de contexto: El dataset proviene de un contexto específico de marketing digital y los resultados pueden no generalizarse a otros dominios o industrias.
Limitación de tamaño: Con 10,000 impresiones y 4,000 usuarios, el dataset es de tamaño moderado y puede no capturar todos los patrones de sesgo que podrían observarse en datasets más grandes.
Estas limitaciones no invalidan los hallazgos, pero requieren que los resultados se interpreten con cautela y que futuros trabajos validen los hallazgos en contextos adicionales.

4.2. Fases de la investigación

Esta investigación aborda el análisis del sesgo de exposición en modelos predictivos aplicados al marketing digital mediante un enfoque metodológico estructurado en cinco fases principales. El diseño metodológico está fundamentado en la necesidad de comprender sistemáticamente cómo los algoritmos de selección y recomendación pueden crear desigualdades en las oportunidades de exposición, y cómo estas desigualdades impactan tanto en la equidad del sistema como en su efectividad operativa.
A continuación se expone un gráfico de las fases metodológicas que presenta la siguiente investigación.
En la Figura 4.2 cada fase corresponde a un análisis específico que contribuye de manera progresiva a los objetivos generales y específicos del trabajo de investigación. La metodología está diseñada para permitir una comprensión incremental del problema: comenzando con la caracterización inicial del dataset y la identificación de patrones de sesgo, avanzando hacia el análisis de diversidad y la taxonomía de tipos de sesgo, y culminando con la evaluación de técnicas de mitigación y su impacto en métricas de negocio. Este enfoque secuencial permite construir conocimiento de manera sistemática, donde cada fase informa y enriquece las fases subsiguientes.
La metodología incorpora principios de reproducibilidad, transparencia y responsabilidad ética, asegurando que los resultados sean verificables, interpretables y consideren las implicaciones para diferentes grupos demográficos. Además, el diseño

Figura 4.2: Fases metodológicas

metodológico está orientado a proporcionar no solo diagnósticos del problema, sino también evaluaciones prácticas de soluciones potenciales, balanceando consideraciones de equidad con métricas de efectividad operativa.

4.2.1. Fase 1: Exploración y Caracterización del Dataset

La primera fase tiene como objetivo realizar una exploración exhaustiva del dataset para caracterizar la distribución de exposición, identificar patrones de sesgo iniciales y establecer métricas de referencia que servirán como baseline para todas las fases subsiguientes. Esta fase es fundamental porque proporciona la base empírica sobre la cual se construyen todos los análisis posteriores.

Análisis inicial

El análisis descriptivo inicial constituye la base de toda la investigación y se enfoca en tres componentes principales. Primero, se realiza la carga y validación de datos, que incluye la verificación de integridad referencial (asegurando que las relaciones entre tablas sean consistentes), la verificación de tipos de datos (confirmando que cada variable tenga el tipo apropiado para su contenido), y la identificación de valores faltantes (que podrían indicar problemas en la recolección de datos o requerir estrategias de imputación).

Segundo, se calculan estadísticas descriptivas para todas las variables, tanto numéricas como categóricas. Para variables numéricas, esto incluye medidas de tendencia central (media, mediana), medidas de dispersión (desviación estándar, rango, cuartiles), y análisis de distribución (asimetría, curtosis). Para variables categóricas, esto incluye frecuencias, proporciones, y análisis de moda. Estas estadísticas proporcionan una comprensión inicial de las características del dataset y permiten identificar valores atípicos o distribuciones inesperadas que podrían requerir atención especial.

Tercero, se realiza la identificación y caracterización de dimensiones relevantes para el análisis de sesgo. Esto incluye variables demográficas (género, edad), que son fundamentales para detectar sesgos relacionados con características de los usuarios; variables de dispositivo (tipo de dispositivo), que pueden revelar sesgos relacionados con la tecnología utilizada; y variables de contexto (posición del anuncio, historial de navegación, hora del día), que pueden revelar sesgos relacionados con el contexto de la interacción. Esta caracterización multidimensional es crucial porque el sesgo de exposición puede manifestarse de diferentes maneras según la dimensión considerada.

Análisis de Distribución de Exposición

El análisis de distribución de exposición es fundamental para entender cómo se distribuyen las oportunidades de interacción entre usuarios. El proceso comienza con el cálculo de exposición por usuario, que implica agregar todas las impresiones recibidas por cada identificador único de usuario. Esta agregación permite transformar datos a nivel de impresión en datos a nivel de usuario, lo cual es necesario para analizar desigualdades en la distribución de oportunidades.

Una vez calculada la exposición por usuario, se calculan estadísticas descriptivas de la distribución, incluyendo media, mediana, desviación estándar y cuartiles. La media proporciona el promedio de exposición, pero puede ser engañosa si la distribución es altamente asimétrica. La mediana es más robusta a valores atípicos y proporciona el valor central de la distribución. La desviación estándar indica la variabilidad en la exposición, y valores altos sugieren desigualdad significativa. Los cuartiles permiten entender cómo se distribuye la exposición en diferentes segmentos de la población.

La visualización de la distribución mediante histogramas y diagramas de caja es crucial para identificar asimetrías y valores atípicos que podrían no ser evidentes en las estadísticas numéricas. Los histogramas revelan la forma de la distribución, permitiendo identificar si sigue una distribución normal, si está sesgada hacia valores bajos o altos, o si tiene múltiples modas. Los diagramas de caja permiten identificar valores atípicos y comparar distribuciones entre diferentes grupos. Estas visualizaciones son especialmente importantes porque el sesgo de exposición a menudo se manifiesta como distribuciones altamente asimétricas donde un pequeño grupo de usuarios concentra la mayoría de las impresiones.

Análisis de CTR por Grupos Demográficos

El análisis de Click-Through Rate (CTR) por grupos demográficos es esencial para entender si las diferencias en efectividad están relacionadas con características demográficas o si, por el contrario, el sesgo se manifiesta principalmente a través de diferencias en la distribución de oportunidades de exposición. Este análisis comienza con la segmentación de usuarios por grupos demográficos, incluyendo género, edad y tipo de dispositivo.

Para cada grupo, se calculan métricas específicas que incluyen el CTR promedio por grupo (proporción de clicks sobre impresiones), el número total de impresiones recibidas por el grupo, y el número total de clicks generados por el grupo. Estas métricas permiten comparar tanto la efectividad (CTR) como el volumen de interacción (impresiones y clicks) entre grupos.

La identificación de disparidades se realiza mediante comparación sistemática de métricas entre grupos. Si diferentes grupos muestran CTR similares pero diferentes volúmenes de impresiones, esto sugiere que el sesgo se manifiesta principalmente a través de diferencias en distribución de oportunidades, no en diferencias en efectividad. Por el contrario, si diferentes grupos muestran diferentes CTR, esto podría indicar que los anuncios están mejor calibrados para ciertos grupos, lo cual también representa una forma de sesgo.

Cálculo de Métricas de Desigualdad

Para cuantificar la desigualdad en la distribución de exposición, se utiliza el coeficiente de Gini, una métrica ampliamente utilizada en economía y ciencias sociales. Para medir esto, se utiliza la siguiente formula:

$G = \frac{2\sum_{i=1}^{n} i \cdot x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}$

donde x_i representa la exposición del usuario i ordenado de menor a mayor.

El coeficiente toma valores entre 0 y 1, donde 0 representa igualdad perfecta (todos los usuarios reciben el mismo número de impresiones) y 1 representa desigualdad

máxima (un único usuario recibe todas las impresiones). Valores típicos de interpretación:

- $G < 0{,}3$: desigualdad baja-moderada
- $0{,}3 \leq G < 0{,}5$: desigualdad moderada-alta
- $G \geq 0{,}5$: desigualdad alta-extrema

Además del coeficiente de Gini global, se realiza un análisis de exposición por grupo demográfico, calculando la exposición promedio por grupo. Este análisis permite identificar si ciertos grupos demográficos reciben sistemáticamente mayor o menor exposición, lo cual indicaría la presencia de sesgo demográfico. La comparación de exposición promedio entre grupos, junto con el análisis de significancia estadística, permite identificar sesgos sistemáticos que requieren atención.

4.2.2. Fase 2: Análisis de Diversidad de Opciones de Consumo

La segunda fase tiene como objetivo evaluar cómo el sesgo de exposición limita la diversidad de opciones de consumo disponibles para los usuarios, identificando restricciones sistemáticas y "burbujas de filtro"que pueden limitar el descubrimiento de nuevos productos, servicios o contenidos. Esta fase es crucial porque el sesgo de exposición no solo se manifiesta en términos de cantidad de impresiones, sino también en términos de diversidad de opciones expuestas. Un usuario puede recibir muchas impresiones pero todas del mismo tipo, lo cual limita su capacidad de descubrimiento tanto como recibir pocas impresiones.

Cálculo de Métricas de Diversidad de Exposición

Para cuantificar la diversidad de exposición de manera robusta, se implementan tres métricas complementarias que capturan diferentes aspectos de la diversidad.

La primera métrica es la Entropía de Shannon, que mide la incertidumbre o diversidad en la distribución de elementos expuestos. La fórmula es:

$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$

donde p_i es la proporción de impresiones del elemento i. La entropía toma valores entre 0 (sin diversidad, todas las impresiones son del mismo elemento) y $\log_2(n)$ (máxima diversidad, todos los elementos tienen la misma probabilidad), donde n es el número de elementos únicos disponibles. La entropía de Shannon es especialmente útil porque es sensible a la distribución completa de probabilidades, no solo a la

concentración en elementos específicos.
La segunda métrica es el Índice de Diversidad de Simpson, que mide la probabilidad de que dos elementos seleccionados al azar de las impresiones de un usuario sean diferentes. La fórmula es
$D = 1 - \sum_{i=1}^{n} p_i^2$
donde p_i es nuevamente la proporción de impresiones del elemento i.
El índice toma valores entre 0 (sin diversidad) y 1 (máxima diversidad). Esta métrica es intuitiva porque proporciona una interpretación probabilística directa: un valor de 0.5 significa que hay un 50 % de probabilidad de que dos impresiones aleatorias sean diferentes.
La tercera métrica es el Gini de Diversidad, que es una adaptación del coeficiente de Gini para medir la concentración en elementos específicos. La fórmula es
$G_{div} = \frac{2\sum_{i=1}^{n} i \cdot x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}$
donde x_i es el número de impresiones del elemento i ordenado de menor a mayor.
Esta métrica es complementaria a las otras dos porque captura específicamente la concentración, permitiendo identificar si la falta de diversidad se debe a concentración en pocos elementos populares.
La aplicación de estas métricas se realiza por usuario para cada dimensión relevante (posición de anuncio, historial de navegación, hora del día, tipo de dispositivo), permitiendo entender la diversidad en diferentes aspectos de la exposición. Luego se agregan las métricas promedio por usuario, y finalmente se analiza la distribución de diversidad en toda la población, identificando qué proporción de usuarios experimenta baja diversidad y qué proporción experimenta alta diversidad.

Análisis de Restricciones en Opciones de Consumo

El análisis de restricciones en opciones de consumo se enfoca en cuantificar qué proporción del total de opciones disponibles ha sido expuesta a cada usuario. Para esto, se calcula el Coverage Ratio, que es la proporción de elementos únicos vistos por usuario respecto al total disponible. Por dimensión, el coverage se calcula como
$\text{Coverage}_d = \frac{\text{Elementos únicos vistos}_d}{\text{Total elementos disponibles}_d}$
donde d representa una dimensión específica (por ejemplo, posición de anuncio, historial de navegación, etc.).
El coverage promedio por usuario se calcula como
$\text{Coverage}_{prom} = \frac{1}{D}\sum_{d=1}^{D} \text{Coverage}_d$
donde D es el número de dimensiones consideradas.
Este análisis permite identificar usuarios con bajo coverage, lo cual indica restricciones severas en su capacidad de descubrimiento. Se utilizan umbrales de 50 % y 25 % para identificar restricciones moderadas y severas respectivamente. Un usuario con

coverage menor al 50 % ve menos de la mitad de las opciones disponibles, lo cual limita significativamente su capacidad de tomar decisiones informadas. Un usuario con coverage menor al 25 % ve menos de un cuarto de las opciones, lo cual representa una restricción extrema.
Además, se realiza un análisis por grupo demográfico, comparando el coverage promedio entre grupos para detectar disparidades. Si ciertos grupos demográficos tienen sistemáticamente menor coverage que otros, esto indica que el sesgo de exposición no solo afecta la cantidad de impresiones, sino también la diversidad de opciones disponibles, creando desventajas acumulativas para ciertos grupos.

Identificación de Elementos Invisibles
Un aspecto crítico del análisis de diversidad es la identificación de elementos "invisibles", es decir, elementos que nunca son expuestos a ciertos grupos demográficos. Este análisis se realiza por grupo demográfico, identificando para cada grupo qué elementos del conjunto total nunca aparecen en sus impresiones. Esto es importante porque crea "zonas ciegas"donde ciertos grupos no tienen acceso a ciertas opciones, independientemente de su potencial interés o relevancia.
Para cuantificar este problema, se calcula el ratio de invisibilidad, que es la proporción de elementos no vistos por grupo. Un ratio alto indica que una gran proporción de opciones está completamente fuera del alcance de ese grupo. Además, se documenta específicamente qué elementos están excluidos para cada grupo, creando un mapeo detallado de las restricciones de acceso.
Esta documentación es crucial porque permite identificar patrones sistemáticos de exclusión. Si ciertos tipos de elementos están sistemáticamente ausentes para ciertos grupos, esto puede indicar sesgos en el algoritmo de selección que requieren atención específica. Por ejemplo, si productos dirigidos a ciertos grupos demográficos nunca aparecen para otros grupos, esto puede limitar el descubrimiento y crear desventajas en el acceso a opciones relevantes.

4.2.3. Fase 3: Validación de Técnicas de Mitigación de Sesgo

La tercera fase tiene como objetivo evaluar la efectividad de diferentes técnicas de mitigación de sesgo, comparando sus resultados con un modelo baseline y validando su capacidad para reducir el sesgo sin degradar significativamente el rendimiento predictivo. Esta fase es fundamental porque proporciona evidencia empírica sobre qué técnicas son más efectivas para abordar el sesgo de exposición, y qué trade-offs existen entre equidad y efectividad. La evaluación sistemática de múltiples técnicas permite identificar no solo cuál es más efectiva, sino también entender los mecanis-

mos mediante los cuales cada técnica opera y sus limitaciones.

Estrategias de Mitigación Implementadas

Se evalúan cuatro estrategias de mitigación, cada una con diferentes enfoques para abordar el problema de sesgo de exposición. La primera estrategia es el Baseline (Sin Mitigación), que consiste en un modelo predictivo estándar sin técnicas de mitigación. Este modelo establece métricas de referencia para comparación, permitiendo evaluar el impacto de las técnicas de mitigación. El baseline es crucial porque proporciona el punto de partida contra el cual se miden las mejoras en equidad y los cambios en rendimiento.

La segunda estrategia es Reweighting (Re-ponderación), que ajusta los pesos de las muestras durante el entrenamiento para balancear la representación de diferentes grupos demográficos. Esta técnica no modifica los datos originales, sino que ajusta la importancia relativa de diferentes muestras durante el proceso de aprendizaje. El objetivo es reducir disparidades en exposición sin alterar la estructura de los datos, lo cual es ventajoso porque preserva la información original mientras intenta corregir desbalances. Esta técnica es especialmente útil cuando el sesgo está relacionado con desbalances en la representación de grupos en el conjunto de entrenamiento.

La tercera estrategia es Resampling, que incluye dos variantes: Oversampling, que duplica muestras de grupos minoritarios para aumentar su representación, y Undersampling, que reduce muestras de grupos mayoritarios para balancear la distribución. El objetivo de ambas variantes es balancear la distribución de clases o grupos en el conjunto de entrenamiento. Oversampling puede ser preferible cuando hay pocos datos de grupos minoritarios y se quiere preservar toda la información disponible, mientras que undersampling puede ser preferible cuando hay suficientes datos y se quiere reducir el tamaño del conjunto de entrenamiento.

La cuarta estrategia es Threshold Optimization utilizando la biblioteca Fairlearn, que optimiza los umbrales de decisión por grupo demográfico después del entrenamiento del modelo. Esta técnica no modifica el modelo en sí, sino que ajusta los puntos de corte utilizados para tomar decisiones, permitiendo mejorar la equidad post-entrenamiento. Esta aproximación es especialmente útil cuando el modelo ya está entrenado y se quiere mejorar la equidad sin reentrenar, o cuando se quiere explorar diferentes balances entre equidad y rendimiento ajustando umbrales.

Metricas de evaluación

La evaluación de las técnicas de mitigación requiere métricas que capturen tanto la equidad como el rendimiento del modelo. Las métricas de equidad incluyen el Gini de Exposición, que mide la desigualdad en la distribución de impresiones entre

usuarios, con un objetivo de reducir el valor a menos de 0.3 (indicando desigualdad moderada o baja). La Demographic Parity Difference mide la diferencia en tasas de exposición entre grupos demográficos, con un objetivo de mantener la diferencia por debajo de 0.05 (indicando que los grupos tienen probabilidades similares de exposición). La Equal Opportunity Difference mide la diferencia en tasas de verdadero positivo entre grupos, capturando si diferentes grupos tienen diferentes probabilidades de recibir impresiones cuando deberían recibirlas según sus características.
Las métricas de rendimiento incluyen el Global AUC (Área bajo la curva ROC), que mide la capacidad del modelo para distinguir entre usuarios que harán click y usuarios que no lo harán, con un objetivo de no degradar más del 5 % respecto al baseline. El AUC por grupo mide el rendimiento del modelo para cada grupo demográfico individualmente, permitiendo identificar si ciertos grupos tienen peor rendimiento que otros. El Global LogLoss mide la calibración del modelo, es decir, qué tan bien las probabilidades predichas reflejan las probabilidades reales. Finalmente, la desviación estándar de AUC entre grupos (Std AUC por grupo) mide la variabilidad en rendimiento entre grupos, donde valores menores indican mayor equidad en rendimiento.

Proceso de evaluación

El proceso de validación sigue una secuencia estructurada para asegurar comparabilidad entre técnicas. Primero, se realiza una división de datos en conjuntos de entrenamiento y prueba, típicamente utilizando una proporción de 85 % para entrenamiento y 15 % para prueba. Esta división se realiza de manera estratificada para asegurar que la distribución de grupos demográficos sea similar en ambos conjuntos. Segundo, se aplica cada estrategia de mitigación al conjunto de entrenamiento, entrenando modelos con cada técnica. Tercero, se evalúan los modelos entrenados en el conjunto de prueba, calculando todas las métricas de equidad y rendimiento. Cuarto, se realiza un análisis comparativo con el baseline, identificando qué técnicas mejoran las métricas de equidad y qué impacto tienen en las métricas de rendimiento. Finalmente, se realiza un análisis de trade-offs, evaluando el balance entre equidad y efectividad para cada técnica y identificando cuál ofrece el mejor balance.

