Universidad Nacional de San Martín

Equidad en aprendizaje automático

TP FINAL

Integrantes:

Arnesano, Gabriel Santiago
Crespi, María Sol
Darnes Pallitto, Santiago
Vidal, Tomas Ramón

1) B)

1. Motivación

Este dataset fue creado con la idea de estudiar cómo un banco podría usar modelos automatizados para decidir si aprobar o no un crédito. O sea, la idea era ver si se podía entrenar un sistema que clasifique a las personas como "buenos" o "malos" pagadores, y así ayudar a tomar decisiones de crédito.

No se armó desde cero para una investigación puntual, sino que se usaron datos reales de un banco regional del sur de Alemania. Más tarde, estos datos fueron compartidos como parte del proyecto europeo Statlog, que se enfocaba en comparar el rendimiento de distintos algoritmos de clasificación. Así que se podría decir que fue creado con fines académicos y prácticos, pensando en tareas de clasificación binaria.

Fuente:

- Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Statlog (German Credit Data). University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

- Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. Data Mining and Knowledge Discovery, 36(5), 1534–1582. https://doi.org/10.1007/s10618-022-00854-z

2. Composición

Cada fila del dataset representa una persona que pidió un préstamo. En total hay 1000 personas, y cada una tiene 20 características: algunas personales (edad, género, estado civil), otras relacionadas al crédito (duración, monto solicitado, historial) y algunas sobre su situación económica (empleo, tipo de vivienda, etc.).

El objetivo es predecir si esa persona es "buena" o "mala" pagadora, que está representado por una variable target binaria. Así que, en resumen, lo que tenés es un dataset con gente que pidió crédito y cómo terminó su situación: si pagó bien o no.

Fuente:

- Dua, D., & Graff, C. (2019). Statlog (German Credit Data) Data Set. UCI Machine Learning Repository. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

3. Proceso de recolección

Los datos no se inventaron, vienen de solicitudes de crédito reales hechas entre 1973 y 1975 en un banco del sur de Alemania. No se detallan los métodos exactos de cómo se recolectaron (por ejemplo, si fue digital o en papel), pero se sabe que fueron decisiones reales y no simuladas.

Un detalle importante es que los datos fueron "balanceados" artificialmente para que haya más casos de mal crédito de los que habría en la realidad. Esto lo hicieron para que los modelos puedan aprender mejor a detectar los malos pagadores, ya que en general son menos. Esto se llama muestreo estratificado.

Fuente:

- Dua, D., & Graff, C. (2019). Statlog (German Credit Data) Data Set. UCI Machine Learning Repository. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

4. Preprocesamiento / limpieza / etiquetado

Sí, hubo preprocesamiento. La versión más conocida del dataset (la que está en el UCI) tiene todos los atributos ya codificados como números, incluso los que eran categóricos, para que se pueda usar directamente en modelos.

Por ejemplo, el estado civil o el tipo de trabajo están representados con códigos numéricos. Además, los valores faltantes ya fueron eliminados, o sea que no vas a tener problemas con los nulos. En algunos trabajos también se ve que se renombran las columnas, ya que los nombres originales son bastante crípticos (como A11, A12, etc.).

Fuente:

- Papers with Code. (s.f.). German Credit Dataset. Retrieved May 29, 2025, from https://paperswithcode.com/dataset/german-credit-dataset

5. Usos

Este dataset se usa muchísimo en investigación, sobre todo para temas de clasificación binaria y fairness. Aparece en trabajos que buscan mejorar modelos de scoring crediticio, pero también en papers que estudian el sesgo en algoritmos, especialmente en género.

Es bastante conocido porque tiene un atributo relacionado al género (aunque de forma codificada), entonces permite estudiar si los modelos tratan de forma distinta a hombres y mujeres.

Hay varios papers que usan este dataset para ver si los modelos discriminan, y también se usa como benchmark para probar métodos de mitigación de sesgos.

