Trabajo Práctico

Equidad en Aprendizaje Automático

2do cuatrimestre 2025

Docente: Mariela Rajngewerc

Alumnos:

- Fraga, Julian
- Miraglia, Morena
- Oliaro, Lucas

Índice

Análisis de los datos

Motivación

Cuando una entidad bancaria recibe una solicitud de préstamo, debe decidir si aprobarla o no basándose en el perfil del solicitante.
Existen dos riesgos principales en esa decisión:

- Si el solicitante **es un buen riesgo crediticio** (es decir, probablemente devolverá el préstamo) y el banco **no aprueba el crédito**, se produce una **pérdida de oportunidad o negocio**.

- Si el solicitante **es un mal riesgo crediticio** (es decir, probablemente no devolverá el préstamo) y el banco **aprueba el crédito**, se genera una **pérdida financiera directa**.

Dado que el segundo riesgo implica un daño económico mayor, el objetivo de la institución crediticia es **minimizar el riesgo total** y **maximizar el beneficio esperado**, estableciendo reglas de decisión adecuadas.

El conjunto de datos **Statlog (German Credit Data)** fue creado para abordar este tipo de problemas de *credit scoring*, utilizando información demográfica, socioeconómica y financiera de solicitantes de préstamos.
Su finalidad es servir como base para construir y comparar modelos predictivos que ayuden a las entidades financieras a decidir **a quién otorgar crédito** de manera más precisa y eficiente.

El dataset fue incluido en el **proyecto europeo Statlog**, orientado a comparar métodos estadísticos y de aprendizaje automático en tareas de clasificación. En el marco de este proyecto, también se introdujo una **matriz de costos** para reflejar la gravedad relativa de los errores de clasificación (penalizando más los falsos positivos, es decir, aprobar a un mal cliente).

Creadores y entidad responsable

El conjunto fue preparado por el **Prof. Dr. Hans Hofmann** del *Institut für Statistik und Ökonometrie*, **Universität Hamburg (Alemania)**, y posteriormente incorporado al repositorio UCI dentro del marco del proyecto **Statlog**, financiado por la Comisión Europea.

Composición del conjunto de datos

Cada instancia representa a **una persona solicitante de crédito bancario**.
El dataset contiene información sobre las características personales, financieras y laborales de cada individuo, utilizadas para predecir si un préstamo debe aprobarse o no.

- **Número de instancias:** 1000.
- **Número de atributos:** 20 (más una variable objetivo).

- **Tipo de datos:** mixtos (numéricos y categóricos).
- **Variable objetivo:**
    - 1 = Buen cliente (Good credit risk).
    - 2 = Mal cliente (Bad credit risk).

Atributos más relevantes (versión CSV legible):

- Edad (numérico).
- Sexo (masculino, femenino).
- Tipo de empleo (0 = no calificado/no residente, 3 = altamente calificado).
- Vivienda (propia, rentada o gratuita).
- Cuentas de ahorro y corriente.
- Monto del crédito (en marcos alemanes).
- Duración del préstamo (en meses).
- Propósito (automóvil, muebles, educación, negocio, vacaciones, etc.).

Además, como ya fue mencionado anteriormente, el dataset incluye una **matriz de costos** para modelar el impacto de los errores de predicción.

Proceso de recopilación

Los datos, como fue mencionado anteriormente, fueron recolectados y preparados por el Prof. Hofmann a partir de solicitudes de crédito bancario en Alemania (probablemente reales o anonimizadas).
No se especifica el banco de origen ni el procedimiento exacto de obtención.
Cada registro refleja características observadas directamente del solicitante, no inferidas ni autorreportadas.
No se incluyen identificadores personales ni variables de localización.

Preprocesamiento, limpieza y etiquetado

- El archivo original ("german.data") usa variables simbólicas (por ejemplo, A11, A12).
- La **Universidad de Strathclyde** generó la versión "german.data-numeric", reemplazando categorías por valores numéricos e indicadores.
- No se reportan valores faltantes.
- Se aplicaron codificaciones ordinales y binarias para variables categóricas.
- Se añadió la **matriz de costos** que penaliza los falsos positivos más que los falsos negativos.

Usos

El dataset se utiliza ampliamente para:

- **Modelos de clasificación supervisada** (regresión logística, análisis discriminante, árboles de decisión, random forest).
- **Estudios de equidad algorítmica**, especialmente para analizar posibles sesgos por género o edad en la aprobación de créditos.