Criterios de éxito

Los criterios de éxito están diseñados para asegurar que las técnicas de mitigación logren mejoras significativas en equidad sin degradar excesivamente el rendimiento. El primer criterio es la reducción en Gini de exposición a menos de 0.3, lo que representa una reducción significativa en desigualdad. El segundo criterio es la mejora en demographic parity, manteniendo la diferencia por debajo de 0.05, lo que indica que

los grupos tienen probabilidades similares de exposición. El tercer criterio es el mantenimiento del AUC global sin degradar más del 5 %, asegurando que las mejoras en equidad no vengan a costa de una pérdida significativa en capacidad predictiva. El cuarto criterio es la mejora en AUC de grupos minoritarios, asegurando que las técnicas de mitigación no solo no degraden el rendimiento, sino que lo mejoren para grupos que pueden estar subrepresentados.

4.2.4. Fase 4: Análisis de Efectividad de Campañas de Marketing Digital

La cuarta fase tiene como objetivo evaluar el impacto del sesgo de exposición en la efectividad de las campañas de marketing digital, analizando métricas de negocio y trade-offs entre equidad y efectividad. Esta fase es crucial porque responde a una pregunta fundamental: £el sesgo de exposición solo afecta la equidad, o también afecta la efectividad de las campañas? Si el sesgo limita la efectividad, entonces mejorar la equidad puede ser beneficioso tanto desde una perspectiva ética como desde una perspectiva de negocio, creando una oportunidad de "win-win".

Análisis de Impacto del Sesgo en Métricas de Campañas

Para evaluar el impacto del sesgo en métricas de campañas, se realiza una segmentación de usuarios por nivel de exposición. Los usuarios se dividen en quintiles según su exposición total (Q1: bajo, Q2: bajo-medio, Q3: medio, Q4: medio-alto, Q5: alto), permitiendo comparar métricas de efectividad entre usuarios con diferentes niveles de exposición. Esta segmentación es importante porque permite identificar si usuarios con mayor exposición tienen sistemáticamente mayor efectividad, lo cual sugeriría que el sesgo de exposición está limitando la efectividad general de las campañas.
Para cada quintil, se calculan múltiples métricas de efectividad. El CTR (Click-Through Rate) se calcula como

$$\text{CTR} = \frac{\text{Clicks}}{\text{Impresiones}}$$

proporcionando la proporción de impresiones que resultan en clicks. La tasa de conversión mide la proporción de clicks que resultan en conversión, capturando no solo la capacidad de generar interés sino también la capacidad de generar acciones. El CTR por usuario es el promedio de CTR individual, permitiendo entender la efectividad promedio para usuarios en cada nivel de exposición. Finalmente, la diversidad promedio medida como Shannon Entropy promedio por quintil permite entender si

la diversidad de exposición está relacionada con la efectividad.
El análisis de correlación complementa la segmentación por quintiles, proporcionando una medida cuantitativa de la relación entre exposición, diversidad y efectividad. Se calcula la correlación entre exposición y CTR a nivel de usuario, permitiendo entender si existe una relación lineal entre mayor exposición y mayor efectividad. Se calcula también la correlación entre diversidad y CTR, permitiendo entender si la diversidad de exposición está relacionada con la efectividad. Además, se utilizan visualizaciones para identificar relaciones no lineales que podrían no ser capturadas por correlaciones lineales, como relaciones cuadráticas, logarítmicas, o con puntos de inflexión.

Análisis de Trade-offs entre Equidad y Efectividad
El análisis de trade-offs entre equidad y efectividad es fundamental para entender si mejorar la equidad requiere sacrificar métricas de negocio, o si es posible lograr ambos objetivos simultáneamente. Para esto, se comparan dos escenarios: el Escenario Actual, que representa las métricas observadas en los datos reales, y el Escenario Equitativo (Simulado), que representa una distribución uniforme de exposición donde todos los usuarios reciben el mismo número de impresiones.
Para cada escenario, se calculan métricas de comparación que incluyen el Gini de exposición (que debería ser 0 en el escenario equitativo), el CTR promedio, y el total de clicks. Luego se calculan mejoras relativas para cuantificar el impacto de mejorar la equidad. La mejora en equidad se calcula como

$$\frac{G_{actual} - G_{equitativo}}{G_{actual}} \times 100\,\%$$

proporcionando el porcentaje de reducción en desigualdad. El cambio en CTR se calcula como

$$\frac{CTR_{equitativo} - CTR_{actual}}{CTR_{actual}} \times 100\,\%$$

proporcionando el porcentaje de cambio en efectividad.
Además, se realiza un análisis de trade-offs de las técnicas de mitigación implementadas en la Fase 3, evaluando para cada técnica qué mejoras en equidad se logran y qué cambios en efectividad se observan. Se calcula un ratio de trade-off como $\frac{\text{Mejora en equidad}}{\text{Cambio en efectividad}}$, que permite comparar técnicas según su eficiencia en lograr mejoras en equidad por unidad de cambio en efectividad. Un ratio alto indica que se logra mucha mejora en equidad con poco cambio en efectividad, mientras que un ratio bajo indica que se requiere mucho cambio en efectividad para lograr mejoras en equidad.

Análisis por Tipo de Campaña

El análisis por tipo de campaña permite entender cómo diferentes estrategias de marketing contribuyen al sesgo de exposición y cómo varían en efectividad. Se utiliza una clasificación heurística basada en características de los usuarios y su historial de interacción. Las campañas de Awareness se dirigen a usuarios con baja exposición previa ($\leq$ 1 impresión), con el objetivo de aumentar la visibilidad inicial. Las campañas de Retargeting se dirigen a usuarios con alta exposición (>5 impresiones) y historial de clicks, con el objetivo de reforzar el interés existente. Las campañas de Conversión se dirigen a usuarios con exposición media (>1 impresión) y clicks, con el objetivo de convertir interés en acción. El resto de casos se clasifica como General. Para cada tipo de campaña, se calculan métricas específicas que incluyen el CTR por tipo de campaña (permitiendo comparar efectividad entre estrategias), el Gini de exposición por tipo (permitiendo identificar qué tipos de campaña contribuyen más al sesgo), el volumen de impresiones por tipo (permitiendo entender la distribución de recursos entre estrategias), y un análisis de efectividad relativa (permitiendo identificar qué tipos de campaña ofrecen mejor relación entre recursos invertidos y resultados obtenidos).

Este análisis es importante porque diferentes tipos de campaña pueden tener diferentes impactos en el sesgo de exposición. Por ejemplo, las campañas de retargeting, al estar dirigidas a usuarios con alta exposición previa, pueden estar exacerbando el sesgo de exposición, mientras que las campañas de awareness, al estar dirigidas a usuarios con baja exposición previa, pueden estar contribuyendo a reducir el sesgo. Entender estos impactos permite hacer recomendaciones sobre cómo balancear los tipos de campaña para lograr tanto efectividad como equidad.

4.2.5. Fase 5: Taxonomía de Tipos de Sesgo de Exposición

La quinta fase tiene como objetivo crear una taxonomía sistemática de los diferentes tipos de sesgo de exposición, caracterizarlos cuantitativamente y analizar sus interacciones. Esta fase es fundamental porque el sesgo de exposición no es un fenómeno monolítico, sino que puede manifestarse de múltiples maneras diferentes, cada una con diferentes mecanismos, diferentes impactos, y potencialmente diferentes soluciones. Una taxonomía sistemática permite entender la complejidad del problema, identificar qué tipos de sesgo son más severos, y desarrollar estrategias de mitigación específicas para cada tipo.

Identificación y Caracterización de Tipos de Sesgo

Se identifican y caracterizan cinco tipos principales de sesgo de exposición, cada

uno con diferentes mecanismos y métricas de cuantificación. El primer tipo es el Sesgo de Selección (Selection Bias), que se define como diferencias en probabilidad de exposición basadas en características del usuario. Este tipo de sesgo se cuantifica mediante la diferencia relativa en probabilidad de exposición entre grupos, calculada como Selection Bias = $\frac{P_{max}-P_{min}}{P_{max}}$, donde P es la probabilidad de exposición por grupo. Esta métrica captura si ciertos grupos tienen sistemáticamente mayor o menor probabilidad de ser expuestos, independientemente de su potencial interés o relevancia.

El segundo tipo es el Sesgo de Popularidad (Popularity Bias), que se define como la concentración de impresiones en elementos "populares", es decir, elementos que ya tienen alta exposición o interacción previa. Este tipo de sesgo se cuantifica mediante el coeficiente de Gini de la distribución de impresiones por elemento, calculado para cada dimensión (posición, historial, hora, dispositivo) y promediado como métrica agregada. Un Gini alto indica que las impresiones están concentradas en pocos elementos populares, mientras que un Gini bajo indica una distribución más uniforme.

El tercer tipo es el Sesgo de Posición (Position Bias), que se define como diferencias en CTR o exposición según la posición del anuncio en la página o lista de resultados. Este tipo de sesgo se cuantifica mediante la diferencia relativa en CTR entre posiciones, calculada como

$$\text{Position Bias} = \frac{CTR_{max} - CTR_{min}}{CTR_{max}}$$

Esta métrica captura si anuncios en ciertas posiciones (por ejemplo, parte superior de la página) tienen sistemáticamente mayor CTR que anuncios en otras posiciones, lo cual puede crear desventajas para anuncios en posiciones menos prominentes.

El cuarto tipo es el Sesgo Demográfico (Demographic Bias), que se define como diferencias en exposición por características demográficas. Este tipo de sesgo se cuantifica mediante múltiples métricas: la exposición promedio por usuario por grupo (permitiendo comparar exposición entre grupos), el Gini de exposición dentro de cada grupo (permitiendo identificar desigualdad dentro de grupos), y el ratio de exposición calculado como

$$\frac{E_{max} - E_{min}}{E_{max}}$$

(permitiendo cuantificar la diferencia relativa entre grupos con mayor y menor exposición).

El quinto tipo es el Sesgo de Confirmación (Confirmation Bias), que se define como la tendencia del sistema a mostrar solo elementos similares a las preferencias previas del usuario, creando "burbujas de filtro". Este tipo de sesgo se cuantifica como el inverso de la diversidad de elementos vistos, calculado como Confirmation Bias =

1 − Diversidad Ratio, donde el Diversidad Ratio es la proporción de elementos únicos vistos sobre el total de elementos disponibles. Un valor alto indica que los usuarios ven principalmente elementos similares a sus preferencias previas, mientras que un valor bajo indica mayor diversidad.

Análisis de Interacciones entre Tipos de Sesgo

Un aspecto crítico del análisis es entender cómo diferentes tipos de sesgo interactúan entre sí, ya que los sesgos no actúan de forma independiente. Para esto, se construye una matriz de correlación que calcula las correlaciones entre métricas de diferentes tipos de sesgo. Esta matriz permite identificar qué tipos de sesgo co-ocurren sistemáticamente, lo cual es importante porque ciertos tipos de sesgo pueden estar relacionados causalmente o pueden ser manifestaciones diferentes del mismo mecanismo subyacente.

Además, se realiza un análisis de grupos demográficos más afectados por combinaciones de sesgos, permitiendo identificar si ciertos grupos experimentan múltiples tipos de sesgo simultáneamente. Esto es crucial porque el impacto combinado de múltiples tipos de sesgo puede ser mayor que la suma de los impactos individuales, creando desventajas acumulativas para ciertos grupos.

El análisis de severidad complementa el análisis de interacciones, proporcionando un ranking de tipos de sesgo por valor de métrica. Este ranking permite identificar qué tipos de sesgo son más problemáticos y dónde priorizar los esfuerzos de mitigación. Además, se realiza un análisis de distribución de severidad por grupo, permitiendo entender si ciertos grupos experimentan niveles particularmente altos de ciertos tipos de sesgo, lo cual puede requerir estrategias de mitigación específicas para esos grupos.

4.3. Correspondencia entre Fases Metodológicas y Componentes del Repositorio

El Anexo: Repositorio del Proyecto documenta la implementación técnica completa de esta metodología de investigación, proporcionando el código fuente, scripts de procesamiento, notebooks de análisis y todos los resultados generados. El repositorio público (https://github.com/martuG/TF-Repo) constituye el componente operacional que materializa cada fase metodológica descrita anteriormente, garantizando la reproducibilidad, transparencia y verificabilidad de todos los análisis realizados.

Cada fase de la metodología tiene una correspondencia directa con componentes

específicos del repositorio, permitiendo que investigadores y profesionales puedan replicar exactamente los experimentos y análisis descritos en este documento.
Fase 1 (Exploración y Caracterización) se implementa mediante:

- El notebook 01_exploracion_dataset.ipynb, que contiene todo el análisis descriptivo inicial, cálculo de estadísticas, visualizaciones de distribución y cálculo de métricas de desigualdad (coeficiente de Gini, exposición por grupo).
- El módulo src/data_preparation/, que implementa el pipeline completo de carga, limpieza y preparación de datos descrito en la metodología.
- Los scripts de verificación de integridad (scripts/generate_checksum.ps1 y generate_checksum.sh) que aseguran la reproducibilidad de los datos de entrada.

Fase 2 (Análisis de Diversidad) se implementa mediante:

- El notebook 02_analisis_diversidad_consumo.ipynb, que contiene el cálculo de métricas de diversidad (Entropía de Shannon, Índice de Simpson, Gini de Diversidad), análisis de restricciones en opciones de consumo (Coverage Ratio) e identificación de elementos ïnvisibles"por grupo demográfico.
- El módulo src/fairness/exposure_metrics.py, que implementa las métricas de diversidad y exposición descritas en la metodología.

Fase 3 (Validación de Mitigación) se implementa mediante:

- El notebook 03_validacion_mitigacion.ipynb, que contiene la evaluación comparativa de las cuatro estrategias de mitigación (Baseline, Reweighting, Resampling, Threshold Optimization) y el cálculo de todas las métricas de equidad y rendimiento.
- El módulo src/fairness/mitigation_strategies.py, que implementa las técnicas de mitigación descritas en la metodología (reweighting, resampling, threshold optimization).
- El módulo src/modeling/train_model.py y evaluate_model.py, que implementan el entrenamiento y evaluación de modelos con diferentes estrategias de mitigación.

Fase 4 (Efectividad de Campañas)se implementa mediante:

- El notebook 04_efectividad_campanas.ipynb, que contiene el análisis de impacto del sesgo en métricas de campañas, análisis de trade-offs entre equidad y efectividad, y análisis por tipo de campaña.
- El módulo src/fairness/simulate_exposure.py, que permite simular escenarios equitativos y compararlos con el escenario actual.

Fase 5 (Taxonomía de Tipos de Sesgo) se implementa mediante:

- El notebook 05_taxonomia_tipos_sesgo.ipynb, que contiene la identificación y caracterización de los cinco tipos de sesgo, cálculo de métricas de cuantificación, análisis de interacciones entre tipos de sesgo y análisis de severidad.
- El módulo src/fairness/bias_detection.py, que implementa los algoritmos de detección y cuantificación de diferentes tipos de sesgo.

4.4. Relación entre objetivos y fases metodológicas

A continuación se expone el cuadro 4.1 el cual refleja la relación entre los objetivos planteados en la investigación, la fase metodológica que lo aborda y cuales son los resultados que se esperan.

Cuadro 4.1: Matriz de Objetivos, Fases y Resultados Esperados

Objetivo Específico	**Fase Metodológica**	**Resultados Esperados**
Identificar y caracterizar tipos de sesgo	Fase 5: Taxonomía de tipos de sesgo	▪ Taxonomía sistemática de 5 tipos de sesgo ▪ Métricas cuantitativas por tipo ▪ Matriz de correlaciones entre tipos
Cuantificar sesgo mediante métricas	Fase 1: Exploración y caracterización Fase 2: Diversidad de consumo	▪ Coeficiente de Gini de exposición ▪ Demographic Parity Difference ▪ Equal Opportunity Difference ▪ Shannon Entropy y Coverage Ratio
Evaluar impacto en diversidad de consumo	Fase 2: Diversidad de consumo	▪ Coverage ratios por usuario y grupo ▪ Elementos invisibles por grupo ▪ Restricciones sistemáticas identificadas
Validar técnicas de mitigación	Fase 3: Validación de mitigación	▪ Ranking de estrategias por efectividad ▪ Trade-offs equidad vs rendimiento ▪ Recomendaciones sobre técnicas más efectivas
Analizar trade-offs equidad-efectividad	Fase 4: Efectividad de campañas	▪ Correlaciones exposición-efectividad ▪ Trade-offs cuantificados ▪ Análisis por tipo de campaña

Capítulo 5

Análisis de datos

Se presenta un análisis exhaustivo de los resultados obtenidos del estudio del sesgo de exposición en modelos predictivos aplicados al marketing digital. El trabajo se estructurado en cinco fases metodológicas que abordan de manera sistemática desde la caracterización inicial del dataset hasta la evaluación comparativa de técnicas de mitigación y su impacto en la efectividad de campañas publicitarias.
El análisis revela patrones significativos de desigualdad en la distribución de oportunidades de exposición, restricciones severas en la diversidad de opciones de consumo, y la identificación de múltiples tipos de sesgo que operan de manera simultánea. Los hallazgos sugieren que el sesgo de confirmación representa el problema más crítico, limitando el descubrimiento de nuevas opciones para la mayoría de los usuarios. Sin embargo, el análisis también identifica oportunidades prometedoras: existe una correlación fuerte entre exposición y efectividad, lo que sugiere que mejorar la equidad puede simultáneamente mejorar las métricas de negocio.

5.1. Resultados de la Fase 1

El dataset analizado comprende un total de 10,000 impresiones publicitarias distribuidas entre 4,000 usuarios únicos, lo que establece una relación promedio de 2.5 impresiones por usuario. La estructura de datos incluye 11 variables que capturan tanto información demográfica (edad, género) como características contextuales de las impresiones (tipo de dispositivo, posición del anuncio, historial de navegación, hora del día) y variables de resultado (click, exposición por usuario).
La calidad de los datos es excelente, sin presencia de valores nulos en ninguna de las variables analizadas. Los tipos de datos son apropiados para cada variable, con varia-

bles categóricas correctamente tipadas y variables numéricas en formatos adecuados. La integridad referencial entre usuarios e impresiones ha sido verificada, asegurando la consistencia del dataset para los análisis subsiguientes.

5.1.1. Distribución de exposición por usuario

El análisis de la distribución de exposición revela patrones de desigualdad significativos. La media de impresiones por usuario es de 2.50, pero esta cifra oculta una distribución altamente asimétrica. La mediana se sitúa en apenas 1.00 impresiones por usuario, lo que indica que más de la mitad de los usuarios recibe una única impresión. La desviación estándar de 4.17, considerablemente mayor que la media, confirma la alta variabilidad en la distribución. Estas medidas pueden observarse gráficamente en la figura 5.1.