Fuente:

- 1. Ramírez López, L. F., & Rebaque Ruiz, M. (2023). Comparative study of bias mitigation techniques for credit scoring in real-world scenarios. In D. Sánchez, J. M. de la Cruz, & A. Bellogín (Eds.), Proceedings of the 1st Workshop on Fairness and Transparency in Artificial Intelligence – FTAI 2023 (Vol. 3908, pp. 143–154). CEUR-WS. http://ceur-ws.org/Vol-3908/paper_15.pdf
- 2. Dua, D., & Graff, C. (2019). Statlog (German Credit Data) Data Set. UCI Machine Learning Repository. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

C) (Realizado en la notebook)

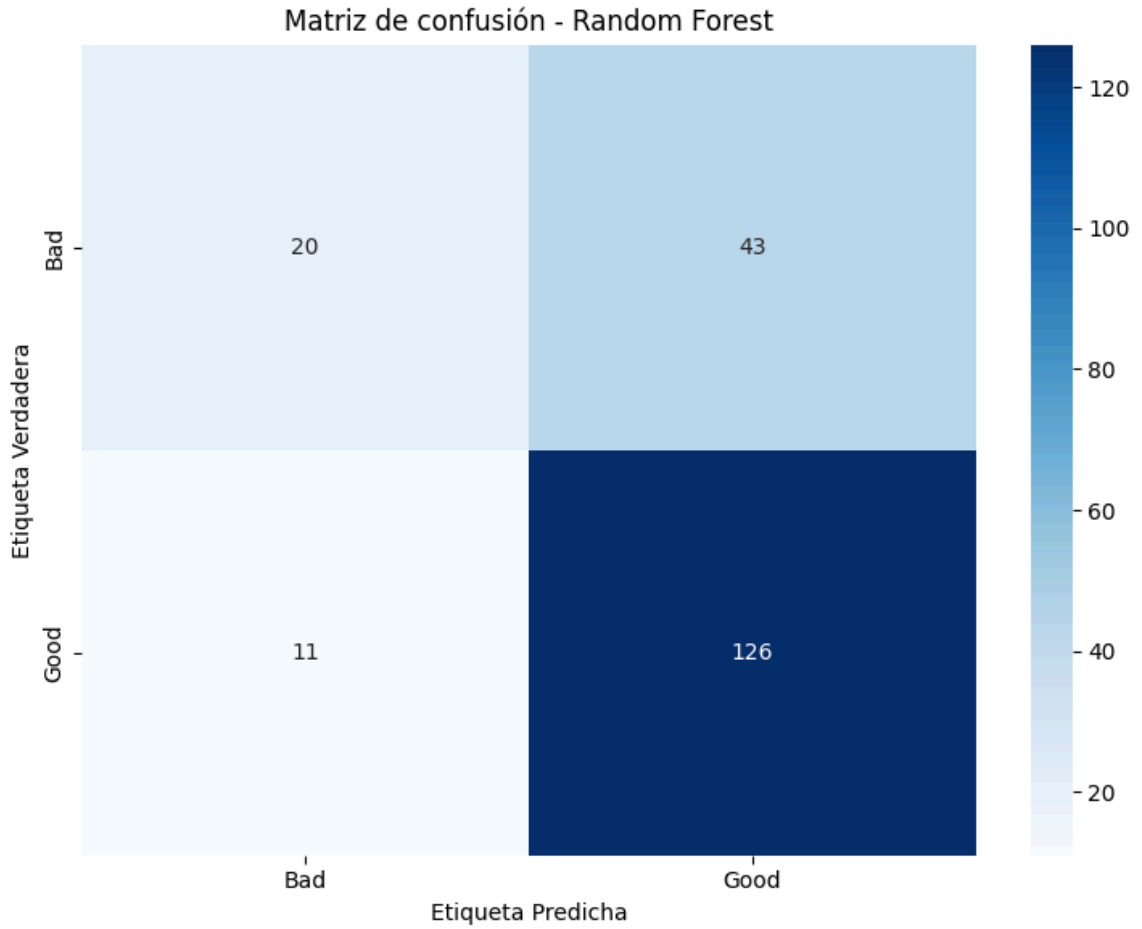
D)