- **Enseñanza en minería de datos, aprendizaje automático y ética de IA**.

Ejemplos de uso académico:

- Kamiran & Calders (2012). *Data Preprocessing Techniques for Classification Without Discrimination.*
- Mehrabi et al. (2022). *Algorithmic Fairness Datasets: The Story So Far.*
- TensorFlow Datasets – german_credit_numeric.

Sesgos potenciales y equidad

El dataset Statlog (German Credit Data) refleja decisiones de crédito tomadas históricamente por entidades bancarias. Como tal, puede incorporar sesgos entendidos como desigualdades presentes en los datos como:

Sesgos Históricos:

- Edad, nivel educativo y empleo
    - Personas más jóvenes, con menor nivel educativo o con empleos menos calificados podrían haber sido sistemáticamente subestimadas en la evaluación de riesgo crediticio, independientemente de su capacidad de pago real.
- Género
    - Históricamente, las mujeres podían tener menor acceso a crédito debido a normas sociales o criterios bancarios que favorecían a los hombres. Este fenómeno es un caso específico de sesgo histórico y es el que se analiza en este trabajo, dado que representa una desigualdad relevante y medible en el dataset.

Sesgo de representación:

- El dataset no está balanceado por género: aproximadamente 70% de los registros corresponden a hombres y 30% a mujeres. Este desequilibrio puede influir en el entrenamiento del modelo, que tiende a optimizar su desempeño para el grupo mayoritario, mientras que el grupo minoritario podría presentar recall o precisión inferiores.

Creación del modelo inicial

Para predecir si un solicitante debía recibir crédito o no, se entrenó un modelo de regresión logística, utilizando el conjunto de datos Statlog (German Credit Data).

Preprocesamiento de los datos

El dataset contiene variables numéricas (edad, monto del crédito, duración del préstamo) y categóricas (sexo y estado civil, tipo de empleo, vivienda, propósito del crédito, cuentas de ahorro).

El primer paso en el procesamiento fue extraer el atributo que consideramos sensible (género) del atributo “sexo y estado civil”, separando esta variable en dos: una binaria y otra categórica.

Luego, las variables categóricas fueron transformadas mediante One Hot Encoding para permitir que el modelo procese correctamente la información no numérica.

Por su parte, las variables numéricas fueron estandarizadas para mejorar la estabilidad y la convergencia del modelo.

Además, se consideró la matriz de costos original del dataset, que penaliza cinco veces más los errores de aprobar a un mal cliente que los de rechazar a uno bueno.

Entrenamiento y evaluación

El conjunto de datos se dividió en entrenamiento y prueba, ajustando el modelo sobre el primero.
Para evaluar el desempeño, se calcularon métricas clásicas de clasificación:

- Accuracy: proporción de predicciones correctas sobre el total.
- Precisión: proporción de predicciones positivas correctas sobre todas las predicciones positivas.
- Recall: proporción de verdaderos positivos correctamente identificados sobre todos los casos positivos reales.
- F1-score: promedio armónico entre precisión y recall, balanceando ambos aspectos.

También se construyó la matriz de confusión para analizar la distribución de verdaderos y falsos positivos y negativos, tanto en general como para cada grupo de género.

Resultados generales del modelo inicial

	Precisión	**Recall**	**F1-Score**	**Support**
Bad	0.43	0.90	0.58	60
Good	0.92	0.49	0.64	140

El modelo mostró un desempeño moderado, con una accuracy global de 0.61.
Para la clase “good” (buen riesgo crediticio), la precisión fue 0.92 y el recall 0.49, mientras que para la clase “bad” la precisión fue 0.43 y el recall 0.90.
Estos resultados indican que el modelo es excesivamente conservador: aprueba pocos préstamos, asegurando que la mayoría sean clientes confiables, pero a costa de rechazar a muchas personas que en realidad sí podrían pagar.

Dado que el objetivo del banco es maximizar la cantidad de personas que efectivamente reciban el préstamo y lo paguen, el error más grave es rechazar a un cliente que sí puede pagar (falso negativo para la clase “good”).
Ese tipo de error representa una pérdida de oportunidad de negocio, ya que se dejan de otorgar créditos a personas solventes que habrían generado ingresos para la institución.

Por tanto, un recall bajo en la clase “good” evidencia un punto clave a mejorar: el modelo debería aumentar la tasa de aprobación entre los buenos clientes sin comprometer en exceso el riesgo.