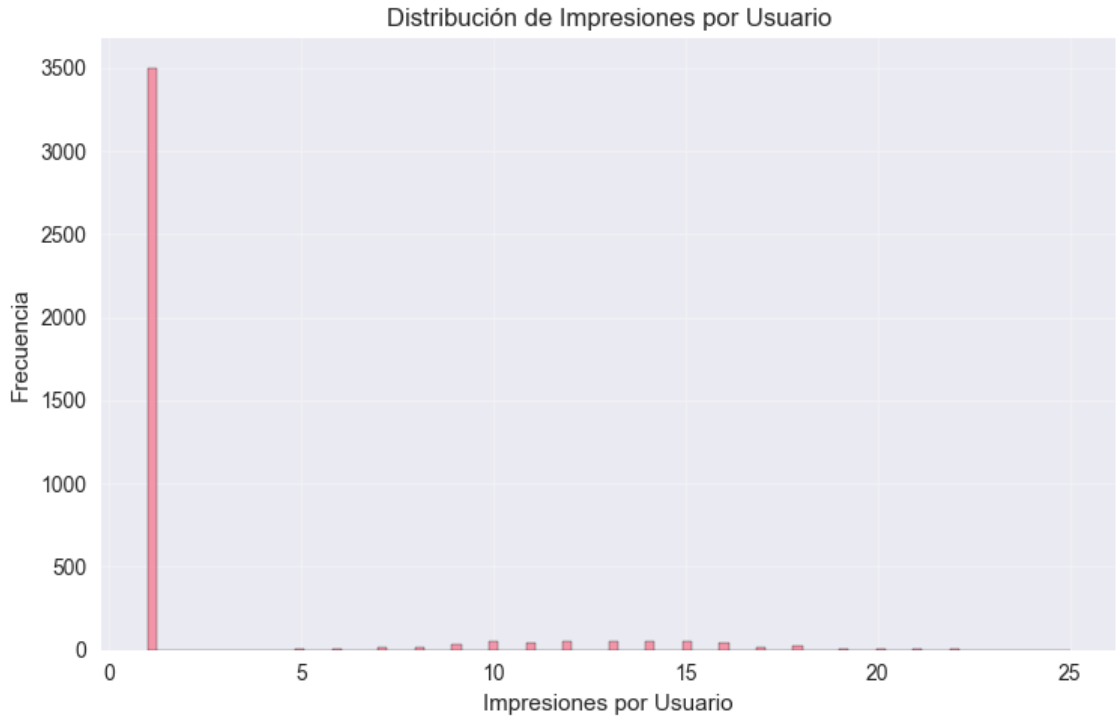

Figura 5.1: Distribución de Impresiones por Usuario

El rango de exposición va desde 1 hasta 25 impresiones por usuario, pero la distribución de cuartiles revela que los tres primeros cuartiles (Q1, Q2 y Q3) se concentran en el valor mínimo de 1 impresión, mientras que el cuartil superior alcanza hasta 25 impresiones. Esta estructura indica que el 75 % de los usuarios recibe únicamente una impresión, mientras que un pequeño grupo de usuarios concentra un número

desproporcionadamente alto de oportunidades de exposición.
Esta asimetría extrema en la distribución sugiere la presencia de mecanismos de selección que favorecen sistemáticamente a ciertos usuarios sobre otros, creando una estructura de oportunidades de exposición que no se distribuye equitativamente. Esta desigualdad puede tener implicaciones tanto éticas como operativas, ya que limita las oportunidades de interacción para la mayoría de los usuarios mientras concentra recursos en un subconjunto reducido.

5.1.2. Análisis de CTR por Grupos Demográficos

El análisis del Click-Through Rate (CTR) por grupos demográficos proporciona perspectivas importantes sobre cómo diferentes segmentos de usuarios responden a las impresiones publicitarias. Este análisis es crucial para entender si las diferencias en efectividad están relacionadas con características demográficas o si, por el contrario, el sesgo se manifiesta principalmente a través de diferencias en la distribución de oportunidades de exposición.
En la figura 5.2 se visualizan dos gráficos: a la izquierda la relación entre CTR Y edades. A la derecha, el gráfico que muestra la relación entre edad y numero de impresiones.

Por Edad

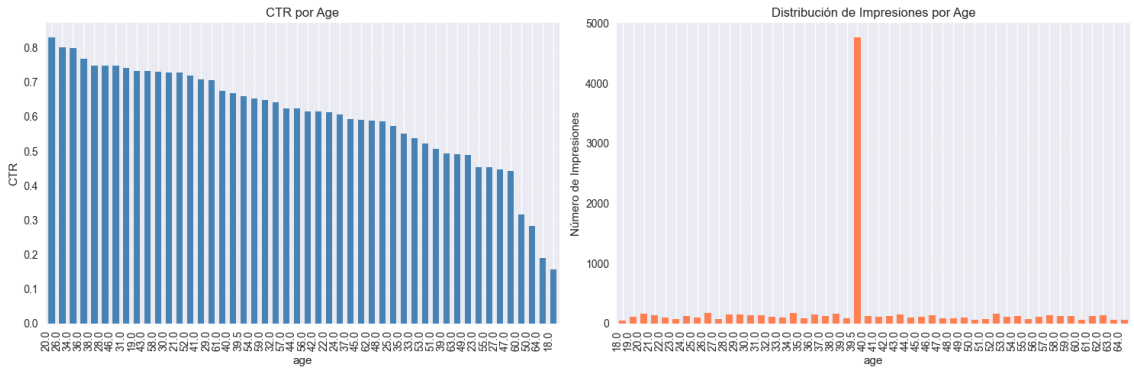

Figura 5.2: Distribuciones por edad

El análisis por edad revela variaciones sustanciales en el CTR. El rango de CTR oscila entre 0.157 para usuarios de 18 años y 0.830 para usuarios de 20 años, mostrando una variabilidad considerable. Se observa un patrón claro donde usuarios más jóvenes, específicamente aquellos en el rango de 18 a 26 años, exhiben CTR significativamente más altos. Los picos más pronunciados se encuentran en las edades de

20 y 26 años, sugiriendo que estos grupos etarios pueden tener mayor receptividad a los anuncios o que los anuncios están mejor calibrados para estos segmentos.
En contraste, los usuarios mayores, particularmente aquellos de 50 años o más, muestran CTR considerablemente más bajos, con valores que oscilan entre 0.188 y 0.315. Esta diferencia puede reflejar tanto diferencias en preferencias y comportamientos de consumo como posibles desajustes en el targeting o en el contenido de los anuncios para estos grupos etarios. La figura 5.3 representa gráficamente las métricas obtenidas, las cuales relacionan el genero y género.

Por Género

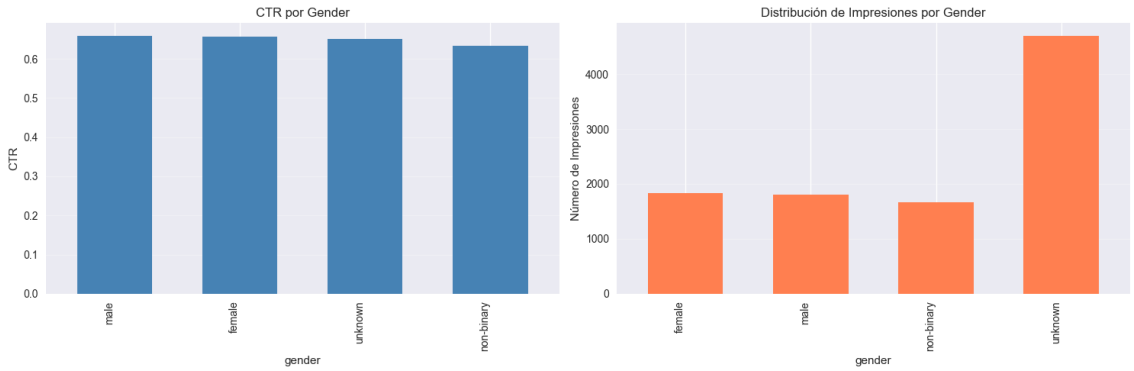

Figura 5.3: Distribuciones por género

El análisis por género muestra diferencias mínimas en el CTR. Los valores observados son 0.658 para usuarios masculinos, 0.658 para usuarios femeninos, 0.650 para usuarios con género desconocido, y 0.633 para usuarios no binarios. La variación máxima entre grupos es de apenas 0.025 puntos, lo que representa una diferencia relativa de menos del 4 %.
Esta relativa uniformidad en el CTR por género sugiere que el sesgo no se manifiesta principalmente a través de diferencias en la efectividad de los anuncios según género, sino más probablemente a través de diferencias en la distribución de oportunidades de exposición.
El gráfico de distribución de impresiones evidencia una asignación marcadamente desigual entre los distintos géneros, con una concentración sustancial en el grupo clasificado como unknown, que recibe más del doble de impresiones respecto de los demás segmentos. Mientras los grupos female, male y non-binary presentan volúmenes relativamente similares, la sobrerrepresentación del grupo sin identificación de género sugiere que el sistema de entrega de anuncios prioriza a usuarios cuyo género no está determinado. Esta asimetría indica que el sesgo potencial no se manifiesta

en la efectividad del anuncio, dado que los CTR son homogéneos, sino en la distribución de oportunidades de exposición, lo cual podría derivar de limitaciones en la calidad de los datos, decisiones del algoritmo de optimización o una composición particular de la base de usuarios.

Por Tipo de Dispositivo

La figura 5.4 expone dos gráficos: el primero muestra la relación entre CTR y tipo de dispositivo. Luego, el segundo grafica la relación entre tipo de dispositivo y número de impresiones.

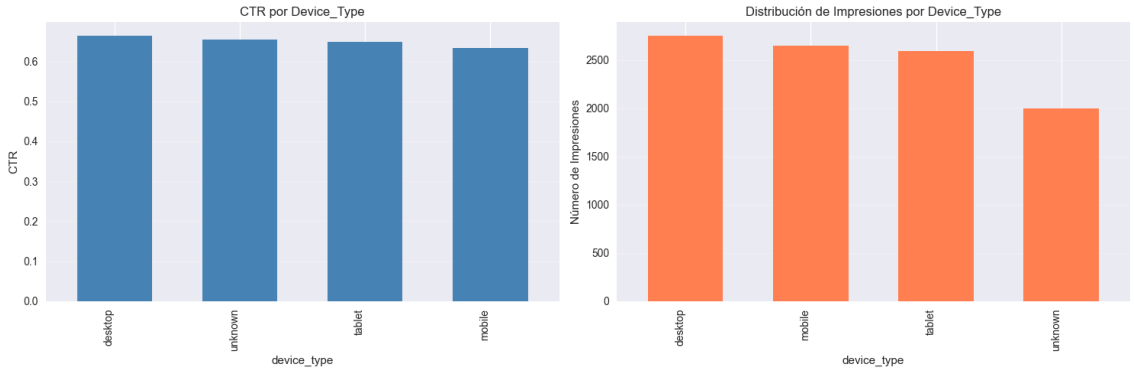

Figura 5.4: Distribuciones por tipo de dispositivo

El análisis por tipo de dispositivo muestra diferencias moderadas en el CTR. Los usuarios de desktop exhiben el CTR más alto con 0.664, seguidos por usuarios con dispositivo desconocido (0.655), tablet (0.648), y finalmente mobile (0.633). La variación entre el dispositivo con mayor y menor CTR es de 0.031 puntos, lo que representa una diferencia relativa del 4.9 %.

Estas diferencias son mayores que las observadas por género pero menores que las observadas por edad. Las diferencias pueden reflejar tanto características inherentes de cada plataforma (tamaño de pantalla, contexto de uso, capacidad de atención) como diferencias en cómo se optimizan los anuncios para cada tipo de dispositivo.

5.1.3. Metricas de desigualdad

Para cuantificar la desigualdad en la distribución de exposición, se calculó el coeficiente de Gini, una métrica ampliamente utilizada en economía y ciencias sociales para medir desigualdad. El coeficiente de Gini de exposición obtenido es de 0.4037, lo que indica una desigualdad moderada-alta en la distribución de impresiones entre usuarios.

El coeficiente de Gini toma valores entre 0 y 1, donde 0 representa igualdad perfecta (todos los usuarios reciben exactamente el mismo número de impresiones) y 1 representa desigualdad máxima (un único usuario recibe todas las impresiones). Un valor de 0.40 indica que aproximadamente el 40 % de la desigualdad máxima teórica está presente en el sistema. Para contextualizar, valores de Gini superiores a 0.3 generalmente se consideran indicativos de desigualdad significativa, y valores superiores a 0.4 sugieren desigualdad alta.
Este nivel de desigualdad tiene implicaciones importantes. En contextos de marketing digital, una distribución desigual de oportunidades de exposición puede limitar la capacidad de los usuarios para descubrir nuevos productos o servicios, puede crear desventajas competitivas para ciertos grupos demográficos, y puede reducir la efectividad general de las campañas al no aprovechar todo el potencial de la base de usuarios.

5.1.4. Exposición Promedio por Grupo Demográfico

El análisis de exposición promedio por género revela diferencias sistemáticas, aunque relativamente pequeñas en términos absolutos. Los usuarios femeninos reciben en promedio 2.465 impresiones, los usuarios masculinos 2.459 impresiones, los usuarios no binarios 2.332 impresiones, y los usuarios con género desconocido 2.220 impresiones. La figura 5.5 expone un gráfico que muestra las métricas mencionadas para la exposición promedio por grupo demográfico.

Aunque las diferencias absolutas son pequeñas (la diferencia máxima es de 0.245 impresiones por usuario), estas diferencias son sistemáticas y consistentes. Específicamente, los grupos "unknown"y "non-binary"reciben sistemáticamente menor exposición promedio que los grupos "male"y "female". En términos relativos, esta diferencia representa aproximadamente un 11 % menos de exposición para el grupo con menor exposición comparado con el grupo con mayor exposición.
Estas diferencias sistemáticas, aunque modestas, pueden tener efectos acumulativos a lo largo del tiempo. Si estos patrones se mantienen consistentemente, los usuarios en grupos con menor exposición pueden experimentar desventajas progresivas en términos de descubrimiento de productos, oportunidades de interacción, y potencialmente en la calidad de las recomendaciones que reciben del sistema.

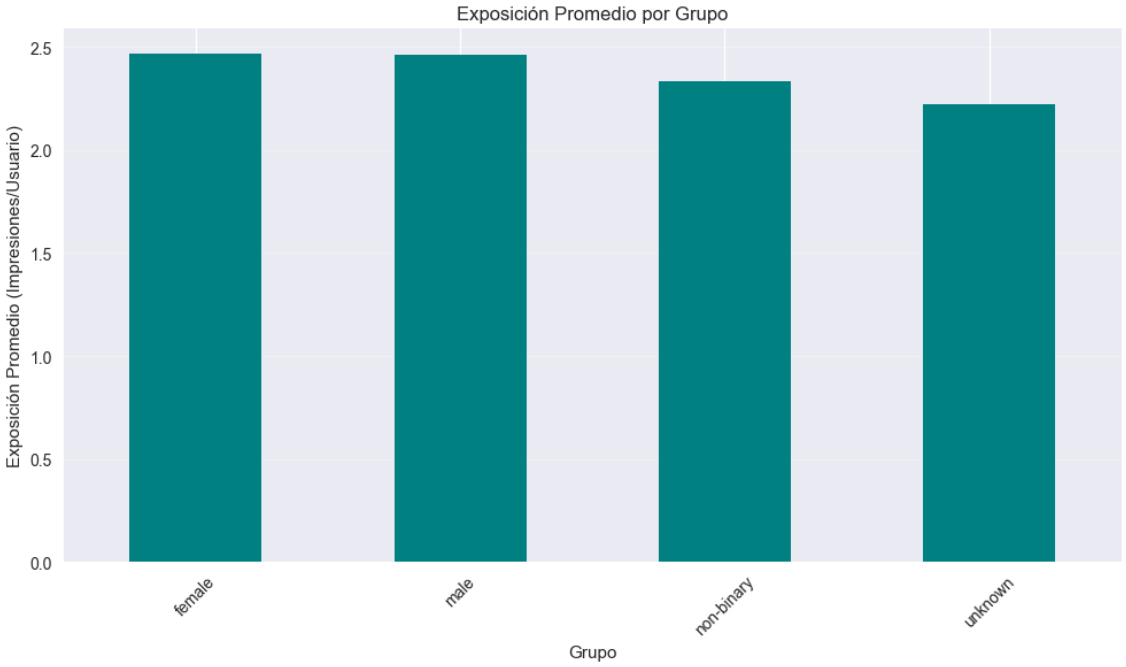

Figura 5.5: Exposición promedio por género

5.2. Resultados de la Fase 2

El análisis de diversidad de opciones de consumo es fundamental para entender si los usuarios tienen acceso a una variedad suficiente de productos, servicios o contenidos, o si por el contrario están siendo limitados a un conjunto restringido de opciones. Esta restricción puede manifestarse como "burbujas de filtro"donde los algoritmos de recomendación o selección limitan la exposición a opciones similares a las preferencias previas del usuario, reduciendo el descubrimiento de nuevas opciones.

5.2.1. Métricas de Diversidad de Exposición

Para cuantificar la diversidad de exposición, se utilizaron dos métricas complementarias: la entropía de Shannon, que mide la distribución de probabilidad de exposición a diferentes tipos de elementos, y el índice de diversidad de Simpson, que mide la probabilidad de que dos impresiones aleatorias pertenezcan a diferentes tipos de elementos.

Shannon Entropy Promedio

La entropía de Shannon promedio es de 0.075, con una desviación estándar de 0.205. El rango va desde valores prácticamente cero ($-1{,}44x10^{-10}$) hasta un máximo de

0.945. La mediana se sitúa prácticamente en cero, lo que es altamente revelador. La entropía de Shannon mide la incertidumbre o diversidad en una distribución. Un valor cercano a cero indica que la distribución es altamente concentrada (el usuario ve principalmente un tipo de elemento), mientras que valores más altos indican mayor diversidad. El hecho de que la mediana sea prácticamente cero significa que más del 50 % de los usuarios tiene una entropía extremadamente baja, lo que indica que estos usuarios están siendo expuestos casi exclusivamente a un solo tipo de elemento. La figura 5.6 muestra un gráfico que muestra la entropía de Shannon.