A partir del análisis preliminar que realizamos sobre el dataset, los posibles sesgos que identificamos son:

- Sesgo de representación e histórico: la variable ‘Personal status and sex’ combina el estado civil con el sexo de las personas y presenta una distribución desigual entre los hombres y las mujeres. En particular, se observa una menor representación de mujeres solteras en comparación con el grupo de los hombres solteros. Esto podría indicar un sesgo de representación y un posible sesgo histórico, ya que los datos reflejan decisiones pasadas que pudieron estar influenciadas por normas sociales discriminatorias.
- Sesgo de discriminación sistemática: la variable ‘Foreing worker’ distingue si los trabajadores son extranjeros o no. En el análisis previo se puede ver que hay una menor proporción de créditos aprobados para los trabajadores extranjeros en comparación con el grupo de los trabajadores nacionales. Este es un posible caso de discriminación sistemática, ya que las decisiones crediticias previas que podrían haber penalizado a personas por su origen, sin relación directa con su capacidad de pago.
- Sesgo histórico: todos los registros del dataset reflejan decisiones humanas pasadas sobre si otorgar o no un crédito. Si estas decisiones estuvieron influenciadas por prejuicios o discriminación, esos sesgos se trasladan al dataset y luego al modelo, generando un sesgo histórico y una posible automatización de decisiones injustas.

2) A) MÉTRICAS Y MATRIZ DE CONFUSIÓN

Se realizó como modelo un Random Forest, que dió esta matriz de confusión y métricas.



- Accuracy: 0.735
- Precision: 0.745
- Recall: 0.919
- F1 Score: 0.825

Podemos observar que tenemos:

True Positives: 126
True Negatives: 20
False Positives: 44
False Negatives: 11

También obtuvimos un Recall alto, por lo que el modelo tiene un buen desempeño al identificar correctamente los positivos. Pero por otro lado, tenemos una cantidad relativamente alta de falsos positivos (FPR aproximadamente del 66%), dejando así un porcentaje elevado de casos negativos clasificados como positivos.

B) Desde el punto de vista del banco, el error más grave es el falso positivo (darle un préstamo a alguien que no paga), ya que implica una pérdida directa de dinero. En cambio, un falso negativo implica dejar pasar una oportunidad de negocio, pero sin pérdida concreta.

Por eso, aunque queremos balancear todas las métricas, es clave que el modelo tenga alta precisión en la clase positiva (buen crédito), aunque eso implique sacrificar un poco de recall, y minimizar el False Positive Rate.

. 3) A)

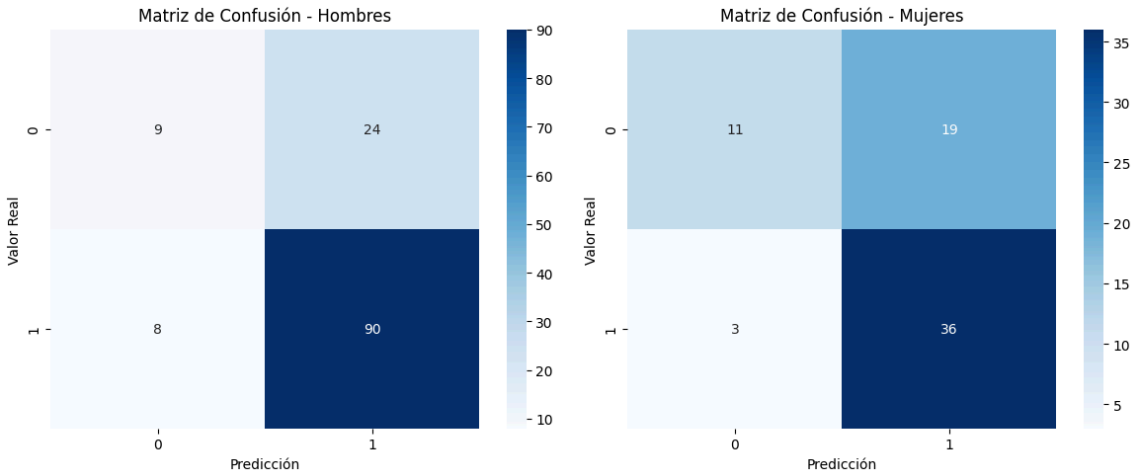
Statistical Parity:
Este criterio implica que la probabilidad de que una persona sea aprobada para el crédito debería ser igual para todos los géneros, sin importar si efectivamente pueden pagar o no el préstamo. En este caso, significa que hombres y mujeres deberían tener la misma tasa de aprobación general. El problema es que no tiene en cuenta si la persona realmente califica o no, solo se enfoca en que las decisiones estén balanceadas entre los grupos.