En este sentido, el modelo cumple con la intención de reducir el riesgo crediticio, pero no con el objetivo institucional de maximizar la aprobación de créditos exitosos, mostrando un sesgo hacia la prudencia que limita su utilidad comercial.

Resultados por subgrupo de género

Género		Precisión	Recall	F1-Score	Support
Mujeres					
	Bad	0.42	0.85	0.57	20
	Good	0.85	0.42	0.57	40
Hombres					
	Bad	0.44	0.93	0.59	40
	Good	0.95	0.52	0.67	100

El modelo tiende a clasificar con menor probabilidad como “buenas” a las mujeres, incluso cuando efectivamente lo son.

Esto sugiere que el modelo podría estar replicando sesgos históricos y de representación presentes en los datos, donde las mujeres tienen menor proporción de registros y, por lo tanto, un ajuste menos preciso.

Evaluación de equidad del modelo inicial

Para evaluar la equidad del modelo entre hombres y mujeres se aplicaron los principales criterios de fairness:

- **Statistical Parity**: Este criterio indica que la probabilidad de que un individuo reciba un préstamo aprobado debe ser la misma para todos los grupos, independientemente de su género. En otras palabras, un modelo cumple SP si la proporción de mujeres y hombres que reciben un crédito es aproximadamente igual. Si bien es un criterio intuitivo y fácil de calcular, en el contexto crediticio puede no ser suficiente, ya que no considera si los individuos realmente tienen capacidad de pago; es decir, podría priorizar la igualdad de acceso sin considerar el riesgo crediticio.
- **Equal Opportunity:** Este criterio se centra exclusivamente en los verdaderos positivos, es decir, aquellas personas que efectivamente pagarán el préstamo. Un modelo cumple Equal Opportunity si la probabilidad de aprobar un crédito para quienes efectivamente lo pagarán es la misma entre mujeres y hombres. Este criterio es especialmente relevante para el banco, ya que maximizar la aprobación entre clientes solventes es coherente con su objetivo de rentabilidad y seguridad financiera.

- **Equalized Odds:** Este criterio amplía Equal Opportunity considerando tanto los verdaderos positivos como los falsos positivos. Para que se cumpla Equalized Odds, la probabilidad de aprobar un crédito debe ser la misma entre géneros tanto para quienes efectivamente pagarán como para quienes no pagarán. Este criterio es más estricto, ya que controla la equidad en ambos tipos de error, y puede ser útil para reducir discriminación en los falsos positivos (aprobar créditos a personas que no los pagarán) y falsos negativos (rechazar créditos a quienes sí pagarían).
- **Predictive Parity**: Este criterio se enfoca en la precisión del modelo. Un modelo cumple Predictive Parity si, dado que se aprueba un crédito, la probabilidad de que efectivamente se pague es la misma entre mujeres y hombres. Esto asegura que las decisiones de aprobación sean igualmente confiables para todos los grupos.

Umbral de equidad

Para determinar si el modelo puede considerarse equitativo (*fair*) bajo cada criterio, se estableció un umbral de disparidad de 0.05. Esto significa que, si la diferencia absoluta entre grupos (por ejemplo, hombres y mujeres) es menor o igual a 0.05, se interpreta que el modelo no presenta una desigualdad estadísticamente relevante. En cambio, valores mayores a 0.05 indican que el modelo trata de manera significativamente distinta a los grupos, por lo que se considera no equitativo (*not fair*) para ese criterio.

Resultados de equidad modelo inicial

	Criterio	**Disparidad Absoluta**	**Umbral**	**Cumple Fairness**
0	Equality of Opportunity Difference	0.095	0.05	Not Fair
3	Accuracy Difference	0.069	0.05	Not Fair
1	False Positive Rate Difference	0.075	0.05	Not Fair
2	Average Odds Difference	0.010	0.05	Fair

Reflexión sobre el modelo inicial

Desde la perspectiva del banco, el criterio de equidad más relevante es **Equal Opportunity**, ya que refleja directamente la posibilidad de que personas que efectivamente pueden pagar el crédito sean aprobadas sin discriminación por género.Sin embargo, la disparidad observada (0.095) evidencia que el modelo actual perjudica sistemáticamente a las mujeres, reduciendo sus probabilidades de acceso al crédito incluso cuando presentan buen perfil financiero.

Este sesgo no sólo tiene implicancias éticas, sino también económicas, ya que el banco podría estar perdiendo oportunidades de otorgar créditos rentables a clientas solventes.Por tanto, resulta necesario aplicar técnicas de mitigación de sesgos, como el reweighting, para ajustar el modelo hacia decisiones más equitativas y alineadas con el objetivo institucional de maximizar la cantidad de préstamos otorgados a personas capaces de devolverlos.