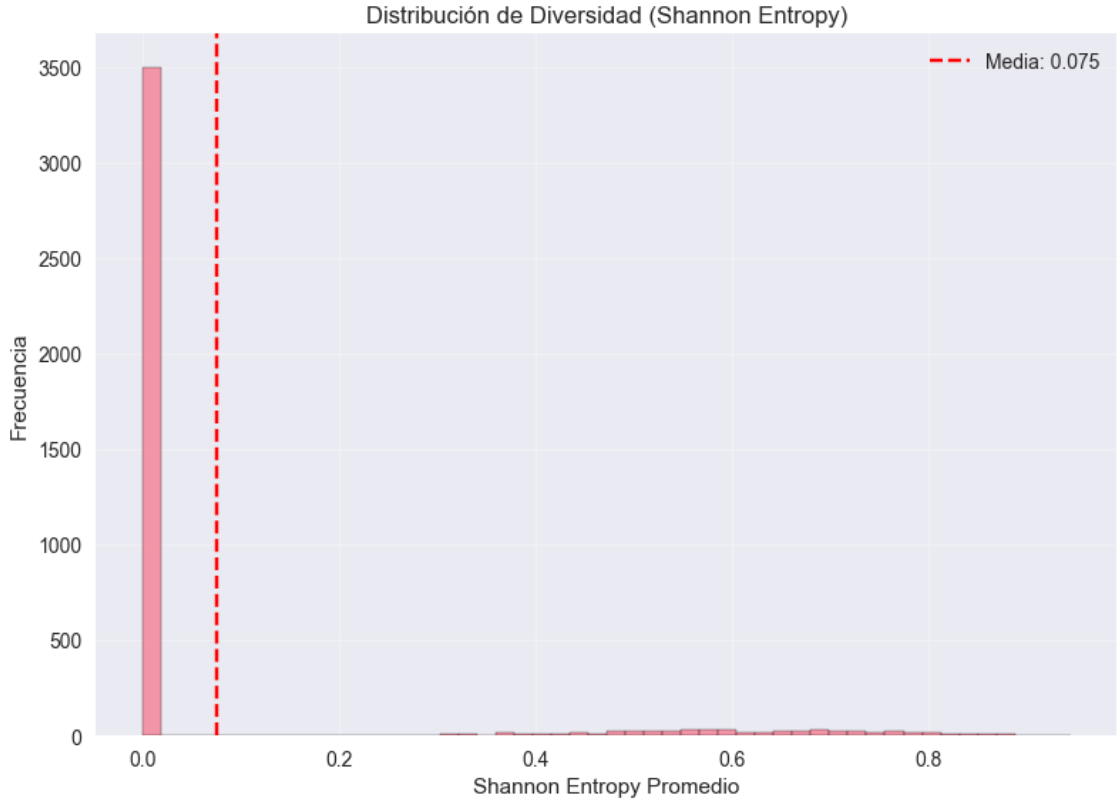

Figura 5.6: Shannon Entropy

La distribución muestra que la mayoría de usuarios (aproximadamente el 75 %) tiene entropía cercana a cero, indicando exposición predominantemente a un solo tipo de elemento. Solo una minoría de usuarios experimenta diversidad significativa, con algunos usuarios alcanzando valores de entropía cercanos a 1. Esta distribución altamente sesgada sugiere la presencia de "burbujas de filtro"para la mayoría de la población, donde los algoritmos de selección limitan sistemáticamente la exposición a opciones similares a las preferencias previas del usuario.

Simpson Diversity Promedio

El índice de diversidad de Simpson promedio es de 0.034, con una desviación estándar de 0.092. El rango va desde 0.000 hasta 0.463, y la mediana es exactamente 0.000. El índice de Simpson mide la probabilidad de que dos impresiones aleatorias de un usuario pertenezcan a diferentes tipos de elementos. Un valor de 0 indica que todas las impresiones son del mismo tipo, mientras que valores más altos indican mayor diversidad. La figura 5.7 muestra los valores obtenidos para la distribución de diversidad de Simpson.

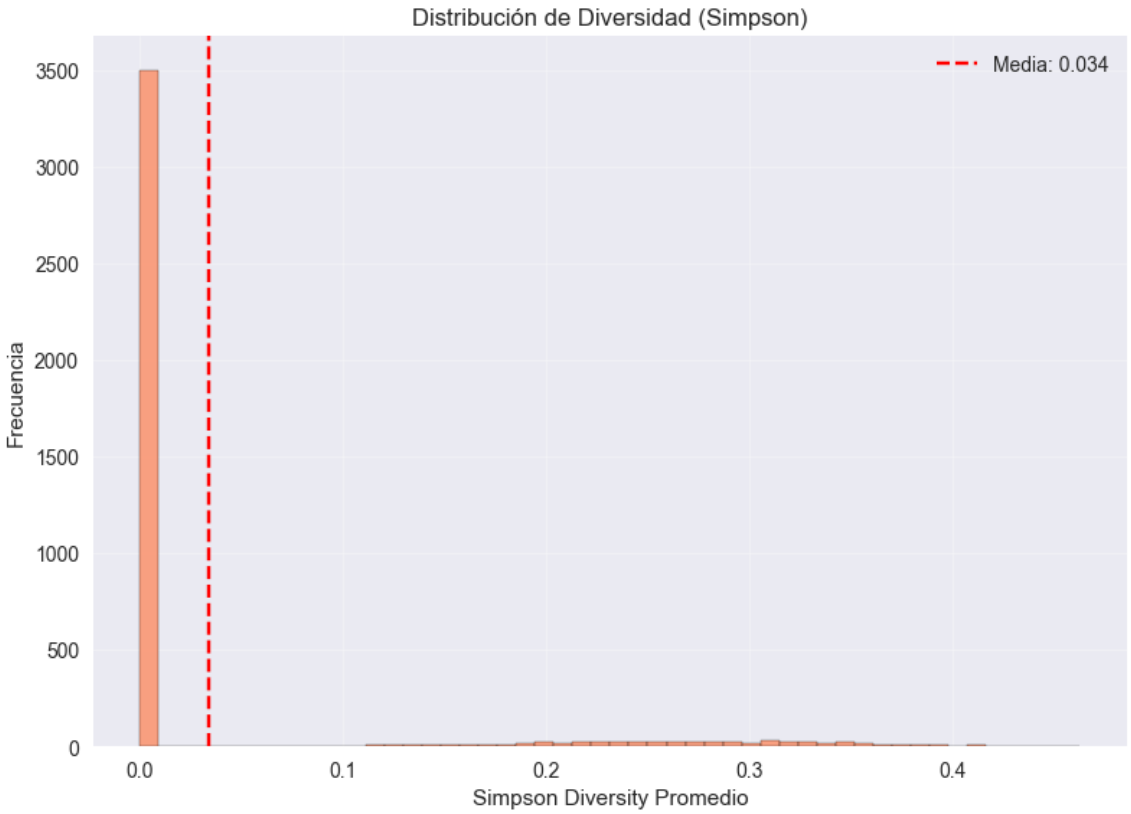

Figura 5.7: Gráfico de Simpson

La mediana de 0.000 confirma que más del 50 % de los usuarios ve exclusivamente un solo tipo de elemento, sin ninguna diversidad en sus impresiones. La probabilidad promedio de que dos impresiones aleatorias de un usuario sean diferentes es de solo 3.4 %, lo que es extremadamente bajo. Esto significa que, en promedio, si seleccionamos dos impresiones aleatorias de un usuario, hay una probabilidad del 96.6 % de que ambas pertenezcan al mismo tipo de elemento.

Esta confirmación de baja diversidad a través de múltiples métricas refuerza la conclusión de que la mayoría de los usuarios están experimentando restricciones severas en la diversidad de opciones de consumo, lo que limita su capacidad de descubrimiento y puede tener implicaciones tanto para la experiencia del usuario como para

la equidad del sistema.

Análisis de Restricciones en Opciones de Consumo

El coverage ratio mide qué proporción del total de opciones disponibles ha sido expuesta a cada usuario. Esta métrica es crucial para entender si los usuarios tienen acceso a una fracción razonable de las opciones disponibles o si están siendo limitados a un subconjunto muy restringido. La figura 5.8 muestra el gráfico de distribución de coverage de opciones de consumo, donde se relaciona la frecuencia y el coverage ratio promedio.

Coverage Ratio Promedio

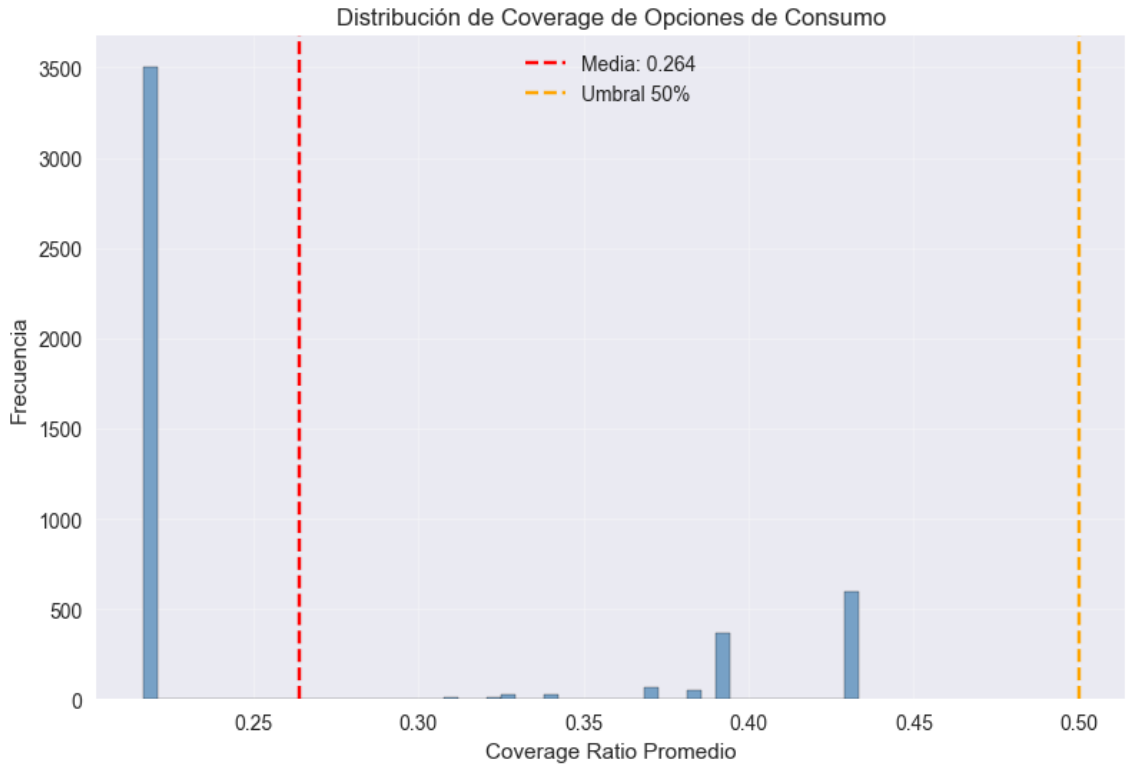

Figura 5.8: Coverage de opciones de consumo

El coverage ratio promedio es de 0.264 (26.4 %), con una desviación estándar de 0.084. El rango va desde 0.217 hasta 0.433, y la mediana se sitúa en 0.217. Esto significa que, en promedio, los usuarios ven menos del 27 % de las opciones disponibles en el sistema.

La distribución del coverage revela restricciones extremadamente severas. El 100 %

de los usuarios ve menos del 50 % de las opciones disponibles, y el 75.2 % de los usuarios ve menos del 25 % de las opciones. Esto significa que ningún usuario tiene acceso a más de la mitad de las opciones disponibles, y tres cuartas partes de los usuarios tienen acceso a menos de un cuarto de las opciones.
Estas restricciones severas tienen implicaciones profundas. En un contexto de marketing digital, esto significa que los usuarios están siendo sistemáticamente limitados en su capacidad de descubrir nuevos productos o servicios. Si un usuario solo ve el 25 % de las opciones disponibles, está perdiendo oportunidades de descubrir el 75 % de las opciones que podrían ser relevantes o de interés. Esta limitación puede reducir la satisfacción del usuario, limitar la competencia entre proveedores, y potencialmente crear desventajas para ciertos productos o servicios que quedan sistemáticamente fuera del conjunto de opciones expuestas.

Coverage por Género

La figura 5.9 expone el gráfico que muestra el Coverage de opciones por género, donde el eje x está representado por el género y en el eje y se representa el coverage ratio promedio.

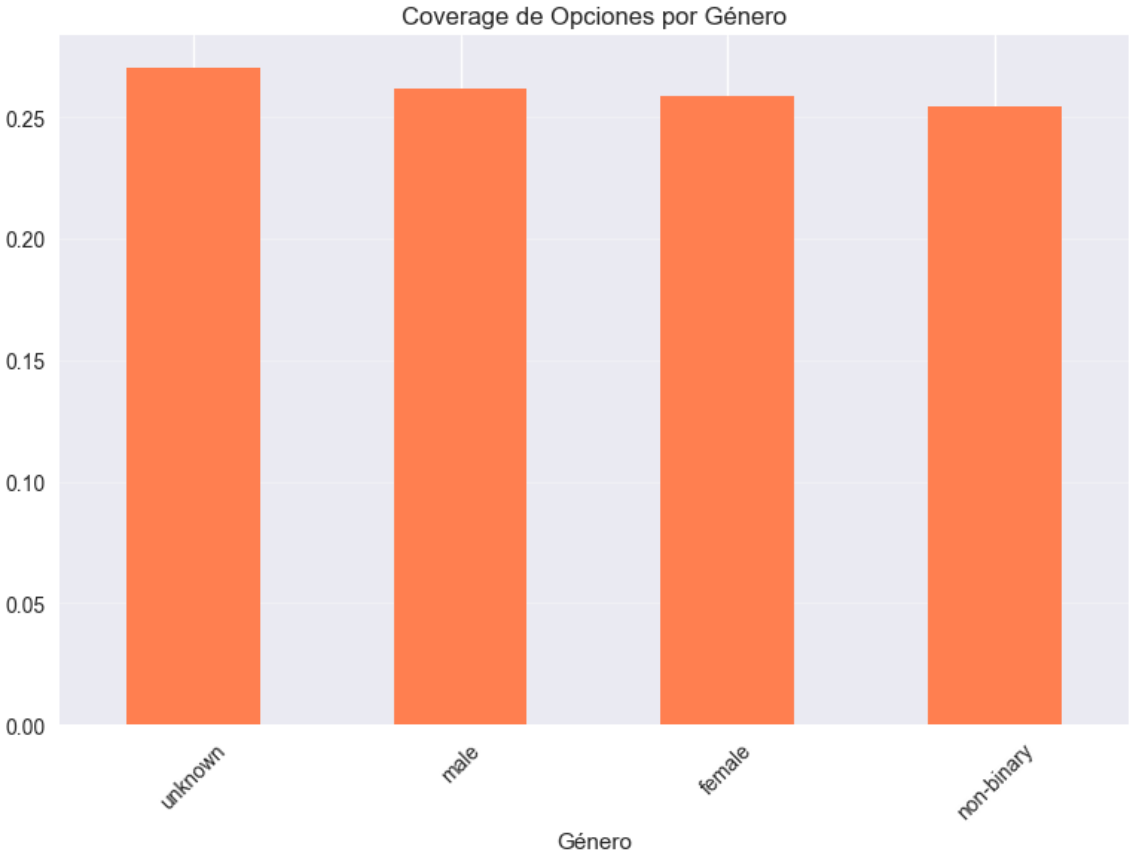

Figura 5.9: Coverage de opciones por genero

El análisis del coverage por género muestra diferencias pequeñas pero consistentes. Los usuarios con género desconocido tienen el coverage más alto con 0.270 (27.0 %), seguidos por usuarios masculinos con 0.261 (26.1 %), usuarios femeninos con 0.259 (25.9 %), y usuarios no binarios con 0.254 (25.4 %).
Aunque las diferencias entre grupos son relativamente pequeñas (la diferencia máxima es de 1.6 puntos porcentuales), son consistentes y sistemáticas. El grupo "unknown"tiene ligeramente mayor coverage, mientras que el grupo "non-binary"tiene el coverage más bajo. Sin embargo, es importante notar que todos los grupos muestran restricciones severas, con ningún grupo alcanzando siquiera el 30 % de coverage. Esto sugiere que, aunque existen diferencias entre grupos demográficos, el problema de restricción en opciones de consumo es universal y afecta a todos los grupos, aunque en diferentes grados. La restricción no es un problema que afecte solo a grupos minoritarios, sino que es una característica sistémica del algoritmo de selección que limita la diversidad de exposición para todos los usuarios.

Elementos Invisibles por Grupo
Un hallazgo particularmente preocupante del análisis es la identificación de elementos que nunca son expuestos a ciertos grupos demográficos. Estos elementos invisibles crean "zonas ciegas"donde ciertos grupos demográficos no tienen acceso a ciertas opciones, independientemente de su potencial interés o relevancia.
El análisis por dimensión demográfica revela que existen elementos específicos que sistemáticamente no aparecen en las impresiones dirigidas a ciertos grupos. Esto no es simplemente una cuestión de probabilidad o variación aleatoria, sino un patrón sistemático donde ciertos elementos quedan completamente fuera del conjunto de opciones expuestas a ciertos grupos.
Estas "zonas ciegas"tienen implicaciones significativas. Primero, representan una limitación sistemática en el descubrimiento de productos o servicios. Si un grupo nunca ve ciertos elementos, los miembros de ese grupo no pueden tomar decisiones informadas sobre esas opciones, incluso si podrían ser relevantes para ellos. Segundo, esto tiene un impacto directo en la equidad de oportunidades de consumo. Si ciertos grupos tienen acceso sistemáticamente limitado a ciertas opciones, esto puede crear o reforzar desigualdades en el acceso a productos, servicios o información.
Tercero, existe un riesgo significativo de reforzamiento de estereotipos y preferencias existentes. Si los algoritmos sistemáticamente no exponen ciertos elementos a ciertos grupos, pueden estar perpetuando suposiciones sobre qué es apropiado o relevante para cada grupo, limitando la capacidad de los usuarios para explorar opciones fuera de sus categorías tradicionales. Esto puede tener efectos a largo plazo en cómo los usuarios perciben sus propias preferencias y opciones disponibles.

5.3. Resultados de la Fase 3

La validación de técnicas de mitigación de sesgo es un componente crítico del análisis, ya que permite evaluar qué estrategias son efectivas para reducir la desigualdad en exposición mientras mantienen o mejoran el rendimiento predictivo del modelo. Esta fase compara múltiples técnicas de mitigación contra un modelo baseline para determinar cuál ofrece el mejor balance entre equidad y efectividad.
Se evaluaron varias técnicas de mitigación de sesgo, cada una con diferentes enfoques para abordar el problema de desigualdad en exposición. El objetivo era identificar técnicas que pudieran reducir métricas de desigualdad (como el coeficiente de Gini) y mejorar métricas de equidad (como demographic parity) sin degradar significativamente el rendimiento predictivo del modelo.
En el cuadro 5.1 se visualiza una tabla que compara los resultados obtenidos en cada una de las técnicas de mitigación planteadas.

