Trabajo Práctico Final
Equidad en Aprendizaje Automático

Búsqueda de Equidad en Clasificador de Riesgo Crediticio

Docente:
Mariela Rajngewerc

Alumnos:
Facundo Latini Leiva,
Matías Ezequiel Raina

15/10/2025

1. Introducción

1.1 Contexto y Objetivos

El aprendizaje automático revolucionó la toma de decisiones en el sector financiero, particularmente en la evaluación de riesgo crediticio. Sin embargo, la implementación de estos sistemas conlleva una responsabilidad ética fundamental: garantizar que las decisiones automatizadas no perpetúen o amplifiquen sesgos históricos existentes en los datos. El **German Credit Data**, conjunto de datos ampliamente utilizado en la literatura, representa un caso de estudio ideal para examinar estos desafíos en un contexto de *scoring* crediticio del mundo real.

La **equidad algorítmica** emerge como una disciplina crítica que busca identificar, medir y mitigar sesgos injustos en modelos predictivos. En el dominio crediticio, donde las decisiones afectan directamente el acceso a oportunidades económicas, la transparencia y justicia de estos sistemas es una consideración técnica, y un imperativo ético y legal.

Este trabajo tiene como objetivo principal desarrollar, evaluar y mejorar un sistema de clasificación de riesgo crediticio, haciendo especial énfasis en la equidad algorítmica hacia diferentes grupos demográficos, particularmente en función del género. Más específicamente, se plantean los siguientes pasos:

1. Analizar exhaustivamente el conjunto German Credit Data, identificando potenciales sesgos históricos y representacionales.
2. Construir un modelo baseline de regresión logística para predicción de riesgo crediticio.
3. Evaluar sistemáticamente la equidad del modelo utilizando múltiples métricas y criterios establecidos.
4. Implementar técnicas de mitigación de sesgos y cuantificar su efectividad.

1.2 Estructura del Informe

El presente informe está estructurado de la siguiente manera:

- Sección 2: Análisis del conjunto de datos e identificación de sesgos potenciales
- Sección 3: Desarrollo y evaluación de un modelo inicial de regresión logística
- Sección 4: Implementación de técnicas de mitigación de sesgos y análisis parcial
- Sección 5: Análisis comparativo de los modelos y resultados finales
- Sección 6: Conclusiones y reflexiones finales

2. Conjunto de Datos y Análisis Exploratorio

2.1 Datasheet del Conjunto de Datos

El conjunto **Statlog (German Credit Data)** fue desarrollado como *benchmark* para comparar métodos de clasificación en el contexto del proyecto Statlog, realizado en 1994. Compilado originalmente por el Prof. Hans Hofmann de la Universidad de Hamburgo, contiene 1,000 solicitudes de crédito anonimizadas de un banco alemán. Desde su creación y difusión, ha sido ampliamente utilizado en investigación para evaluar algoritmos de clasificación.

Las características principales del conjunto de datos son:

- Tamaño: 1,000 instancias con 20 variables independientes
- Variable objetivo: Clasificación binaria "buen riesgo"/"mal riesgo" crediticio
- Atributos: Mixtos (numéricos y categóricos) sobre información personal y financiera
- Versión utilizada: Original con variables categóricas para preservar interpretabilidad

2.2 Análisis Exploratorio de Datos (EDA)

Nuestro análisis comenzó examinando la estructura fundamental del conjunto de datos, donde identificamos un desbalance significativo en la variable objetivo:

Cantidad de 'buenos' pagadores: 700

Este hallazgo inicial reveló una distribución 70/30 entre buenos y malos pagadores, alertándonos sobre la necesidad de considerar técnicas para manejar datos desbalanceados durante el modelado.

Al profundizar en la composición demográfica, descubrimos patrones preocupantes en la representación:

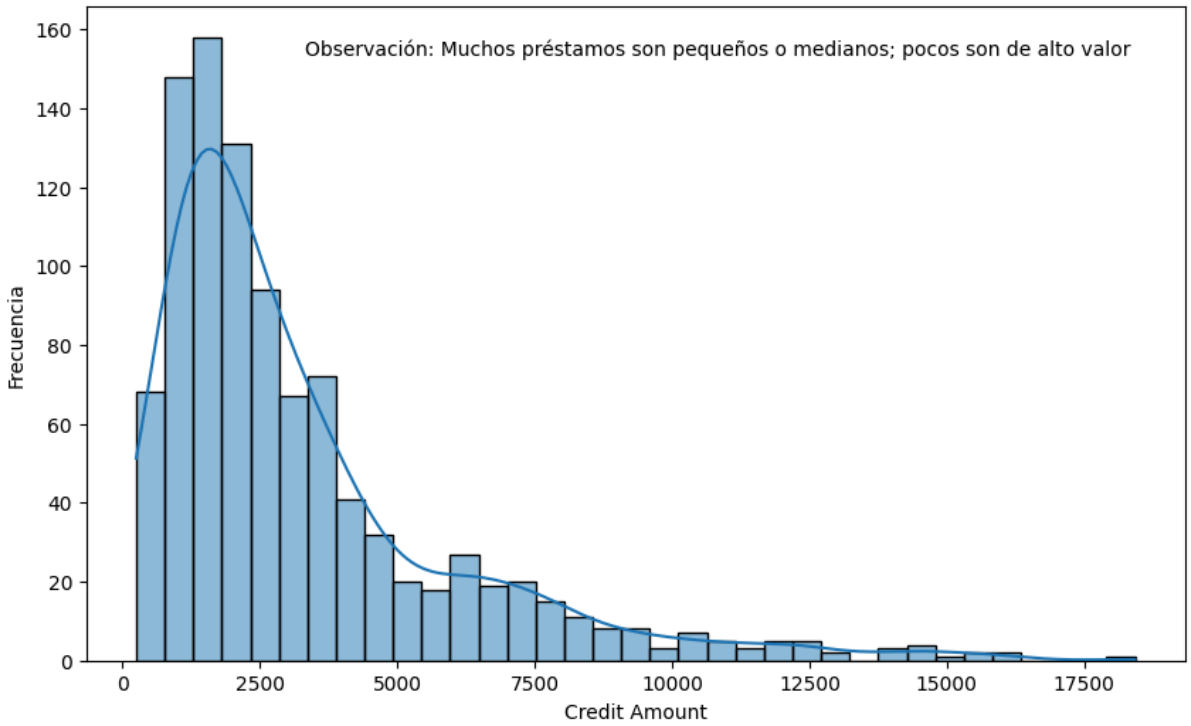
Cantidad de hombres: 690
Cantidad de extranjeros (A202): 37

La sobrerrepresentación masculina (69% vs 31% de mujeres) y la mínima presencia de trabajadores extranjeros plantearon inmediatas banderas rojas sobre la capacidad del modelo para generalizar de manera justa.

Un hallazgo particularmente interesante surge al examinar la variable de estado de trabajador extranjero: la documentación dice que 'A201' representa trabajadores extranjeros, pero de ser así se vería una sobrerrepresentación de los mismos. Da la sensación de que la variable está invertida, por lo que interpretamos que así es.

Por otro lado, se graficaron otras variables de interés, como lo son el monto del crédito (“*credit_amount*”) y la edad del solicitante (“*age*”) para ver sus distribuciones:

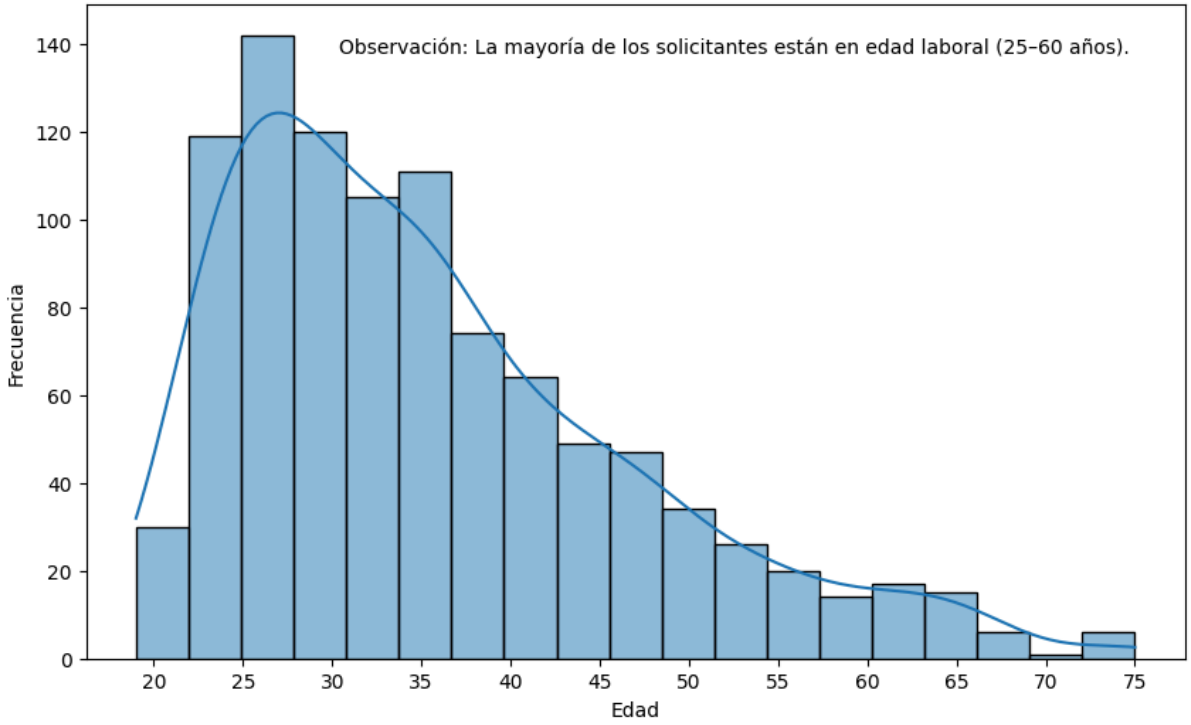
Distribución del monto del crédito solicitado:



Observación: Muchos préstamos son pequeños o medianos; pocos son de alto valor

Esta distribución sugiere que el dataset representa principalmente operaciones crediticias de bajo a medio riesgo, lo que podría limitar la capacidad predictiva del modelo para préstamos mayores.

Distribución de las Edades de los solicitantes:



Observación: La mayoría de los solicitantes están en edad laboral (25-60 años).

La concentración en edades productivas refleja el contexto histórico de solicitud de créditos, pero plantea desafíos para evaluar *fairness* en grupos etarios extremos.

2.3 Identificación de Sesgos Potenciales

Ya realizada la exploración inicial del conjunto de datos, pudimos identificar las siguientes fuentes de sesgo potenciales:

1. Sesgo Histórico: Los datos fueron recolectados en un contexto en el que era más común que un préstamo lo pidiera un hombre, por lo que el modelo puede aprender y perpetuar patrones discriminatorios del pasado, reflejando desigualdades estructurales en el acceso al crédito de la época.

2. Sesgo de Representación: Hay varias categorías que están subrepresentadas en el dataset, como la cantidad de mujeres, de extranjeros, y de adultos mayores; esto puede hacer que el modelo aprenda patrones poco fiables o erróneos para grupos minoritarios.

3. Sesgo de Aprendizaje: Es posible que el modelo priorice a la clase 'good credit' dado que es mayoritaria, en cuyo caso se podría amplificar la desigualdad entre clases presente en los datos, penalizando aún más a los grupos ya desfavorecidos en el contexto crediticio.

3. Modelo Inicial: Regresión Logística

3.1 Preprocesamiento de Datos

Se optó por utilizar el conjunto de datos en su formato original para preservar la interpretabilidad de las variables, un aspecto crucial dado que el análisis de equidad requiere un conocimiento exhaustivo y transparente de los atributos.

En primer lugar, se convirtió la variable "*personal_status_sex*" a "*gender*", de forma que se facilitara el análisis de equidad de nuestros modelos.

Luego, se dividió el conjunto de datos en subconjuntos de entrenamiento y evaluación, una práctica estándar en machine learning que permite evaluar la capacidad de generalización del modelo utilizando datos no vistos durante el entrenamiento.

Para realizar la transformación necesaria de las variables categóricas, la opción utilizada fue usar la implementación de scikit-learn del One Hot Encoder. Las variables numéricas no fueron transformadas como parte del preprocesamiento.

Por otro lado, la variable target fue mapeada de forma que tuviera el rango clásico de las variables binarias. De esta manera, a los "buenos pagadores" se los representa con cero (en lugar de uno), mientras que a los "malos pagadores" con uno (en lugar de dos).

3.2 Entrenamiento del Modelo

El modelo se entrenó usando un Pipeline para los datos, de forma que se empleara el preprocesador de las variables categóricas antes de llegar a la implementación del regresor logístico. Nuevamente, la librería utilizada para el modelo fue scikit-learn.

La mayoría de los hiper parámetros se dejaron en el estado default que provee scikit-learn, la única modificación fue definir "*max_iter*=10_000". Este parámetro especifica el número máximo de iteraciones que el algoritmo de optimización puede realizar para encontrar los coeficientes óptimos del modelo. Si se llega al resultado óptimo antes de alcanzar la cantidad máxima de iteraciones, el algoritmo se detiene. Por el contrario, si no se encuentra una solución óptima antes de alcanzar dicho número, igualmente se detiene pero se imprime una advertencia (*ConvergenceWarning*).