Mitigación de sesgos

Tras evidenciar que el modelo inicial presentaba disparidades de género, en especial una menor probabilidad de aprobación para mujeres con buen perfil crediticio, se aplicaron técnicas de mitigación de sesgos con el objetivo de mejorar la equidad sin deteriorar el desempeño predictivo. Estas estrategias buscan ajustar el modelo o los datos de entrenamiento para reducir la influencia de sesgos históricos o de representación, promoviendo decisiones más justas y consistentes con los objetivos institucionales del banco.

En este trabajo se implementaron y compararon dos enfoques complementarios: Reweighting y Correlation Remover.

Mitigación de sesgos mediante Reweighting

La técnica de reweighting (preprocesamiento) ajusta los pesos de las instancias de entrenamiento de modo que las combinaciones entre el atributo protegido (género) y la variable objetivo (riesgo crediticio) se vuelvan más balanceadas. En otras palabras, otorga mayor peso a los casos subrepresentados (por ejemplo, mujeres “buenas pagadoras”) y menor a los sobrerrepresentados, reduciendo así la influencia de patrones históricos injustos.

Tras aplicar este método, el modelo fue entrenado con la misma arquitectura de regresión logística y evaluado con las métricas clásicas de clasificación y fairness

Resultados métricas generales

	Precisión	**Recall**	**F1-Score**	**Accuracy**
Good	0.92	0.52	0.67	0.63
Bad	0.44	0.88	0.59	

Resultados métricas fairness

	Criterio	Disparidad Absoluta	Umbral	Cumple Fairness
0	Equality of Opportunity Difference	0.040	0.05	Fair
3	Accuracy Difference	0.0024	0.05	Fair
1	False Positive Rate Difference	0.15	0.05	Not Fair
2	Average Odds Difference	0.09	0.05	Not Fair

Reflexiones modelo Reweighting

El modelo ajustado mediante Reweighting logra un avance significativo en términos de equidad. La mejora en la Equality of Opportunity Difference (de 0.095 a 0.040) indica que ahora hombres y mujeres con buen perfil crediticio tienen prácticamente la misma probabilidad de ser aprobados. Esta métrica es clave desde la perspectiva del banco, ya que refleja directamente la capacidad del modelo para identificar clientes solventes sin sesgos de género. Además, la Accuracy Difference se reduce casi a cero, lo que sugiere que el desempeño general del modelo es consistente entre grupos.

Por otro lado, aunque la False Positive Rate Difference aumentó (de 0.075 a 0.15), este cambio puede interpretarse como una consecuencia del reequilibrio: al aumentar la aprobación para mujeres, también se incrementa el riesgo de aprobar a algunas clientas que no pagarán. Sin embargo, este riesgo adicional parece estar controlado, ya que la precisión para la clase “good” se mantiene alta (0.92) y el recall mejora (de 0.49 a 0.52), lo que implica una mayor capacidad para detectar oportunidades de negocio.

Mitigación de sesgos mediante Correlation Remover

La técnica Correlation Remover busca eliminar o reducir la correlación que existe entre el atributo protegido (en nuestro caso el género), se diferencia de Reweighting ya que actúa directamente sobre las variables predictores, en vez de ajustar los pesos, quitando la correlación. Se espera que con esta técnica, las correlaciones históricas entre el género y la obtención del crédito tengan menor influencia al momento de otorgar un crédito.

Tras aplicar este método, el modelo fue entrenado con la misma arquitectura de regresión logística y evaluado con las métricas clásicas de clasificación y fairness

Resultados métricas generales

	Precisión	Recall	F1-Score	Accuracy
Good	0.93	0.53	0.67	0.64
Bad	0.45	0.87	0.59	

Resultados métricas fairness

	Criterio	Disparidad Absoluta	Umbral	Cumple Fairness
0	Equality of Opportunity Difference	0.065	0.05	Not Fair
3	Accuracy Difference	0.0014	0.05	Fair
1	False Positive Rate Difference	0.1500	0.05	Not Fair
2	Average Odds Difference	0.1075	0.05	Not Fair

El ajuste nos da una reducción en todos los indicadores, logró una mejora importante en Accuracy Difference, reduciendo de 0.069 a 0.014, dando un comportamiento similar para ambos grupos. Sin embargo, ninguna otra métrica se ubica dentro del umbral de equidad (0.05). No se logró eliminar completamente las diferencias entre los grupos

Reflexión modelo Correlation Remover

El modelo ajustado con Correlation Remover también muestra mejoras respecto al baseline, aunque más moderadas en términos de equidad. La Equality of Opportunity Difference se reduce de 0.095 a 0.065, lo que indica una mejora, pero aún fuera del umbral de equidad definido (0.05). Esto significa que persiste una diferencia significativa en la probabilidad de aprobación entre hombres y mujeres con buen perfil crediticio.