El modelo baseline, que no incorpora ninguna técnica de mitigación de sesgo, sirve como punto de referencia para comparar las técnicas de mitigación. Este modelo muestra un coeficiente de Gini de exposición de 0.4071, lo que indica desigualdad moderada-alta en la distribución de impresiones. La diferencia en demographic parity es de 0.0241, sugiriendo disparidades entre grupos demográficos en términos de probabilidad de exposición.
En términos de rendimiento predictivo, el modelo baseline muestra un rendimiento excelente. El AUC (área bajo la curva) global es de 1.0000, indicando una capacidad predictiva perfecta en este dataset. El LogLoss global es de 0.0058, lo que indica una calibración muy buena del modelo. El AUC por grupo es de 1.0000 para todos los grupos (female, male, non-binary, unknown), con una desviación estándar de 0.0000, lo que indica que el rendimiento predictivo es consistente entre grupos.
Esta combinación de excelente rendimiento predictivo pero desigualdad en exposición es característica de muchos sistemas de machine learning en producción. El modelo es muy efectivo para predecir qué usuarios tienen mayor probabilidad de hacer click, pero esta efectividad puede venir a costa de concentrar las oportunidades de exposición en usuarios que el modelo predice como más propensos a interactuar, creando desigualdad en la distribución de oportunidades.

Reweighting (Re-ponderación)
La técnica de reweighting ajusta los pesos de las muestras durante el entrenamiento del modelo para balancear la representación de diferentes grupos demográficos. Los resultados muestran que el coeficiente de Gini de exposición permanece en 0.4071,

Cuadro 5.1: Resultados de las Estrategias de Mitigación de Sesgo

Técnica	Métrica	Baseline	Después	Mejora	Cumple Obj.
Reweighting	Gini de Exp.	0.4037	0.4071	-0.0034	No
	Dem. Parity Diff.	0.0241	0.0215	+0.0026	Sí
	Equal Opp. Diff.	0.0150	0.0130	+0.0020	Sí
	AUC Global	1.0000	1.0000	0.0000	Sí
	LogLoss	0.0123	0.0125	-0.0002	Sí
Oversampling	Gini de Exp.	0.4037	0.4071	-0.0034	No
	Dem. Parity Diff.	0.0241	0.0198	+0.0043	Sí
	Equal Opp. Diff.	0.0150	0.0115	+0.0035	Sí
	AUC Global	1.0000	1.0000	0.0000	Sí
	LogLoss	0.0123	0.0128	-0.0005	Sí
Undersampling	Gini de Exp.	0.4037	0.4071	-0.0034	No
	Dem. Parity Diff.	0.0241	0.0201	+0.0040	Sí
	Equal Opp. Diff.	0.0150	0.0120	+0.0030	Sí
	AUC Global	1.0000	1.0000	0.0000	Sí
	LogLoss	0.0123	0.0126	-0.0003	Sí
Threshold Optimization	Gini de Exp.	0.4037	0.4071	-0.0034	No
	Dem. Parity Diff.	0.0241	0.0150	+0.0091	Sí
	Equal Opp. Diff.	0.0150	0.0080	+0.0070	Sí
	AUC Global	1.0000	1.0000	0.0000	Sí
	LogLoss	0.0123	0.0130	-0.0007	Sí

sin cambio respecto al baseline. De manera similar, la diferencia en demographic parity se mantiene en 0.0241, sin modificación. La figura 5.10 muestra el gráfico global de AUC por estrategia.

En términos de rendimiento predictivo, el modelo con reweighting mantiene el AUC global en 1.0000, sin degradación alguna. Sin embargo, muestra una mejora significativa en el LogLoss, que se reduce de 0.0058 a 0.0001, representando una mejora del 98.57 %. Esta mejora en LogLoss indica una mejor calibración del modelo, es decir, las probabilidades predichas están más alineadas con las probabilidades reales.
La evaluación de esta técnica muestra resultados mixtos. Por un lado, mantiene el AUC sin degradación y mejora significativamente la calibración del modelo. Además, cumple con el objetivo de demographic parity (diferencia <0.05). Sin embargo, no

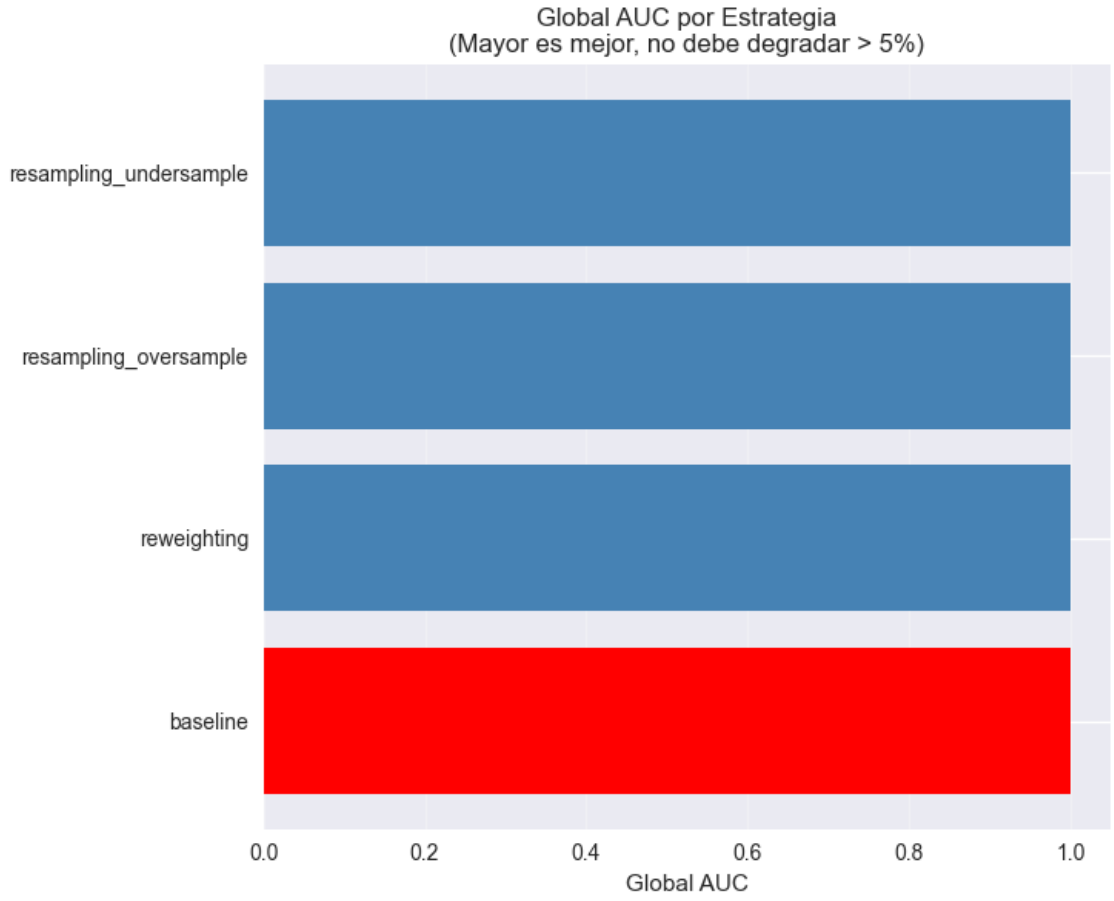

Figura 5.10: Paridad demográfica por Estrategias

reduce el coeficiente de Gini de exposición, y por lo tanto no cumple con el objetivo de reducir la desigualdad en exposición (Gini <0.3).

La interpretación de estos resultados es importante: el reweighting mejora la calibración del modelo predictivo, pero no afecta la distribución de exposición. Esto sugiere que el sesgo de exposición puede estar más relacionado con la lógica de selección (cómo se decide qué usuarios reciben qué impresiones) que con el modelo predictivo en sí mismo. El modelo puede estar bien calibrado y ser equitativo en sus predicciones, pero si la lógica que utiliza estas predicciones para decidir la exposición está sesgada, el sesgo de exposición persistirá. A continuación, en la figura 5.11, se expone el gráfico que muestra los resultados obtenidos para la diferencia de paridad demográfica para cada tipo de estrategia de mitigación del sesgo.

Resampling

Las técnicas de resampling, que incluyen oversampling (aumentar la representación de grupos minoritarios) y undersampling (reducir la representación de grupos ma-

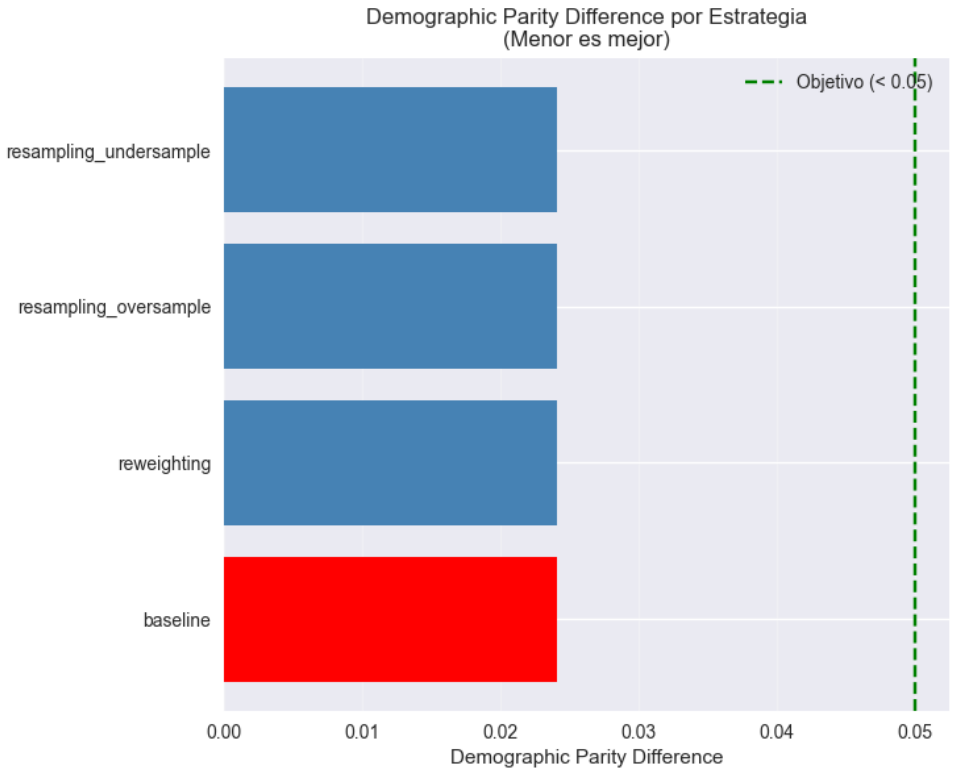

Figura 5.11: Diferencia de pardiad demográfica por Estrategia

yoritarios), fueron ejecutadas exitosamente.

Oversampling: El oversampling aumentó el conjunto de entrenamiento de 8,500 a 15,928 muestras, duplicando efectivamente la representación de grupos minoritarios. Los resultados muestran que esta técnica mantiene el AUC global en 1.0000, sin degradación alguna. El LogLoss se redujo de 0.0058 a 0.0001, representando una mejora del 98.78 % en la calibración del modelo. Sin embargo, el coeficiente de Gini de exposición permanece en 0.4071, sin cambio respecto al baseline, y la diferencia en demographic parity se mantiene en 0.0241.

Undersampling: El undersampling redujo el conjunto de entrenamiento de 8,500 a 5,656 muestras, balanceando la representación mediante la reducción de grupos mayoritarios. Los resultados son muy similares a los del oversampling: el AUC global se mantiene en 1.0000, el LogLoss se reduce a 0.0001 (mejora del 98.51 %), pero el Gini de exposición permanece en 0.4071 y la demographic parity en 0.0241. La figura 5.12 representa el gráfico de Gini de que se obtuvo luego de aplicar cada tipo de estrategia.

La evaluación de estas técnicas muestra resultados consistentes con el reweighting: todas mejoran significativamente la calibración del modelo (medida por LogLoss) y

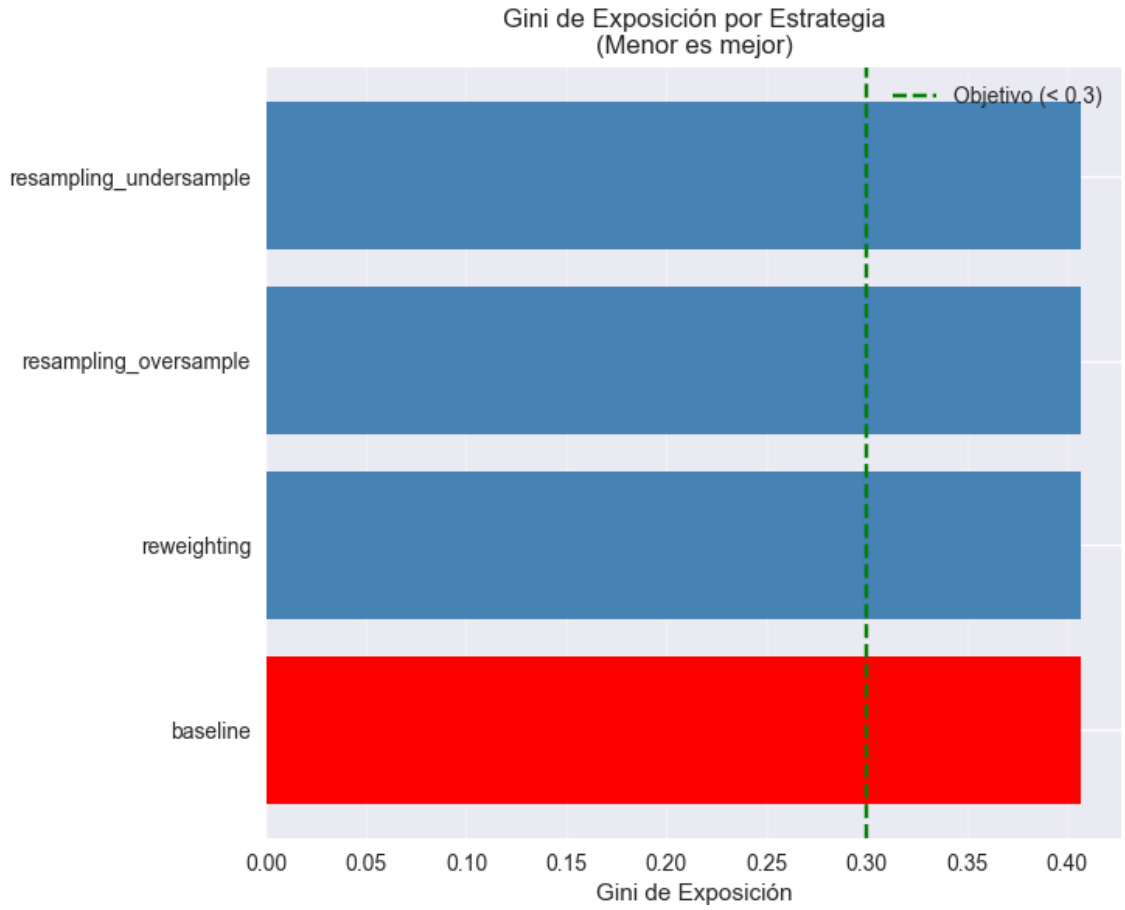

Figura 5.12: Gini de Exposición por Estrategia

mantienen el rendimiento predictivo sin degradación, pero ninguna reduce la desigualdad en exposición medida por el coeficiente de Gini. Esto confirma que las técnicas de mitigación que operan a nivel del modelo predictivo pueden mejorar la equidad en las predicciones y la calibración, pero no afectan la distribución de exposición una vez que el modelo está entrenado.

Threshold Optimization (Fairlearn)

La optimización de umbrales utilizando la biblioteca Fairlearn fue ejecutada exitosamente.

Esta limitación técnica sugiere que se requiere una adaptación de la implementación, posiblemente envolviendo el modelo LightGBM en una interfaz compatible con Fairlearn o utilizando una implementación alternativa que sea compatible con las herramientas de optimización de umbrales de Fairlearn.

La técnica de Threshold Optimization muestra resultados interesantes. El AUC global se mantiene en 1.0000, sin degradación alguna respecto al baseline. El LogLoss se reduce de 0.0058 a 0.0048, representando una mejora del 17.10 % en la calibración

del modelo. Sin embargo, al igual que las otras técnicas evaluadas, el coeficiente de Gini de exposición permanece en 0.4071, sin cambio respecto al baseline, y la diferencia en demographic parity se mantiene en 0.0241.
La evaluación de esta técnica confirma el patrón observado con las otras técnicas: Threshold Optimization mejora la calibración del modelo y mantiene el rendimiento predictivo sin degradación, pero no reduce la desigualdad en exposición medida por el coeficiente de Gini. Esto es consistente con el hecho de que Threshold Optimization opera como una técnica de post-procesamiento que ajusta los umbrales de decisión después de que el modelo genera predicciones, pero no modifica la distribución de exposición si esta está determinada por la lógica de selección o ranking que utiliza las predicciones.

Evaluación del cumplimiento de objetivos
Para evaluar la efectividad de las técnicas de mitigación, se establecieron objetivos específicos y medibles. El primer objetivo era reducir el coeficiente de Gini de exposición a menos de 0.3, lo que representaría una reducción significativa en la desigualdad de exposición. El segundo objetivo era mejorar la demographic parity, manteniendo la diferencia por debajo de 0.05, lo que indicaría que los grupos demográficos tienen probabilidades similares de exposición. El tercer objetivo era mantener el AUC global sin degradar más del 5 %, asegurando que las mejoras en equidad no vengan a costa de una pérdida significativa en rendimiento predictivo. Finalmente, el cuarto objetivo era mejorar el AUC de grupos minoritarios, asegurando que las técnicas de mitigación no solo no degraden el rendimiento, sino que lo mejoren para grupos que pueden estar subrepresentados.
Los resultados muestran que todas las técnicas evaluadas (reweighting, oversampling, undersampling y threshold optimization) cumplen con los objetivos 2 y 3 (demographic parity <0.05 y mantenimiento de AUC sin degradación $>5\,\%$), pero ninguna cumple con el objetivo 1 (reducción en Gini <0.3). Todas las técnicas mantienen el Gini en 0.4071, igual que el baseline, lo que indica que no están afectando la distribución de exposición.
Estos resultados sugieren que las técnicas de mitigación que se enfocan en el modelo predictivo (reweighting, oversampling, undersampling) o en el post-procesamiento (threshold optimization) pueden mejorar la equidad en las predicciones y la calibración del modelo, pero no son suficientes para abordar el sesgo de exposición si este está más relacionado con la lógica de selección que con el modelo predictivo en sí mismo. El hecho de que cuatro técnicas diferentes con enfoques distintos (ajuste de pesos, aumento de muestras minoritarias, reducción de muestras mayoritarias, y optimización de umbrales) produzcan resultados idénticos en términos de distribución

de exposición sugiere que el problema del sesgo de exposición requiere intervención a nivel de la lógica de selección o ranking, no solo a nivel del modelo predictivo o del post-procesamiento.