En nuestro caso, no hubo problemas de convergencia durante el entrenamiento con la cantidad de iteraciones que definimos, por lo que podemos afirmar que se encontró una solución óptima.

3.3 Evaluación del Modelo Inicial

Una vez finalizada la etapa de entrenamiento del modelo, se realizó la evaluación del mismo utilizando métricas estándar, como *accuracy*, *recall*, *f1-score* y precisión, y adicionalmente, se evaluó el costo teórico de los errores del modelo usando una matriz de costo (recomendada en la documentación original del dataset). Ésta define el costo que tienen los

errores de la matriz de confusión, teniendo un ratio de penalización Falsos Negativos:Falsos Positivos de 5:1.

La decisión de emplear la matriz de costo, se da como resultado de dos factores fundamentales: la documentación asegura que se debe usar como parte del uso correcto del conjunto de datos, y, en la consigna del trabajo, se pide que trabajemos desde el punto de vista de un equipo del banco que busca minimizar los préstamos impagos. Naturalmente, y como los objetivos se alinean, se eligió calcular el costo del modelo usando la matriz recomendada.

Los resultados obtenidos fueron los siguientes:

Matriz de Confusión:

	Predicho Negativo	Predicho Positivo
Real Negativo	126	15
Real Positivo	24	35

- Se acierta en la mayoría de los casos.
- Hay más Falsos Negativos que Falsos Positivos: esto resulta problemático desde el punto de vista del costo, ya que estos errores son los más penalizados.
    - Con una cuenta rápida se puede ver que la mayor parte del costo total del modelo viene de los Falsos Negativos:

      $$total\,cost = fp * 1 + fn * 5 = 15 + 120 = 135,$$

      donde $fp =$ Falso Positivo, y $fn =$ Falso Negativo

Accuracy global:

- Se acierta en un 81% de los casos,
- pero como el dataset está desbalanceado 70/30, este resultado puede ser engañoso, ya que un clasificador trivial que únicamente predice la clase mayoritaria tendría un *accuracy* del 70%.

Análisis por clase:

Clase 0 = Good Credit:

- $precisión = 0.84$: Cuando el modelo predice que alguien es “buen cliente”, acierta el 84% de las veces.
- $recall = 0.89$: El modelo detecta correctamente al 89% de los buenos clientes reales.
- $f1 - score = 0.87$: Buen compromiso entre precisión y exhaustividad en el grupo.

Clase 1: Bad Credit:

- $precisión = 0.70$: Sólo el 70% de los rechazados son “malos pagadores” realmente
- $recall = 0.59$: Se están detectando apenas más de la mitad de los clientes realmente riesgosos; es particularmente crítico porque se definió de antemano que los Falsos Negativos son más costosos.
- $f1 - score = 0.64$: Performance deficiente arrastrada por el bajo *recall*.

Conclusiones del modelo inicial:

El modelo inicial, aunque técnicamente competente en términos de accuracy global, resulta financieramente subóptimo y potencialmente riesgoso para la operación crediticia. La posible causa es que la clase *Bad Credit* está subrepresentada en el dataset, haciendo que el modelo tenga menos datos de entrenamiento para esos casos y, como resultado, no sea capaz de un rendimiento adecuado.