Por otro lado, la Accuracy Difference es muy baja (0.0014), lo que sugiere que el modelo tiene un desempeño global muy similar entre géneros. Esta consistencia puede ser valiosa para el banco, ya que garantiza que el modelo no favorece sistemáticamente a un grupo en términos de rendimiento general. Sin embargo, las métricas más sensibles, como la Average Odds

Difference y la False Positive Rate Difference, siguen mostrando disparidades importantes, lo que limita el impacto positivo de esta técnica en decisiones individuales de crédito.

Comparación entre modelos

Al recopilar los resultados de los distintos modelos obtenemos la siguiente tabla, en la que la mayoría de métricas se ven mejoradas por la mitigación de sesgos sobre el *baseline*.

Metric	**Baseline**	**Preprocessing Mitigator Reweighting**	**Preprocessing Mitigator Correlation Remover**	**Ref.**
Statistical Parity	0.059524	0.054762	0.071429	0
Disparate Impact	0.848485	1,144654	1.188679	1
Four Fifths Rule	0.848485	0.873626	0.841270	1
Cohen D	0.123139	0.112175	0.146137	0
2SD Rule	0.796819	0.725984	0.944911	0
Equality of Opportunity Difference	0.095000	0.040000	0.065000	0
False Positive Rate Difference	0.075000	0.150000	0.150000	0
Average Odds Difference	0.010000	0.095000	0.107500	0

Accuracy Difference	0.069048	0.002381	0.014286	0

En términos de métricas generales, los tres modelos muestran una precisión elevada para la clase “good” (clientes que efectivamente pagan), lo que indica que las decisiones de aprobación son confiables. Sin embargo, el recall, es decir, la capacidad de detectar a todos los buenos clientes, es más bajo en el modelo baseline (0.49) y mejora levemente en los modelos ajustados: 0.52 con Reweighting y 0.53 con Correlation Remover. Esto implica que los modelos con mitigación recuperan oportunidades de negocio que el modelo original dejaba pasar, permitiendo que más personas solventes accedan al crédito. La accuracy también mejora: pasa de 0.615 en el baseline a 0.635 con Reweighting y 0.64 con Correlation Remover. Aunque la diferencia es moderada, en contextos bancarios incluso pequeñas mejoras pueden traducirse en impactos económicos significativos, especialmente cuando se trata de decisiones que afectan directamente la rentabilidad de la institución.

Desde la perspectiva del banco, el criterio de equidad más relevante es Equality of Opportunity, ya que refleja la posibilidad de que personas que efectivamente pueden pagar el crédito sean aprobadas sin discriminación por género. En el modelo baseline, la disparidad observada en esta métrica (0.095) evidencia que el sistema perjudicaba sistemáticamente a las mujeres, reduciendo sus probabilidades de acceder al crédito incluso con un perfil financiero favorable. Este sesgo no solo tiene implicancias éticas, al reproducir desigualdades históricas, sino también económicas, ya que el banco podría estar perdiendo oportunidades de otorgar créditos rentables a clientas solventes.

El modelo con Reweighting logra reducir esta disparidad a 0.040, ubicándose dentro del umbral de equidad definido en el trabajo (0.05). Además, la diferencia de accuracy entre géneros se reduce casi a cero (0.002), lo que indica que el modelo tiene un desempeño equilibrado. Aunque la False Positive Rate Difference aumenta (0.15), este efecto puede considerarse un costo aceptable frente a la ganancia en inclusión y justicia. El modelo ajustado mediante Reweighting logra un balance más justo entre géneros, manteniendo un desempeño adecuado en la identificación de buenos y malos clientes. Desde una perspectiva institucional, este resultado es sumamente positivo: se reduce el riesgo de rechazar a buenas clientas sin comprometer la seguridad financiera del banco. Esto demuestra que la equidad puede alcanzarse sin pérdida de rendimiento predictivo, lo que refuerza la idea de que los modelos justos también pueden ser eficientes.