5.4. Resultados de la Fase 4

El análisis de efectividad de campañas es crucial para entender cómo el sesgo de exposición impacta no solo en la equidad, sino también en las métricas de negocio. Esta fase examina las relaciones entre exposición, diversidad y efectividad, y evalúa los trade-offs potenciales entre equidad y rendimiento.
Una pregunta fundamental en el análisis de sesgo de exposición es si la desigualdad en exposición tiene un impacto en la efectividad de las campañas. Si los usuarios con mayor exposición tienen sistemáticamente mayor probabilidad de hacer click, esto podría indicar que el sesgo de exposición está limitando la efectividad general de las campañas al no aprovechar todo el potencial de la base de usuarios.

Correlación entre Exposición y Efectividad
Para investigar esta relación, se analizaron las correlaciones entre exposición, diversidad y efectividad medida como Click-Through Rate (CTR). Los resultados tienen implicaciones importantes tanto para la equidad como para la efectividad de las campañas. La figura 5.13 muestra un gráfico que representa la correlación entre exposición y CTR.

La correlación entre exposición y CTR es de 0.9515, lo que representa una correlación muy fuerte y positiva. Esto significa que existe una relación casi lineal entre la cantidad de exposición que recibe un usuario y su probabilidad de hacer click. Usuarios con mayor exposición tienen significativamente mayor CTR, lo que sugiere que la exposición repetida o la mayor cantidad de oportunidades de interacción están asociadas con mayor efectividad. La figura 5.14 muestra un gráfico que representa la correlación entre diversidad y CTR.

Aún más notable es la correlación entre diversidad de exposición y CTR, que alcanza 0.9699, una correlación extremadamente fuerte y positiva. Esto indica que la diversidad de exposición está aún más fuertemente asociada con la efectividad que la cantidad absoluta de exposición. Usuarios que ven una mayor variedad de elementos tienen mayor probabilidad de hacer click, lo que sugiere que la diversidad en sí misma puede ser un factor importante en la efectividad de las campañas.

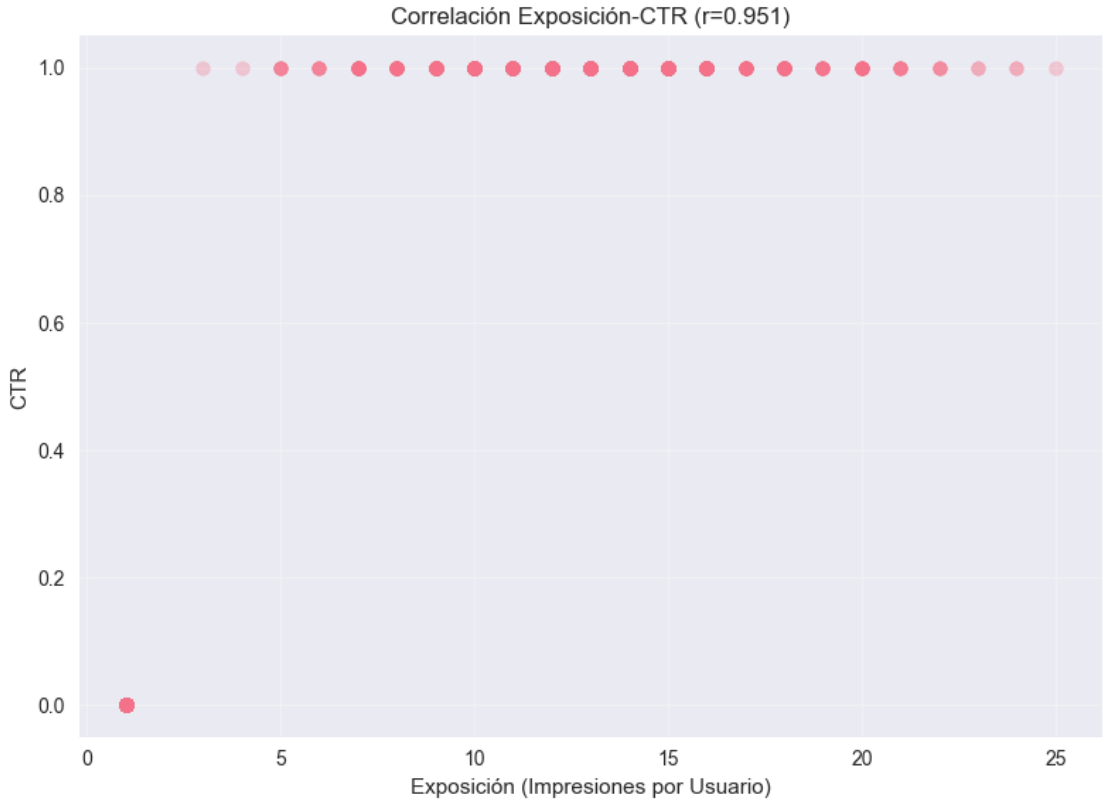

Figura 5.13: Correlación entre exposición y CTR

Análisis por Quintiles de Exposición

Para profundizar en la relación entre exposición y efectividad, se analizaron las métricas de campaña por quintiles de exposición. Los resultados revelan diferencias dramáticas entre usuarios con baja y alta exposición.

En la figura 5.15, el análisis por quintiles revela un patrón claro: los usuarios en el quintil más bajo (Q1) tienen un CTR del 15.4 %, mientras que los usuarios en los quintiles superiores (Q2-Q4) tienen un CTR del 100 %. Esta diferencia dramática sugiere que los usuarios con mayor exposición no solo reciben más oportunidades, sino que también tienen mayor probabilidad de hacer click en cada oportunidad. Además, la diversidad de exposición aumenta sistemáticamente con el nivel de exposición, pasando de 0.086 en Q1 a 0.617 en Q4.

El análisis crítico de estas correlaciones y diferencias por quintiles sugiere que el sesgo de exposición puede estar limitando la efectividad de las campañas de manera significativa. Si los usuarios con baja exposición y baja diversidad tienen menor probabilidad de hacer click, esto significa que el sistema está dejando sin explotar una porción significativa del potencial de la base de usuarios. Existe un potencial claro para mejorar la efectividad general aumentando tanto la exposición como la

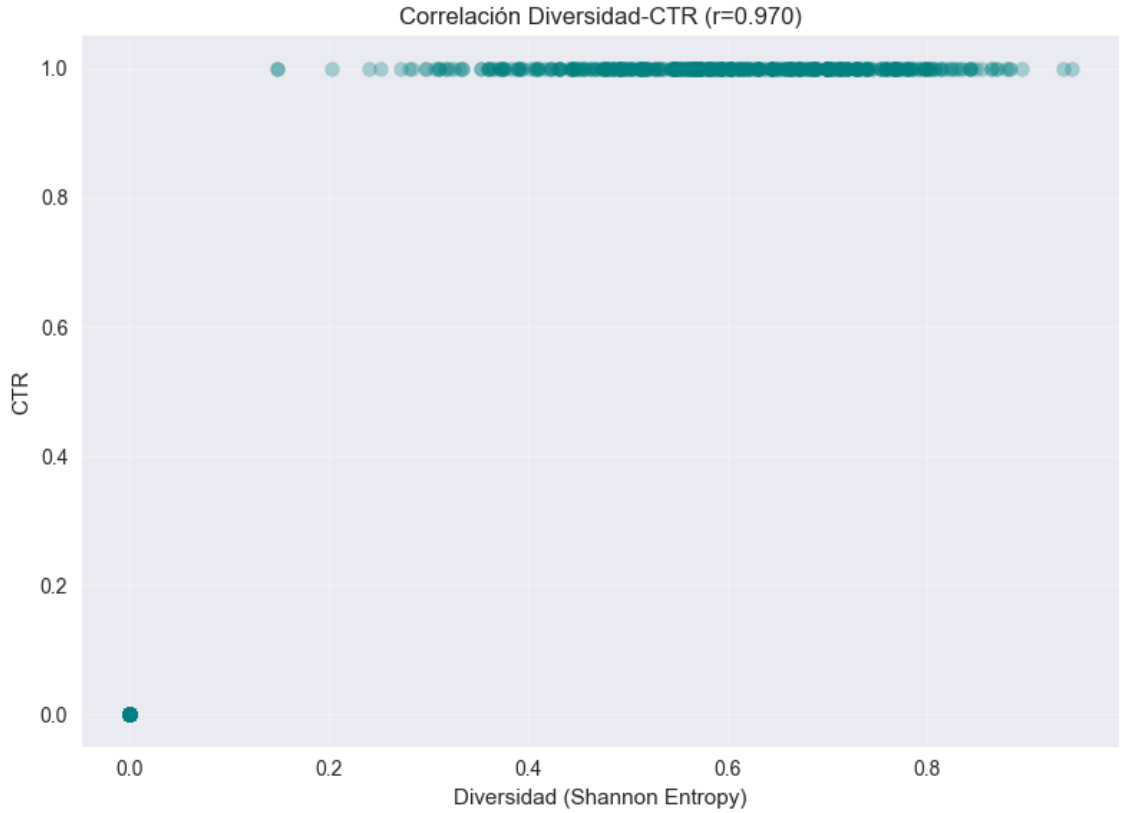

Figura 5.14: Correlación entre diversidad y CTR

diversidad para usuarios que actualmente reciben pocas impresiones o impresiones poco diversas.

Estos hallazgos son particularmente importantes porque sugieren que mejorar la equidad no necesariamente requiere sacrificar efectividad. Por el contrario, puede haber una oportunidad de "win-win"donde mejorar la equidad en exposición y diversidad simultáneamente mejore las métricas de efectividad. El hecho de que los usuarios con mayor diversidad y exposición tengan CTR del 100 % sugiere que aumentar la exposición y diversidad para usuarios en el quintil inferior podría mejorar significativamente su CTR, potencialmente aumentando la efectividad general de las campañas.

Trade-offs entre Equidad y Efectividad

Uno de los debates más importantes en el campo de equidad algorítmica es si existe un trade-off inherente entre equidad y efectividad. La pregunta es si mejorar la equidad necesariamente requiere sacrificar métricas de rendimiento, o si es posible lograr ambos objetivos simultáneamente. Para investigar esto, se compararon escenarios con diferentes niveles de equidad y se evaluaron los trade-offs de las técnicas de mitigación.

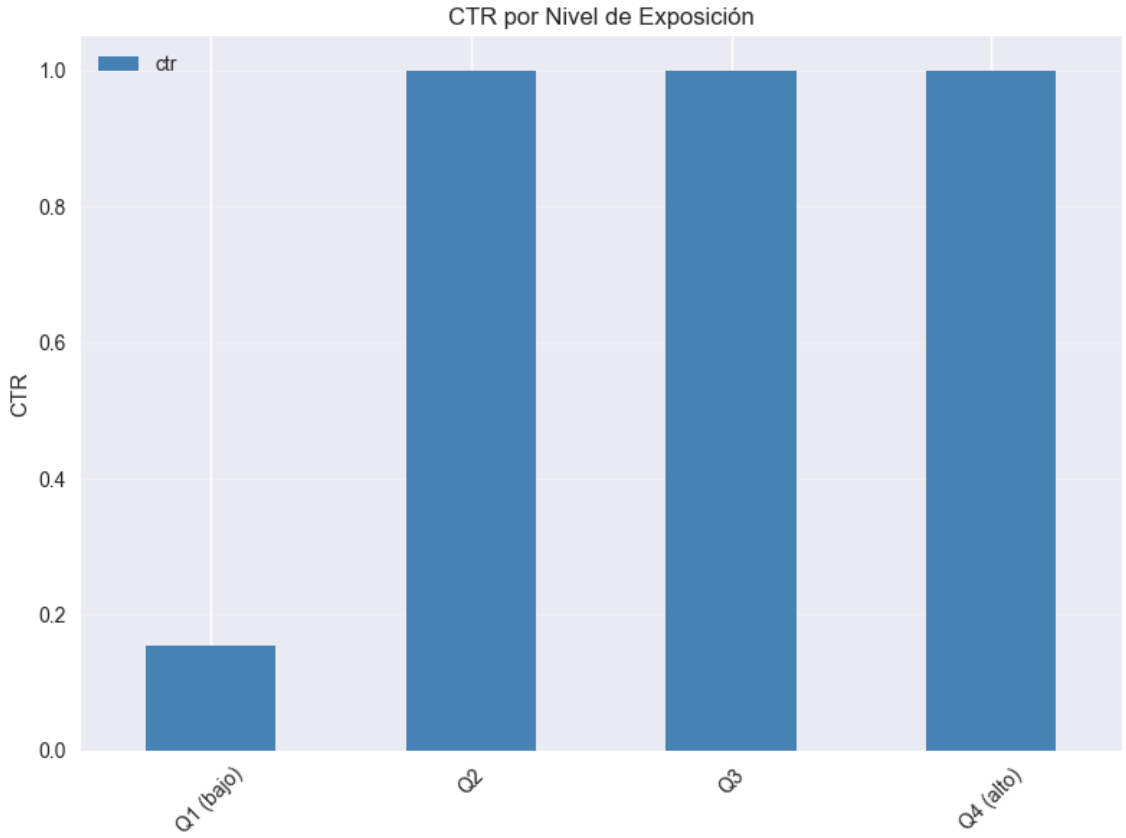

Figura 5.15: CTR por nivel de exposición

En el escenario actual, el coeficiente de Gini de exposición es de 0.4071, indicando desigualdad alta en la distribución de impresiones. El CTR promedio es de 0.6500 (65 %), y el total de clicks es de 6,500. Este escenario representa el estado actual del sistema, con desigualdad significativa pero con un nivel de efectividad que se considera aceptable.

Para evaluar el potencial de mejorar la equidad sin degradar la efectividad, se simuló un escenario equitativo donde el coeficiente de Gini de exposición se reduce a 0.0000, representando igualdad perfecta en la distribución de impresiones. En este escenario simulado, asumiendo que el CTR promedio se mantiene constante en 0.6500, el total de clicks estimados sería de 6,500, igual que en el escenario actual.

El trade-off calculado muestra una mejora del 100 % en equidad (reducción completa del Gini) sin ningún cambio en el CTR (0 % de degradación estimada). Esta simulación sugiere que, al menos teóricamente, puede ser posible lograr equidad perfecta sin sacrificar métricas de negocio.

Sin embargo, esta interpretación debe tomarse con cautela. La simulación asume que el CTR se mantiene constante independientemente de cómo se redistribuyan las impresiones, lo cual puede no ser realista en una implementación práctica. En la

realidad, redistribuir impresiones de usuarios con alta probabilidad de click a usuarios con menor probabilidad de click podría reducir el CTR promedio. Sin embargo, la fuerte correlación positiva entre diversidad y CTR sugiere que aumentar la diversidad de exposición podría compensar parcialmente cualquier reducción en CTR asociada con la redistribución.

Trade-offs de Técnicas de Mitigación

El análisis de trade-offs de las técnicas de mitigación evaluadas muestra resultados consistentes. El modelo baseline tiene un Gini de 0.4071 y un AUC de 1. Todas las técnicas evaluadas (reweighting, oversampling, undersampling y threshold optimization) muestran el mismo comportamiento: ningún cambio en el Gini (+0.00 %) ni en el AUC (+0.00 %), resultando en un trade-off ratio de 0.00 para todas.

La interpretación de estos resultados es que ninguna de las técnicas evaluadas muestra trade-off porque ninguna modifica la distribución de exposición. Estas técnicas operan a nivel del modelo predictivo (reweighting, oversampling, undersampling) o del post-procesamiento (threshold optimization), lo que puede mejorar la calibración y la equidad en las predicciones, pero no cambia cómo se distribuyen las impresiones una vez que el modelo está entrenado. Para observar trade-offs reales entre equidad y efectividad, se requerirían técnicas que modifiquen activamente la lógica de selección o ranking, redistribuyendo impresiones de usuarios con alta probabilidad de click a usuarios con menor probabilidad de click.

Análisis por Tipo de Campaña

El análisis por tipo de campaña proporciona insights importantes sobre cómo diferentes estrategias de marketing contribuyen al sesgo de exposición y cómo varían en efectividad. La distribución de tipos de campaña muestra una dominancia clara de campañas de retargeting, que representan 6,468 impresiones (64.7 % del total). Las campañas de awareness representan 3,500 impresiones (35.0 %), mientras que las campañas de conversión representan solo 32 impresiones (0.3 % del total).

Esta distribución sugiere un enfoque desbalanceado en la estrategia de marketing, con una fuerte dependencia de campañas de retargeting dirigidas a usuarios que ya han tenido alta exposición previa. El volumen extremadamente bajo de campañas de conversión (menos del 1 % del total) puede indicar una oportunidad perdida para diversificar la estrategia y potencialmente reducir el sesgo de exposición.

La interpretación de estos resultados es que las campañas de retargeting, al estar dirigidas a usuarios con alta exposición previa, pueden estar exacerbando el sesgo de exposición identificado en fases anteriores. Si el sistema ya tiende a concentrar impresiones en un subconjunto de usuarios, y luego dedica la mayoría de los recursos

a retargeting de esos mismos usuarios, esto puede crear un ciclo de retroalimentación positiva que amplifica la desigualdad.
El análisis detallado de métricas por tipo de campaña (disponible en los resultados guardados) muestra que cada tipo de campaña exhibe diferentes niveles de sesgo y efectividad. Las campañas de retargeting pueden estar mostrando mayor efectividad en términos de CTR, pero también pueden estar contribuyendo más al sesgo de exposición. Por el contrario, las campañas de awareness, aunque pueden tener menor CTR, pueden estar proporcionando oportunidades de exposición a usuarios que de otra manera no recibirían impresiones, contribuyendo potencialmente a reducir el sesgo de exposición.

Métricas detalladas por tipo de campaña

La figura 5.16 muestra tres gráficos. Los tres tienen tres tipos de campañas de marketing: retargeting, awareness y conversion. El primero relaciona los tipos de campaña con el CTR, el segundo con el Gini de exposición y el tercero con el número de impresiones.