Equalized Odds:
Este criterio busca que el modelo tenga el mismo comportamiento en términos de errores para cada género. Es decir, tanto la tasa de falsos positivos (aprobar a quien no debería) como la de falsos negativos (rechazar a quien sí debería) deberían ser similares entre hombres y mujeres. En este contexto, asegura que el modelo no sea más riguroso o permisivo con un género que con otro.

Equal Opportunity:
Este criterio es una versión más relajada del anterior. Se enfoca solo en que, entre las personas que realmente pueden pagar el préstamo (es decir, que pertenecen a la clase positiva), la tasa de aprobación sea igual entre géneros. En este problema, significa que hombres y mujeres que tienen un buen perfil crediticio deberían tener la misma probabilidad de ser aprobados. Es un criterio más justo desde el punto de vista práctico, porque no exige forzar aprobaciones, sino tratar igual a quienes realmente califican.

Predictive Parity:
Este criterio dice que la precisión del modelo (es decir, la proporción de personas aprobadas que realmente pagan el préstamo) debería ser igual para hombres y mujeres. En otras palabras, si el modelo aprueba a alguien, la confianza en que esa persona efectivamente va a pagar debería ser la misma sin importar su género.

B) Basándonos en las matrices de confusión que tenemos, evaluamos si el modelo inicial es fair calculando las métricas anteriores. Teniendo en cuenta que elegimos un umbral del 0,1



Métricas para Hombres:
Accuracy: 0.7557
Precision: 0.7895
Recall: 0.9184
F1-Score: 0.8491

Métricas para Mujeres:
Accuracy: 0.6812
Precision: 0.6545
Recall: 0.9231
F1-Score: 0.7660

Statistical Parity:

- Hombres: (90 + 24) / (9 + 24 + 8 + 90) = 114 / 131 ≅ 0.8702
- Mujeres: (36 + 19) / (11 + 19 + 3 + 36) = 55 / 69 ≅ 0.7971
  Por lo que nos queda:
  |0.8702 - 0.7971| ≅ 0.0731 < 0.1 → Se cumple

Por lo que podemos concluir que sí se cumple Statistical Parity.

Equalized Odds:

- TPR(hombres) = 90 / (90 + 8) = 90 / 98 ≅ 0.918
- TPR(mujeres) = 36 / (36 + 3) = 36 / 39 ≅ 0.923
  |0.918 - 0.923| ≅ 0.005 < 0.1 → Se cumple
- FPR(hombres) = 24 / (24 + 9) = 24 / 33 ≅ 0.7273
- FPR(mujeres) = 19 / (19 + 11) = 19 / 30 ≅ 0.6333
  |0.7273 - 0.6333| ≅ 0.094 < 0.1 → Se cumple

Por lo que podemos concluir que se cumple Equalized Odds.

Equal Opportunity:

- Solo compara TPR:

- Ya calculado: |0.918 - 0.923| ≅ 0.005 < 0.1 → Se cumple

Como vimos en cuando analizamos Equalized Odds, el módulo de la diferencia es menor al umbral por lo que sí se cumple Equal Opportunity.