Aunque Reweighting y Correlation Remover mejoran significativamente la equidad en métricas como Equality of Opportunity y Accuracy Difference, ambos modelos presentan una False Positive Rate Difference elevada (0.15), lo que indica que podrían estar aprobando más créditos a personas que no los pagarán. Este efecto puede representar un riesgo financiero adicional para el banco. Sin embargo, considerando que la precisión para la clase “good” se mantiene alta y que el recall mejora, este costo puede interpretarse como un compromiso razonable frente a la ganancia en inclusión y justicia.

Si bien Correlation Remover mejora la uniformidad del desempeño global (Accuracy Difference de 0.0014), no logra cumplir con el umbral de equidad en la métrica clave de Equality of Opportunity (0.065). Esto indica que aún persiste una diferencia significativa en la probabilidad de aprobación entre géneros para clientes solventes. Por lo tanto, si bien es útil para reducir correlaciones históricas, no alcanza el mismo nivel de equidad que Reweighting, lo que limita su aplicabilidad en contextos donde la justicia en decisiones individuales es prioritaria.

Impacto en la vida real

Los resultados obtenidos en este trabajo no solo tienen relevancia técnica, sino también implicancias concretas en el mundo real. En el contexto bancario, donde las decisiones automatizadas afectan directamente el acceso al crédito, la equidad de los modelos se traduce en consecuencias tangibles para las personas.

Cuando un modelo presenta sesgos de género, como se evidenció en el baseline, puede rechazar sistemáticamente a mujeres con buen perfil financiero. Esto implica que clientas solventes quedan excluidas del sistema crediticio, limitando su capacidad de inversión, emprendimiento o mejora de calidad de vida. En muchos casos, el acceso al crédito es una herramienta clave para el desarrollo personal y económico, por lo que una decisión injusta puede perpetuar desigualdades estructurales y restringir oportunidades. Por el contrario, un modelo más equitativo, como el ajustado mediante Reweighting, permite que las decisiones de aprobación reflejen mejor la capacidad real de pago de cada persona, sin que el género influya negativamente. Esto promueve una mayor inclusión financiera, permitiendo que más personas accedan a recursos que les permitan emprender, estudiar, mejorar su vivienda o afrontar situaciones imprevistas. En términos sociales, esto contribuye a reducir brechas históricas y a construir un sistema financiero más justo y accesible.

Desde el punto de vista institucional, aplicar modelos equitativos también tiene beneficios concretos. Un sistema que reconoce correctamente a los buenos clientes, sin sesgos, mejora la rentabilidad del banco al reducir el rechazo de perfiles solventes. Además, fortalece la reputación de la entidad, mostrando compromiso con prácticas responsables y alineadas con principios éticos. En un entorno regulatorio cada vez más exigente, contar con modelos auditables y equitativos también reduce riesgos legales y reputacionales.

En definitiva, la equidad en modelos de machine learning no es solo una cuestión técnica o académica, sino una necesidad práctica. Impacta directamente en la vida de las personas, en la eficiencia del negocio y en la sostenibilidad de las decisiones automatizadas. Por eso, integrar criterios de fairness y aplicar técnicas de mitigación de sesgos es un paso fundamental para construir sistemas que funcionen bien y de manera justa para todos los grupos sociales.

Conclusiones

A lo largo del trabajo se evidenció que el modelo inicial presentaba disparidades de género significativas, especialmente en la probabilidad de aprobación para mujeres con buen perfil crediticio. Este sesgo, además de ser éticamente cuestionable, representa una pérdida de oportunidades de negocio para el banco.

La aplicación de técnicas de mitigación de sesgos en la etapa de preprocesamiento permitió mejorar la equidad del modelo sin comprometer su desempeño. Reweighting se destacó por reducir la disparidad en Equality of Opportunity por debajo del umbral definido, logrando un comportamiento más justo entre géneros. Correlation Remover, si bien no alcanzó el mismo nivel de equidad, mejoró la uniformidad del desempeño global y aportó valor en términos de transparencia.

Los resultados muestran que es posible construir modelos que sean al mismo tiempo eficientes y equitativos. En contextos reales, esto se traduce en decisiones crediticias más inclusivas, mayor confianza en los sistemas automatizados y una mejora en la rentabilidad institucional. La equidad, lejos de ser un ideal abstracto, se consolida como un criterio técnico y estratégico fundamental para el desarrollo de sistemas de aprendizaje automático responsables.