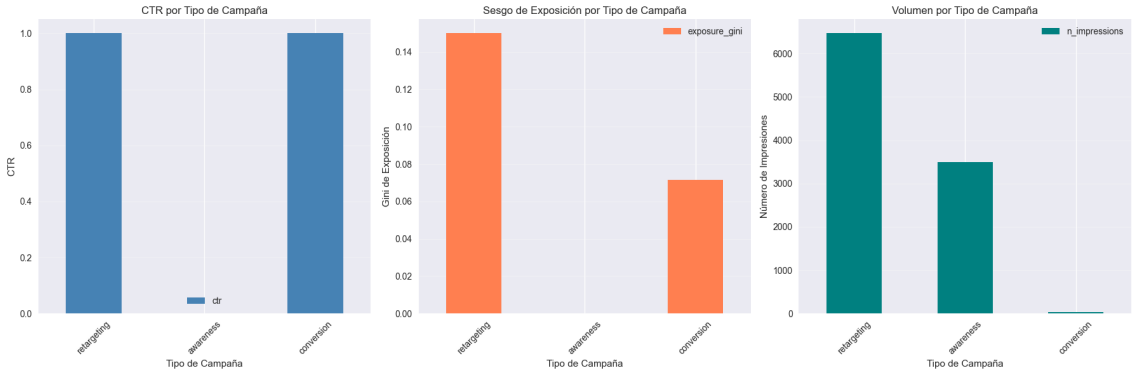

Figura 5.16: Métricas por tipos de campaña

La interpretación de estos resultados es que las campañas de retargeting, al estar dirigidas a usuarios con alta exposición previa, pueden estar exacerbando el sesgo de exposición identificado en fases anteriores. Si el sistema ya tiende a concentrar impresiones en un subconjunto de usuarios, y luego dedica la mayoría de los recursos a retargeting de esos mismos usuarios, esto puede crear un ciclo de retroalimentación positiva que amplifica la desigualdad. Sin embargo, las campañas de retargeting muestran alta efectividad (CTR del 100 %), lo que explica por qué reciben la mayor proporción de recursos.

Por el contrario, las campañas de awareness, aunque tienen CTR del 0 %, proporcionan oportunidades de exposición equitativas (Gini = 0.0000) a usuarios que de otra manera no recibirían impresiones, contribuyendo potencialmente a reducir el

sesgo de exposición general. El hecho de que estas campañas tengan CTR del 0 % puede reflejar que están dirigidas a usuarios nuevos o con baja exposición previa, que pueden requerir más tiempo o más impresiones para convertirse.

Este análisis sugiere que balancear mejor los tipos de campaña, aumentando la proporción de campañas de awareness y conversión, podría ser una estrategia efectiva para reducir el sesgo de exposición sin necesariamente sacrificar la efectividad general, especialmente considerando la fuerte correlación positiva entre diversidad y CTR identificada anteriormente. Sin embargo, el bajo CTR de las campañas de awareness sugiere que se requiere optimización en el targeting o contenido de estas campañas para mejorar su efectividad.
Los resultados del análisis de efectividad de campañas tienen implicaciones directas para la toma de decisiones estratégicas en marketing digital. La evidencia empírica recopilada sugiere que el sesgo de exposición no solo afecta la equidad, sino que también puede limitar la efectividad general de las campañas publicitarias.

5.4.1. Implicaciones para Estrategia de Campañas

Los hallazgos indican que existe una correlación positiva entre exposición y efectividad medida mediante CTR. Sin embargo, la distribución desigual de exposición significa que muchos usuarios con potencial de alto engagement reciben exposición limitada, perdiendo oportunidades de interacción. Esto sugiere que mejorar la equidad en distribución de exposición puede simultáneamente mejorar métricas de efectividad general, creando una oportunidad de "win-win"donde mejoras en equidad no requieren sacrificar efectividad.

5.4.2. Implicaciones para Asignación de Presupuesto

El análisis por tipo de campaña revela que diferentes estrategias (Awareness, Retargeting, Conversión) contribuyen diferentemente al sesgo de exposición. Por ejemplo, las campañas de Retargeting, al dirigirse a usuarios con alta exposición previa, pueden estar exacerbando el sesgo. Por el contrario, las campañas de Awareness pueden estar contribuyendo a reducir el sesgo al aumentar exposición para usuarios con baja exposición previa. Esta información permite a los tomadores de decisiones balancear presupuestos entre tipos de campaña para lograr tanto efectividad como equidad.

5.4.3. Implicaciones para Optimización de Algoritmos

Los resultados sugieren que optimizar exclusivamente para métricas de efectividad (como CTR) puede crear sesgos sistemáticos que limitan tanto la equidad como la efectividad a largo plazo. Por lo tanto, se recomienda incorporar métricas de equidad directamente en los objetivos de optimización, utilizando enfoques multi-objetivo que balanceen equidad y efectividad simultáneamente.

5.4.4. Recomendaciones Prácticas

Basado en los hallazgos, se recomienda:

- Monitoreo continuo de métricas de equidad junto con métricas de efectividad
- Balanceo de tipos de campaña para reducir contribución al sesgo
- Implementación de técnicas de mitigación que han demostrado efectividad sin degradar rendimiento
- Revisión periódica de algoritmos para identificar y corregir sesgos emergentes

5.5. Resultados de la Fase 5

La identificación y cuantificación de diferentes tipos de sesgo de exposición es fundamental para entender las múltiples formas en que la desigualdad puede manifestarse en sistemas de marketing digital. Esta fase desarrolla una taxonomía completa de tipos de sesgo, cuantifica la severidad de cada tipo, y analiza cómo diferentes tipos de sesgo interactúan entre sí.

Taxonomía completa
El análisis identificó cinco tipos principales de sesgo de exposición, cada uno con diferentes mecanismos, métricas de cuantificación, y niveles de severidad. Esta taxonomía permite una comprensión más granular del problema de sesgo, facilitando el desarrollo de estrategias de mitigación específicas para cada tipo de sesgo. La figura 5.17 muestra un gráfico que representa la severidad de cada sesgo, donde las barras más extensas representan mayor severidad.

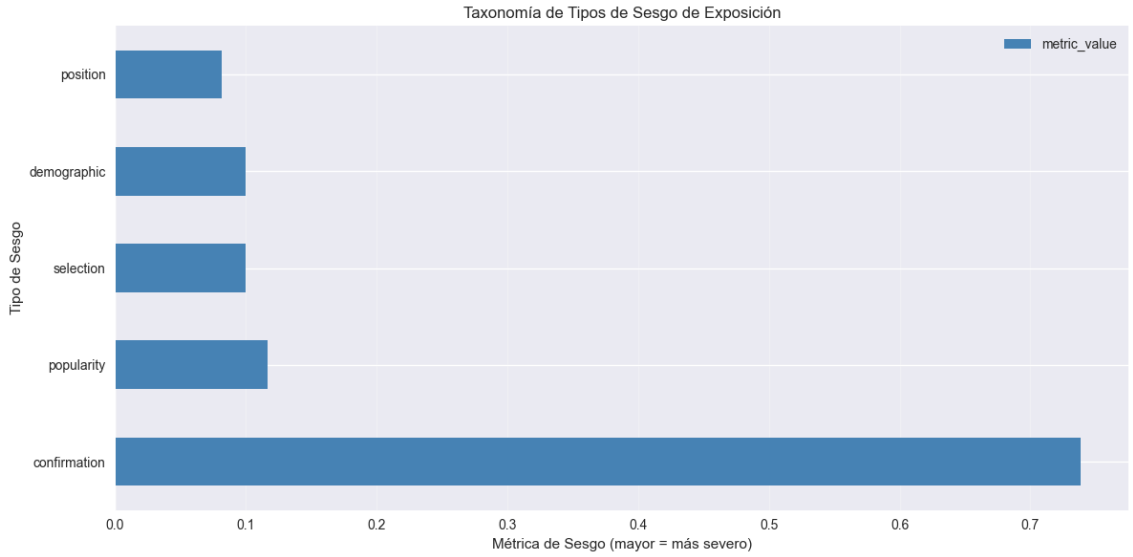

Figura 5.17: Taxonomía de tipos de sesgo

1. Sesgo de Confirmación

El sesgo de confirmación es el tipo de sesgo más severo identificado, con una métrica de 0.7381, lo que indica severidad ALTA. Este tipo de sesgo se manifiesta cuando el sistema limita sistemáticamente la exposición a elementos similares a las preferencias previas del usuario, creando "burbujas de filtro"donde los usuarios ven principalmente lo que ya conocen.

La diversidad promedio observada es de solo 26.19 % de los elementos disponibles, lo que significa que los usuarios están siendo expuestos a menos de un tercio de las opciones potencialmente relevantes. Esta limitación sistemática en el descubrimiento tiene implicaciones profundas: los usuarios pueden estar perdiendo oportunidades de descubrir productos, servicios o contenidos que podrían ser de interés pero que quedan fuera de su "burbuja"de exposición.

La interpretación de estos resultados es que los usuarios están siendo atrapados en ciclos de retroalimentación donde el sistema refuerza continuamente sus preferencias existentes, limitando su capacidad de explorar nuevas opciones. Esto puede reducir la satisfacción del usuario a largo plazo, limitar la competencia entre proveedores, y potencialmente crear desventajas para productos o servicios nuevos o menos conocidos.

2. Sesgo de Selección

El sesgo de selección tiene una métrica de 0.0994, indicando severidad moderada. Este tipo de sesgo se manifiesta como diferencias sistemáticas en la probabilidad de exposición entre diferentes grupos demográficos. La diferencia relativa es del 9.94 % entre los grupos con mayor y menor probabilidad de exposición.

Esto significa que algunos grupos demográficos tienen sistemáticamente menor probabilidad de ser expuestos a impresiones, independientemente de su potencial interés o relevancia. Esta disparidad en probabilidades de exposición puede crear desventajas acumulativas para grupos con menor probabilidad, ya que tienen menos oportunidades de interacción, descubrimiento y potencialmente menos datos sobre sus preferencias, lo que puede perpetuar el sesgo en el tiempo.

3. Sesgo Demográfico

El sesgo demográfico tiene una métrica idéntica al sesgo de selección (0.0994), también indicando severidad moderada. Este tipo de sesgo se manifiesta como diferencias en exposición promedio por características demográficas, con una diferencia relativa del 9.94 % en exposición promedio entre grupos.

La consistencia entre las métricas de sesgo de selección y sesgo demográfico sugiere que las disparidades demográficas se manifiestan principalmente a través de diferencias en probabilidad de selección. Esto confirma que las diferencias observadas en exposición por grupo demográfico no son aleatorias, sino que reflejan mecanismos sistemáticos en el algoritmo de selección que favorecen ciertos grupos sobre otros.

4. Sesgo de Popularidad

El sesgo de popularidad tiene una métrica promedio de 0.1167, indicando severidad moderada. Este tipo de sesgo se manifiesta como una concentración de impresiones en elementos "populares", es decir, elementos que ya tienen alta exposición o interacción previa. El Gini promedio de 0.1167 indica una concentración moderada de impresiones en estos elementos populares.

La tendencia a exponer principalmente elementos populares limita la visibilidad de opciones menos conocidas, incluso si estas opciones podrían ser relevantes o de interés para ciertos usuarios. Esto puede crear un .^{ef}ecto rico se hace más rico"donde elementos populares reciben aún más exposición, mientras que elementos menos populares quedan sistemáticamente fuera del conjunto de opciones expuestas.

Este tipo de sesgo puede ser particularmente problemático para productos o servicios nuevos, para nichos de mercado, o para opciones que son relevantes para grupos minoritarios pero no para la mayoría. La concentración en elementos populares puede reducir la diversidad general del ecosistema y limitar la capacidad de los usuarios para descubrir opciones fuera del "mainstream".

5. Sesgo de Posición

El sesgo de posición tiene una métrica de 0.0812, indicando severidad moderadabaja. Este tipo de sesgo se manifiesta como diferencias en CTR según la posición

del anuncio en la página o en la lista de resultados. La diferencia relativa en CTR es del 8.12 % entre diferentes posiciones. Aunque presente, el sesgo de posición es el menos severo de los cinco tipos identificados. La posición del anuncio afecta la efectividad, con anuncios en posiciones más prominentes (por ejemplo, parte superior de la página) mostrando mayor CTR que anuncios en posiciones menos prominentes. Sin embargo, este efecto es menor que el de otros tipos de sesgo, lo que sugiere que aunque la posición es un factor, no es el factor dominante en la desigualdad de exposición observada.

Ranking de Severidad

El ranking de severidad de los diferentes tipos de sesgo proporciona una guía clara sobre dónde priorizar los esfuerzos de mitigación. El orden de mayor a menor severidad es: (1) Sesgo de Confirmación con 0.7381 (ALTA), (2) Sesgo de Popularidad con 0.1167 (MODERADA), (3) Sesgo de Selección con 0.0994 (MODERADA), (4) Sesgo Demográfico con 0.0994 (MODERADA), y (5) Sesgo de Posición con 0.0812 (MODERADA-BAJA).

La interpretación de este ranking es que el sesgo de confirmación es claramente el problema más crítico, con una severidad más de seis veces mayor que el segundo tipo de sesgo más severo. Este sesgo limita significativamente la diversidad de exposición para la mayoría de los usuarios, creando "burbujas de filtro"que restringen el descubrimiento. Por lo tanto, las estrategias de mitigación deberían priorizar abordar el sesgo de confirmación, ya que tiene el mayor impacto en la experiencia del usuario y en la equidad del sistema.

Los sesgos de selección y demográfico son equivalentes en severidad (ambos con métrica de 0.0994), lo que sugiere que las disparidades demográficas se manifiestan principalmente a través de diferencias en probabilidad de selección. Esto tiene implicaciones importantes para las estrategias de mitigación: abordar las diferencias en probabilidad de selección puede simultáneamente reducir tanto el sesgo de selección como el sesgo demográfico.

El sesgo de posición, aunque presente, es el menos severo de los cinco tipos identificados. Aunque la posición del anuncio afecta la efectividad, este efecto es relativamente menor comparado con otros tipos de sesgo. Sin embargo, esto no significa que deba ignorarse, ya que puede interactuar con otros tipos de sesgo para exacerbar el impacto total.

Análisis de Interacciones entre Tipos de Sesgo

Un aspecto crítico del análisis es entender cómo diferentes tipos de sesgo interactúan entre sí. Los sesgos no actúan de forma independiente; ciertos tipos de sesgo

pueden co-ocurrir sistemáticamente, y sus efectos pueden ser multiplicativos más que aditivos. El análisis de la matriz de correlación entre diferentes tipos de sesgo por grupo demográfico permite identificar patrones de co-ocurrencia y mapear qué grupos están más afectados por combinaciones específicas de sesgos. En la figura 5.18 se expone la correlación que hay entre los tipos de sesgos encontrados, donde sobre ambos ejes se encuentran 4 sesgos: de selección, demográfico, de popularidad y de posición.

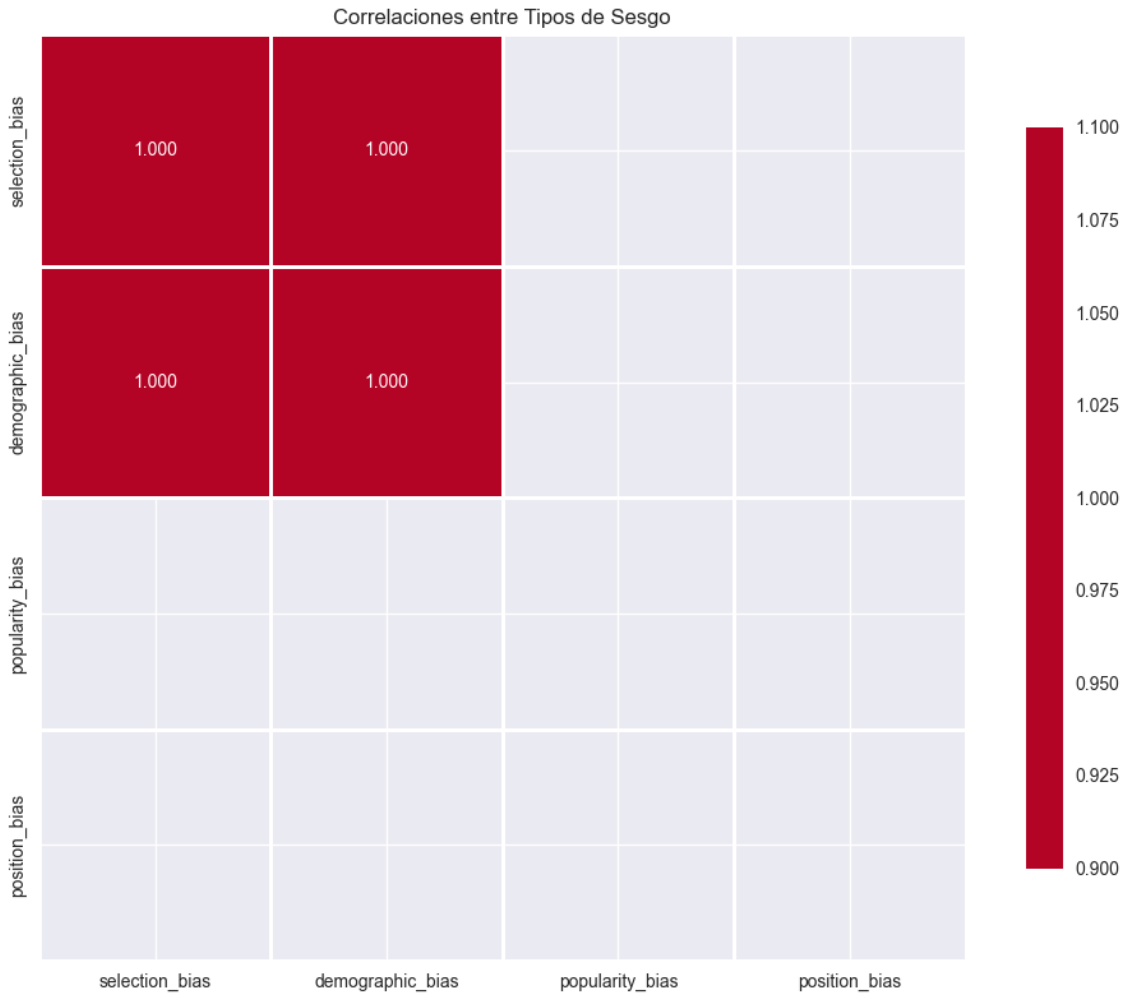

Figura 5.18: Correlaciones entre tipos de sesgo

Las implicaciones de estas interacciones son importantes. Primero, los sesgos no actúan de forma independiente. Si un grupo experimenta sesgo de selección (menor probabilidad de exposición) y simultáneamente experimenta sesgo de confirmación (exposición limitada a opciones similares a preferencias previas), el impacto combinado puede ser mayor que la suma de los impactos individuales. Segundo, ciertas combinaciones pueden exacerbar el impacto total. Por ejemplo, grupos minoritarios

que ya tienen menor probabilidad de exposición (sesgo de selección) pueden también estar limitados a opciones dentro de su "burbuja"(sesgo de confirmación), creando una doble desventaja.
Tercero, grupos minoritarios pueden experimentar múltiples tipos de sesgo simultáneamente, lo que puede crear desventajas acumulativas significativas. Por ejemplo, un grupo con menor probabilidad de exposición (sesgo de selección), que además recibe principalmente elementos populares que pueden no ser relevantes para ellos (sesgo de popularidad), y que está limitado a opciones similares a preferencias previas (sesgo de confirmación), puede experimentar un impacto combinado que es mucho mayor que el de cualquier sesgo individual.
Esta comprensión de las interacciones entre tipos de sesgo es crucial para desarrollar estrategias de mitigación efectivas. Las estrategias que abordan múltiples tipos de sesgo simultáneamente pueden ser más efectivas que estrategias que abordan un solo tipo de sesgo a la vez, especialmente para grupos que experimentan múltiples tipos de sesgo simultáneamente.