Predictive Parity:

- PPV(hombres) = 90 / (90 + 24) = 90 / 114 ≅ 0.7895
- PPV(mujeres) = 36 / (36 + 19) = 36 / 55 ≅ 0.6545

|0.7895 - 0.6545| ≅ 0.135 > 0.1 → No se cumple

Al ser valores iguales, podemos concluir que no se cumple Predictive Parity.

C) Desde el rol del banco, nos interesa que el modelo sea justo sin comprometer la calidad de las decisiones. Como planteamos anteriormente, nuestro enfoque principal es minimizar el FPR, ya que esta clase de errores son los más perjudiciales para nuestro negocio.

En términos de fairness, una métrica que tiene en cuenta el FPR es Equalized Odds, en la cual se tiene

$$P(\hat{Y} = 1|A = a, Y = y) = P(\hat{Y} = 1|A = b, Y = y) \ \forall \ a, b \in A, y \in \{0,1\}$$

Desde el punto de vista del negocio y buscando mantener la equidad en el modelo para las clases sensibles, esta métrica es una opción interesante para trabajar y tratar de minimizar.

4) A) En esta etapa aplicamos dos técnicas de mitigación de sesgos en la fase de preprocesamiento: Reweighting y Correlation Remover, ambas vistas en clase. Estas técnicas permiten intervenir en los datos antes del entrenamiento del modelo, con el objetivo de reducir los sesgos detectados en el análisis de fairness sin comprometer el rendimiento predictivo.

Reweighting:

La técnica de Reweighting ajusta los pesos asignados a cada observación en función de su pertenencia al grupo sensible y su clase verdadera. La idea es que, al alterar la importancia relativa de cada muestra durante el entrenamiento, el modelo se vea incentivado a aprender patrones menos sesgados, compensando desequilibrios estructurales del dataset original.

Correlation Remover:

Correlation Remover busca eliminar o reducir la correlación entre la variable sensible y el resto de las variables predictoras. Esto es especialmente útil cuando la variable sensible está indirectamente asociada con otras que afectan la predicción, como el tipo de empleo o el historial crediticio. Al eliminar esta correlación, se limita la capacidad del modelo para aprender patrones discriminatorios, incluso si el género no está explícitamente incluido en el modelo.

Interpretación de resultados

Después de aplicar ambas técnicas, reentrenamos el modelo de Random Forest y evaluamos tanto el rendimiento general (accuracy, precision, recall, F1-score) como las métricas de equidad, utilizando umbrales de disparidad definidos previamente (0.1) y evaluando los criterios de Statistical Parity, Equal Opportunity y Equalized Odds.

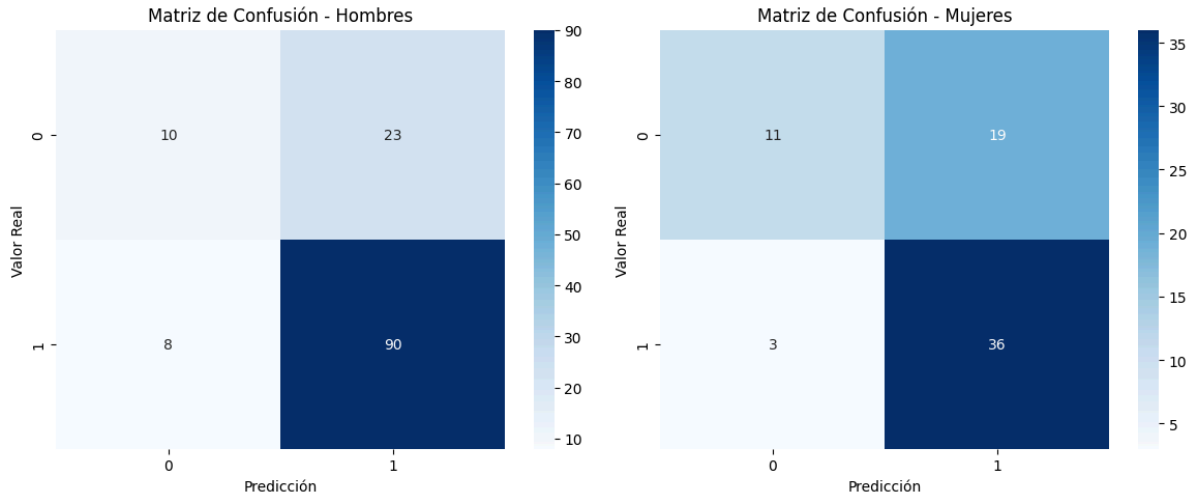
Los principales hallazgos cuando utilizamos el Correlation Remover fueron que solo una métrica mejoró, *Equality of Opportunity Difference* pasó de -0.0047 a 0.020931. Pero al mismo tiempo, *Statistical Parity Difference* aumentó de 0.073128 a 0.123465 y tanto Disparate Impact como *2SD Rule ambas subieron.* Esto muestra la tensión entre distintos criterios de fairness: mejorar una dimensión puede empeorar otra.