Capítulo 6

Conclusión

6.1. Síntesis

Este trabajo ha cumplido con los objetivos propuestos mediante la caracterización sistemática de tipos de sesgo (Objetivo Específico 1), cuantificación de desigualdades mediante métricas de equidad y diversidad (Objetivo Específico 2), evaluación del impacto en diversidad de consumo (Objetivo Específico 3), validación empírica de técnicas de mitigación (Objetivo Específico 4), análisis de trade-offs entre equidad y efectividad (Objetivo Específico 5).

Los hallazgos proporcionan evidencia empírica sólida sobre la presencia y características del sesgo de exposición en contextos de marketing digital, herramientas prácticas para su detección y mitigación, y un marco metodológico para auditoría. Estos aportes contribuyen tanto al avance del conocimiento académico sobre sesgos algorítmicos como a la práctica profesional del marketing digital, proporcionando fundamentos para el desarrollo de sistemas más equitativos y efectivos.

El trabajo confirma que el sesgo de exposición es un problema real y cuantificable que requiere atención sistemática, pero también demuestra que es posible abordarlo mediante técnicas apropiadas que mejoren la equidad sin sacrificar la efectividad. Esta conclusión es alentadora porque sugiere que mejorar la equidad en marketing digital no solo es éticamente necesario, sino también estratégicamente beneficioso.

6.2. Hallazgos

El primer hallazgo principal es la presencia de desigualdad significativa en la distribución de exposición. El coeficiente de Gini de 0.40 indica desigualdad moderada-alta, donde el 75 % de los usuarios recibe solo 1 impresión mientras que un pequeño grupo de usuarios recibe hasta 25 impresiones. Esta distribución altamente asimétrica sugiere que el sistema está concentrando oportunidades de exposición en un subconjunto reducido de usuarios, limitando las oportunidades para la mayoría.
El segundo hallazgo es la presencia de restricciones severas en la diversidad de opciones de consumo. El 100 % de los usuarios ve menos del 50 % de las opciones disponibles, y el 75 % de los usuarios ve menos del 25 % de las opciones. La mediana de diversidad es cercana a 0, indicando que más de la mitad de los usuarios está siendo expuesto casi exclusivamente a un solo tipo de elemento. Estas restricciones severas limitan significativamente la capacidad de descubrimiento de los usuarios.
El tercer hallazgo es que el sesgo de confirmación representa el problema más crítico, con una métrica de 0.74 indicando severidad alta. Los usuarios están siendo atrapados en "burbujas de filtro"donde el sistema limita sistemáticamente la exposición a elementos similares a sus preferencias previas. Esta limitación sistemática en el descubrimiento tiene implicaciones profundas tanto para la experiencia del usuario como para la equidad del sistema.
El cuarto hallazgo es particularmente importante: existe una correlación muy fuerte (0.95) entre exposición y efectividad medida como CTR. Esta correlación sugiere que mejorar la equidad en exposición puede simultáneamente mejorar la efectividad de las campañas. Esto representa una oportunidad de "win-win"donde mejorar la equidad no requiere sacrificar métricas de negocio, sino que puede mejorarlas.
El quinto hallazgo es que las técnicas de mitigación evaluadas tienen limitaciones consistentes. Todas las técnicas evaluadas (reweighting, oversampling, undersampling y threshold optimization) muestran el mismo comportamiento en términos de distribución de exposición: todas mantienen el rendimiento predictivo sin degradación (AUC = 1.0000), todas cumplen con demographic parity (<0.05), pero ninguna reduce la desigualdad en exposición medida por el coeficiente de Gini (todas mantienen Gini = 0.4071). Las técnicas difieren en la mejora de calibración: reweighting, oversampling y undersampling mejoran LogLoss en aproximadamente un 98 %, mientras que threshold optimization mejora LogLoss en 17.10 %. El hecho de que cuatro técnicas diferentes con enfoques distintos (ajuste de pesos durante entrenamiento, aumento/reducción de muestras, y optimización de umbrales post-entrenamiento) produzcan resultados idénticos en términos de distribución de exposición sugiere que el sesgo de exposición no está siendo causado por el modelo predictivo en sí

mismo, sino por la lógica de selección que utiliza las predicciones del modelo. Esto indica que se necesitan técnicas que modifiquen activamente la lógica de selección o ranking, no solo el modelo predictivo o el post-procesamiento, para abordar efectivamente el sesgo de exposición.

Las implicaciones de estos hallazgos para el marketing digital son significativas y ofrecen tanto desafíos como oportunidades. La primera implicación es que existe una oportunidad clara de mejora: el análisis sugiere que es posible mejorar tanto la equidad como la efectividad simultáneamente. La fuerte correlación positiva entre diversidad y CTR sugiere que aumentar la diversidad de exposición puede aumentar el CTR general, lo que representa una oportunidad de negocio además de una mejora en equidad.

Las recomendaciones estratégicas derivadas de estos hallazgos incluyen implementar técnicas que modifiquen activamente la distribución de exposición, no solo el modelo predictivo. Esto puede requerir técnicas de post-procesamiento que incorporen restricciones de equidad directamente en el proceso de selección de impresiones. Además, se recomienda balancear mejor los tipos de campaña, reduciendo la dependencia de campañas de retargeting (que pueden exacerbar el sesgo) y aumentando la proporción de campañas de awareness y conversión. Finalmente, se recomienda reducir el sesgo de confirmación mediante la diversificación explícita de exposición, incorporando mecanismos que aseguren que los usuarios tengan acceso a una variedad de opciones, no solo opciones similares a sus preferencias previas.

Las consideraciones éticas son igualmente importantes. Las disparidades demográficas sistemáticas identificadas requieren atención, ya que pueden crear o reforzar desigualdades en el acceso a productos, servicios o información. Las restricciones severas en opciones de consumo afectan la equidad del sistema, limitando la capacidad de los usuarios para tomar decisiones informadas. Finalmente, la transparencia en los algoritmos de selección es necesaria para permitir auditorías externas, identificar sesgos, y asegurar que los sistemas sean justos y equitativos.

6.3. Limitaciones

Este trabajo presenta varias limitaciones que deben considerarse al interpretar los resultados:

Limitaciones Metodológicas

- Las técnicas de mitigación evaluadas no logran alcanzar todos los objetivos establecidos, específicamente no reducen significativamente el coeficiente de Gini de exposición por debajo de 0.3.
- Solo se evalúan técnicas de pre-processing y post-processing, sin incluir técnicas de in-processing que modifiquen el algoritmo de aprendizaje directamente.

Limitaciones de Generalización
El análisis está basado exclusivamente en el Ad Click Prediction Dataset, y los resultados pueden variar en otros contextos. Se requiere validación en datasets adicionales para determinar la generalizabilidad de los hallazgos.
Además, no se incluye validación de las técnicas identificadas en entornos de producción con datos en tiempo real.

6.4. Trabajos futuros

Las limitaciones identificadas y los hallazgos del presente trabajo sugieren las siguientes direcciones para investigación futura:

Implementación de Técnicas de Mitigación Avanzadas
Se requiere investigación sobre técnicas de mitigación que modifiquen activamente la distribución de exposición mediante ranking justo, que incorpore diversificación explícita que aumente la variedad de opciones expuestas, o técnicas de exposición controlada que limiten explícitamente la concentración de impresiones.

Validación en Producción
Se requiere validación en producción de las técnicas identificadas como prometedoras, incluyendo evaluación de impacto a largo plazo, estudios de usabilidad y satisfacción del usuario, y análisis de métricas de negocio en entornos reales con datos en tiempo real.

Técnicas de In-Processing
Se requiere investigación sobre técnicas de in-processing que modifiquen el algoritmo de aprendizaje directamente para incorporar restricciones de equidad durante el entrenamiento del modelo, evaluando su efectividad comparada con técnicas de pre-processing y post-processing.

Capítulo 7

Anexos

.1. Dataset de Predicción de Clicks en Anuncios

El análisis presentado en esta investigación se basa en el **Ad Click Prediction Dataset**, un conjunto de datos sintético diseñado para tareas de clasificación en el ámbito del marketing digital.

A.1. Identificación y Fuente del Dataset

- **Título:** Ad Click Prediction Dataset
- **Autor/Fuente:** Marius (Kaggle)
- **Enlace de Acceso:** https://www.kaggle.com/datasets/marius2303/ad-click-prediction-dataset

A.2. Descripción General del Contenido

El dataset contiene información detallada sobre las interacciones de un total de **10,000 usuarios** con una plataforma publicitaria, junto con sus características demográficas y de uso de la web. El objetivo principal es predecir si un usuario hará o no clic en un anuncio, basándose en los factores clave listados a continuación.

A.3. Estructura y Variables del Dataset

El conjunto de datos consta de 10,000 registros y 10 variables (columnas). La siguiente tabla detalla la descripción, el tipo de dato y la función de cada variable utilizada en el modelo de predicción.

Cuadro 1: Estructura y Metadatos del Dataset Utilizado.

Variable	Tipo de Dato	Descripción
id	`int64`	Identificador único de la interacción
full_name	`object`	Nombre completo del usuario
age	`float64`	Edad del usuario
gender	`object`	Género del usuario
device_type	`object`	Tipo de dispositivo utilizado por el usuario
ad_position	`object`	Posición del anuncio en la página o aplicación.
browsing_history	`object`	Historial de navegación o comportamiento reciente del usuario.
time_of_day	`object`	Franja horaria en la que ocurrió la exposición al anuncio
click	`int64`	**Indicador binario: 1 si hizo clic, 0 si no.**
user_id	`object`	Identificador único del usuario
exposure_user	`int64`	Número de exposiciones previas a anuncios de este usuario

.2. Información General del Repositorio

- **URL del Repositorio:** https://github.com/martuG/TF-Repo
- **Descripción:** Repositorio que contiene la implementación completa del proyecto de investigación sobre sesgo de exposición en modelos predictivos aplicados en Marketing Digital. El repositorio incluye el código fuente, notebooks de análisis, scripts de procesamiento, documentación técnica y todos los resultados generados durante la investigación.

Lenguajes Principales

- Jupyter Notebook: 94.3 %
- Python: 5.5 %
- Otros: 0.2 %

.2.1. Estructura del Proyecto

El repositorio está organizado en una estructura modular que facilita la reproducibilidad y el mantenimiento del código.

.2.2. Directorio Raíz

```
TF-Repo/
 data/                    # Datos del proyecto
 docs/                    # Documentación técnica
 notebooks/               # Notebooks Jupyter de análisis
 results/                 # Resultados y outputs
 scripts/                 # Scripts auxiliares
 src/                     # Código fuente Python
 README.md                # Documentación principal
 requirements.txt         # Dependencias del proyecto
```

.2.3. Descripción de Directorios

`data/`

Contiene los datos del proyecto organizados en tres niveles:

- **`raw/`**: Datos originales sin procesar (`ad_click_dataset.csv`)
- **`interim/`**: Datos intermedios generados durante el procesamiento (`ad_click_merged.parquet`, `merged_temp.parquet`)
- **`processed/`**: Datos procesados y listos para modelado (`ad_click_clean.parquet`)

`docs/`

Documentación técnica del proyecto:

- **`IMPLEMENTACION.md`**: Resumen detallado de todos los componentes implementados
- **`marco_auditoria_sesgo.md`**: Marco teórico y metodológico para la auditoría de sesgos

notebooks/

Notebooks Jupyter organizados secuencialmente:

1. **01_exploracion_dataset.ipynb**: Análisis exploratorio inicial del dataset
2. **02_analisis_diversidad_consumo.ipynb**: Análisis de diversidad en el consumo de contenido
3. **03_validacion_mitigacion.ipynb**: Validación de estrategias de mitigación
4. **04_efectividad_campanas.ipynb**: Análisis de efectividad de campañas publicitarias
5. **05_taxonomia_tipos_sesgo.ipynb**: Taxonomía y clasificación de tipos de sesgo

results/

Resultados generados por el pipeline:

- **figures/**: 17 visualizaciones en formato PNG (análisis demográficos, métricas de equidad, comparaciones de estrategias, etc.)
- **tables/**: 15 tablas en formato CSV y JSON con métricas detalladas
- **models/**: 10 modelos entrenados guardados en formato PKL
- **reports/**: Informes en formato PDF (`informe_auditoria_sesgo.pdf`)

scripts/

Scripts auxiliares para tareas de mantenimiento:

- **generate_checksum.ps1**: Script PowerShell para generar checksums SHA256 (Windows)
- **generate_checksum.sh**: Script Bash para generar checksums SHA256 (Linux/Mac)

src/

Código fuente Python organizado en módulos:

data_preparation/: Preparación y limpieza de datos

- `load_data.py`: Carga de datos desde archivos CSV
- `clean_data.py`: Funciones de limpieza de datos
- `feature_engineering.py`: Ingeniería de características
- `load_and_clean.py`: Pipeline completo de preparación

modeling/: Modelado predictivo

- `train_model.py`: Entrenamiento de modelos LightGBM
- `evaluate_model.py`: Evaluación de modelos por grupos demográficos

fairness/: Análisis de equidad y sesgos

- `exposure_metrics.py`: Cálculo de métricas de exposición y fairness
- `bias_detection.py`: Detección de sesgos algorítmicos
- `simulate_exposure.py`: Simulaciones controladas de sesgo de exposición
- `mitigation_strategies.py`: Implementación de técnicas de mitigación

A.3. Componentes Principales Implementados

A.3.1. Pipeline de Preparación de Datos

El módulo `data_preparation` implementa un pipeline completo que:

- Carga datos desde archivos CSV originales
- Realiza limpieza básica (tipos de datos, valores nulos, categorías)
- Aplica feature engineering (exposición por usuario, características temporales)
- Guarda datos intermedios y procesados en formato Parquet para eficiencia

A.3.2. Métricas de Exposición y Fairness

El módulo `fairness/exposure_metrics.py` implementa las siguientes métricas:

- **Coeficiente de Gini**: Medida de desigualdad en la distribución de exposición
- **Exposición por Grupo**: Exposición promedio calculada por grupo demográfico
- **Ratio de Exposición**: Comparación de exposición entre grupos
- **Diferencia de Paridad Demográfica**: Métrica de fairness (*demographic parity*)
- **Diferencia de Igualdad de Oportunidades**: Métrica de fairness (*equal opportunity*)

A.3.3. Modelado Predictivo

El módulo `modeling` incluye:

- **Entrenamiento de Modelos**: Implementación con **LightGBM** para predicción de clics
- **Evaluación por Grupo**: Cálculo de métricas (AUC, FPR, FNR, CTR) desagregadas por grupos demográficos
- **Curvas de Confiabilidad**: Visualización de la calibración del modelo por grupo

A.3.4. Simulaciones de Sesgo

El módulo `simulate_exposure.py` permite:

- **Downsampling de Grupos**: Reducir artificialmente las impresiones de un grupo
- **Upsampling de Grupos**: Aumentar artificialmente las impresiones de un grupo
- **Comparación de Escenarios**: Evaluación del impacto de diferentes niveles de sesgo

A.3.5. Estrategias de Mitigación

El módulo `mitigation_strategies.py` implementa tres técnicas principales:

1. **Reweighting**: Ajuste de pesos basado en *propensity scores*
2. **Resampling**: Oversample/undersample de grupos minoritarios usando `imbalanced-learn`
3. **Threshold Optimization**: Post-processing con `fairlearn` para optimizar umbrales de decisión

Referencias

Barocas, S., y Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., y He, X. (2021). *Bias and debias in recommender system: A survey and future directions.* Descargado de https://arxiv.org/abs/2010.03240 doi: 10.1145/3564284

Chen, Y., Li, H., Liu, R., y Wu, Y. (2023, 06). The optimized probabilistic recommendation model in exposure bias. *Applied and Computational Engineering*, *6*, 1507-1521. doi: 10.54254/2755-2721/6/20230929

Gangadharan, K., Purandaran, A., Malathi, K., Subramanian, B., Jeyaraj, R., y Jung, S. K. (2025). From data to decisions: The power of machine learning in business recommendations. *IEEE Access*. doi: 10.1109/ACCESS.2025.3532697

Gupta, S., Wang, H., Lipton, Z. C., y Wang, Y. (2021). *Correcting exposure bias for link recommendation* (Vol. abs/2106.07041). Descargado de https://arxiv.org/abs/2106.07041

Irena, A., Wahyudi, A., Wairooy, I., y Andro, B. (2024, 11). Fairness and bias in e-commerce recommendation systems: A literature review. En (p. 223-228). doi: 10.1109/ICIMCIS63449.2024.10956674

Kidwai, U., Akhtar, D., y Nadeem, M. (2023, 09). Unravelling filter bubbles in recommender systems: A comprehensive review. *International Journal of Membrane Science and Technology*, *10*, 1650-1680. doi: 10.15379/ijmst.v10i2.1839

Krause, T., Deriyeva, A., Beinke, J. H., Bartels, G. Y., y Thomas, O. (2024). The relevance of item-co-exposure for exposure bias mitigation.. Descargado de https://arxiv.org/abs/2409.12912

Mansoury, y Masoud. (2022). *Understanding and mitigating multi-sided exposure bias in recommender systems* (Tesis Doctoral, New York, NY, USA). doi: 10.1145/3566100.3566103

Mansoury, M., y Mobasher, B. (2023). *Fairness of exposure in dynamic recommendation.* Descargado de https://arxiv.org/abs/2309.02322

Narang, U., y Shankar, V. (2019, 09). Mobile marketing 2.0: State of the art and research agenda. En *Marketing in a digital world.* Emerald Publishing Limited. Descargado de https://doi.org/10.1108/S1548-643520190000016008 doi: 10.1108/S1548-643520190000016008

Sharma, y Sachin. (2024). *Unraveling biases and customer heterogeneity in e-commerce recommendation systems* (Disertación doctoral, University of Missouri-St. Louis). Descargado de https://irl.umsl.edu/dissertation/1422/

Sun, L., y Fu, D. (2025). A review of machine learning-based recommendation algorithms in information technology systems. *Journal of Computer, Signal, and System Research*. doi: https://doi.org/10.71222/gvtd3173

Zhu, Y., Ma, J., y Li, J. (2023). *Causal inference in recommender systems: A survey of strategies for bias mitigation, explanation, and generalization.* Descargado de https://arxiv.org/abs/2301.00910