Reweighting, en cambio, mostró mejoras significativas en casi todas las métricas de equidad. Logró reducir la Statistical Parity Difference a 0.0510, y también mejoró la Average Odds Difference y la 2SD Rule. Sin embargo, no se observaron avances en la métrica de Equality of Opportunity Difference, que se mantuvo sin cambios relevantes.

Las métricas de equidad mostraron mejoras leves pero consistentes y el impacto fue moderado. Esto sugiere que el tipo de sesgo presente en este dataset se puede reducir parcialmente mediante el balanceo de pesos, aunque no de forma completa.

En cuanto al rendimiento predictivo, las métricas como precisión, recall y F1-score se mantuvieron prácticamente inalteradas en ambos modelos ajustados. Esto indica que es posible aplicar técnicas de mitigación sin comprometer la capacidad del modelo para identificar buenos y malos pagadores, lo cual es clave desde una perspectiva institucional.

B) Antes de realizar la mitigación de sesgos, se obtuvo la siguiente matriz de confusión para el atributo protegido

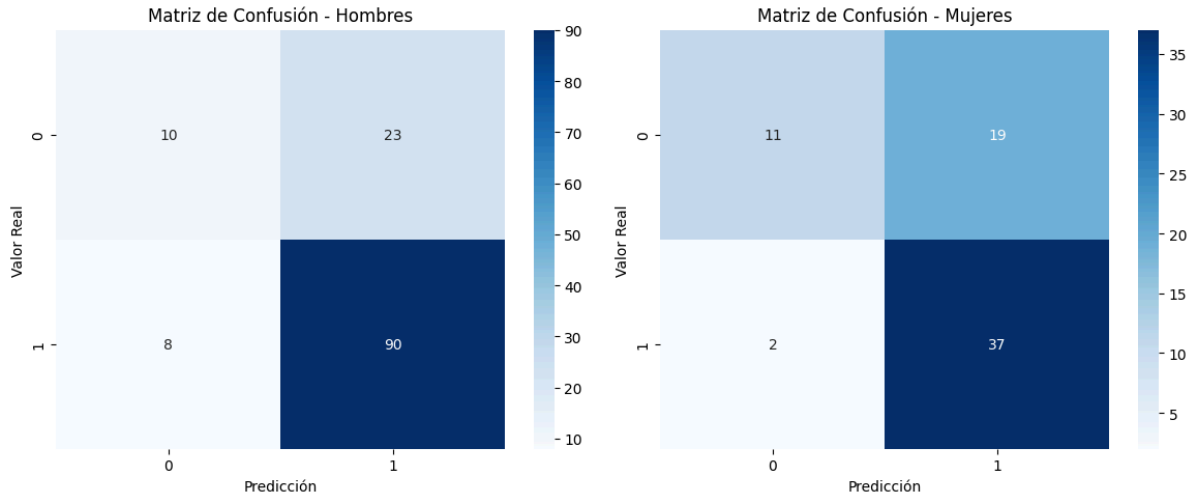


Y se obtuvieron las siguientes métricas de fairness, utilizando un umbral de 0.1:

- Statistical parity difference: 0.0731
- Equalized Odds difference: 0.0939
- Equal opportunity difference: 0.0047
- Predictive parity difference: 0.1349

Para las cuales vemos que con el umbral de 0.1 se cumplen:
Statistical parity difference, Equalized Odds difference, Equal opportunity difference

Luego se realizó una mitigación de sesgos, a través del método de rewieghting, el cual luego de llevarse a cabo devolvió la siguiente matriz de confusión:



Y se obtuvieron las siguientes métricas:

- Statistical parity difference: 0.0579
- Equalized Odds difference: 0.0667
- Equal opportunity difference: 0.0304
- Predictive parity difference: 0.1308

Para las cuales vemos que con el umbral de 0.1 se cumplen:
Statistical parity difference, Equalized Odds difference, Equal opportunity difference

También se realizó una mitigación por el método Correlation Remover, el cual devolvió las siguientes métricas:

- Statistical parity difference: 0.1235
- Equalized Odds difference: 0.1636
- Equal opportunity difference: 0.0209
- Predictive parity difference: 0.1102

Para las cuales vemos que con el umbral de 0.1 se cumple solamente Equalized Odds difference.

Vemos que luego de mitigar el sesgo con el método de rewieghting obtenemos valores más pequeños en las métricas, lo cual indica que el sesgo disminuyó. En cambio en el caso del método Correlation Remover, luego de aplicarlo solamente mejoró la métrica Equal opportunity, mientras que las otras superaron el umbral y dejaron de cumplirse.

5) A) Comparación entre el modelo original y los ajustados

Al comparar el modelo original con los modelos ajustados mediante técnicas de mitigación de sesgos, observamos que el modelo baseline ya presentaba un comportamiento razonablemente equitativo según varias métricas analizadas. La aplicación de Correlation Remover produjo algunas mejoras puntuales en el indicador equal opportunity, pero también generó pequeñas desventajas para el grupo protegido en otras métricas como Statistical Parity. Por su parte, Reweighting logró mejoras en todas las métricas menos en Equal Opportunity,, no obstante el impacto fue moderado. Es por ello que nuestro equipo decidió quedarse con el modelo de IA mitigado por reweighting.

B) Discusión sobre mejoras en fairness y performance

Aunque se intentaron técnicas de mitigación en la etapa de preprocesamiento, no se observaron mejoras significativas en la equidad del modelo. Las métricas de performance (accuracy, precision, recall, F1-score) se mantuvieron estables, lo cual es positivo, pero las intervenciones no aportaron beneficios concretos que justifiquen reemplazar el modelo original. De hecho, algunas métricas de fairness incluso se vieron levemente afectadas de forma negativa tras el ajuste. Esto refuerza la idea de que las técnicas de mitigación deben aplicarse con cuidado y que su efectividad depende del tipo y estructura del sesgo presente en los datos.

C) Reflexión sobre el impacto en el mundo real

En contextos reales como el otorgamiento de créditos, es fundamental lograr un equilibrio entre rendimiento y equidad. Un modelo que discrimina injustamente puede tener consecuencias sociales graves, pero uno que pierde precisión también puede afectar la rentabilidad y la confianza en el sistema. En este trabajo, vimos que aplicando mitigación de sesgos se pueden lograr predicciones más justas de un modelo. En nuestro caso, elegimos implementar una mitigación por el método reweighting que logra una buena capacidad predictiva y un trato relativamente equitativo entre géneros. Además, seguimos considerando que Equalized Odds es el criterio más relevante, ya que garantiza que quienes tienen el perfil adecuado no sean discriminados por su género, sin forzar decisiones que comprometan el objetivo del sistema, y también tiene en cuenta la tasa de FPR que es decisiva para el negocio.