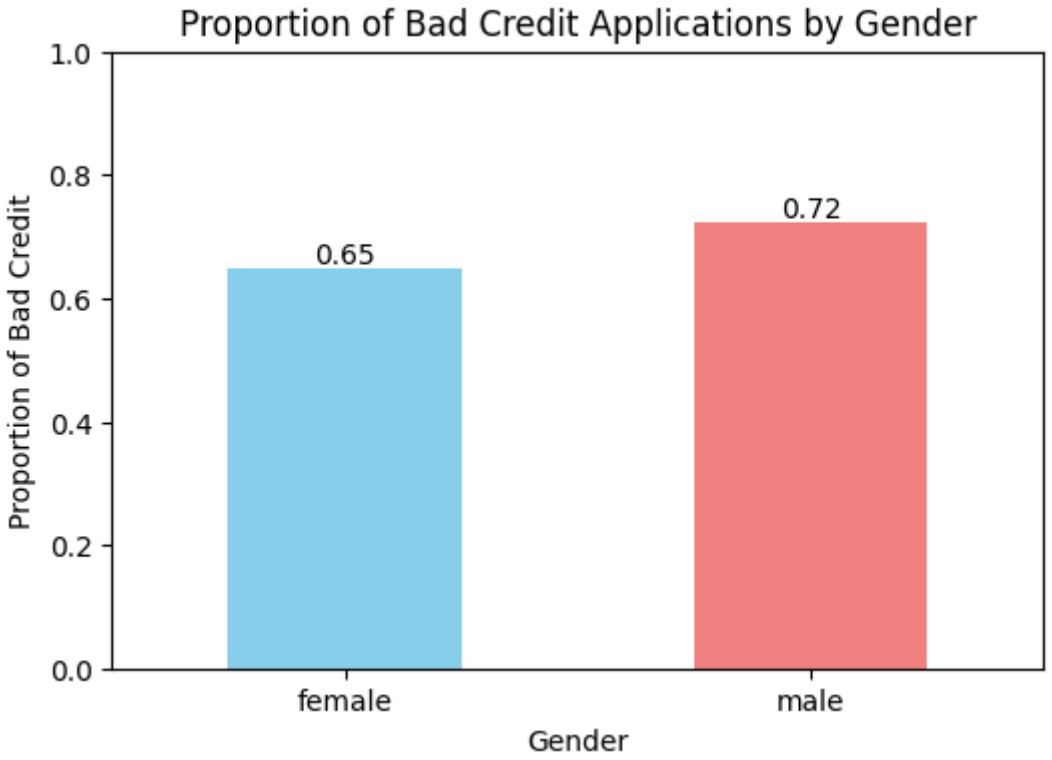
3.4. Evaluación de Equidad del Modelo Inicial

3.4.1. Métricas de Equidad

Primeramente, nos parece adecuado definir y contextualizar las métricas de equidad a analizar:

- **Statistical parity**: Compara la tasa de predicciones positivas (en este caso, de “mal cliente”) entre grupos sensibles. En este caso se cumple *statistical parity* si hombres y mujeres reciben el mismo porcentaje de predicciones de mal cliente.
- **Equalized odds**: Exige igualdad tanto en TPR (tasa de verdaderos positivos) como en FPR (tasa de falsos positivos) entre los grupos. En este caso se cumple si para ambos géneros, el modelo detecta correctamente a los malos clientes con la misma frecuencia (TPR) y la proporción de buenos clientes marcados como malos es igual (FPR).
- **Equal opportunity**: Compara la tasa de verdaderos positivos (TPR) entre grupos. Es decir que para ambos géneros compara la proporción de los malos clientes que el modelo identifica correctamente.
- **Predictive parity**: Compara la precisión (PPV) de las predicciones positivas entre grupos, es decir, de todos los que el modelo predijo como malos clientes, cuántos realmente lo son. Si la precisión es igual para hombres y mujeres, significa que una predicción de mal cliente tiene el mismo nivel de confiabilidad sin importar el género.

Como paso inicial, comparamos el Base Rate para los géneros, lo que representa la cantidad de créditos solicitados por “malos pagadores”. Los resultados indican que, como era de esperar, hay diferencias en la prevalencia entre géneros:



Para evaluar el cumplimiento de los criterios de equidad en nuestros modelos, calculamos las métricas de *fairness* comparando el desempeño entre los grupos de hombres y mujeres. Específicamente, utilizamos el módulo de la diferencia entre las métricas obtenidas para cada grupo.

La interpretación de estas diferencias requiere considerar el contexto proporcionado por el análisis del Base Rate, que revela una diferencia inherente de aproximadamente 7.5 puntos porcentuales en el riesgo crediticio real entre géneros. Esta disparidad subyacente implica que, incluso un modelo perfectamente preciso, produciría necesariamente diferencias en métricas como Statistical Parity.

Por lo tanto, en lugar de exigir una igualdad perfecta (diferencia = 0) que forzaría al modelo a distorsionar sus predicciones y ser menos preciso, establecimos un umbral de tolerancia práctico de **0.1**, o lo que es lo mismo, 10%. Este enfoque reconoce la existencia de diferencias reales en el riesgo crediticio, al mismo tiempo que limita la magnitud de las disparidades permitidas para evitar una discriminación algorítmica significativa.

3.4.2. Evaluación por Género

Los resultados de las métricas a evaluar son:

Género	Statistical parity	TPR (Equal opportunity)	FPR	Precisión (Predictive parity)
Femenino	0.232	0.588	0.077	0.769
Masculino	0.257	0.595	0.118	0.676

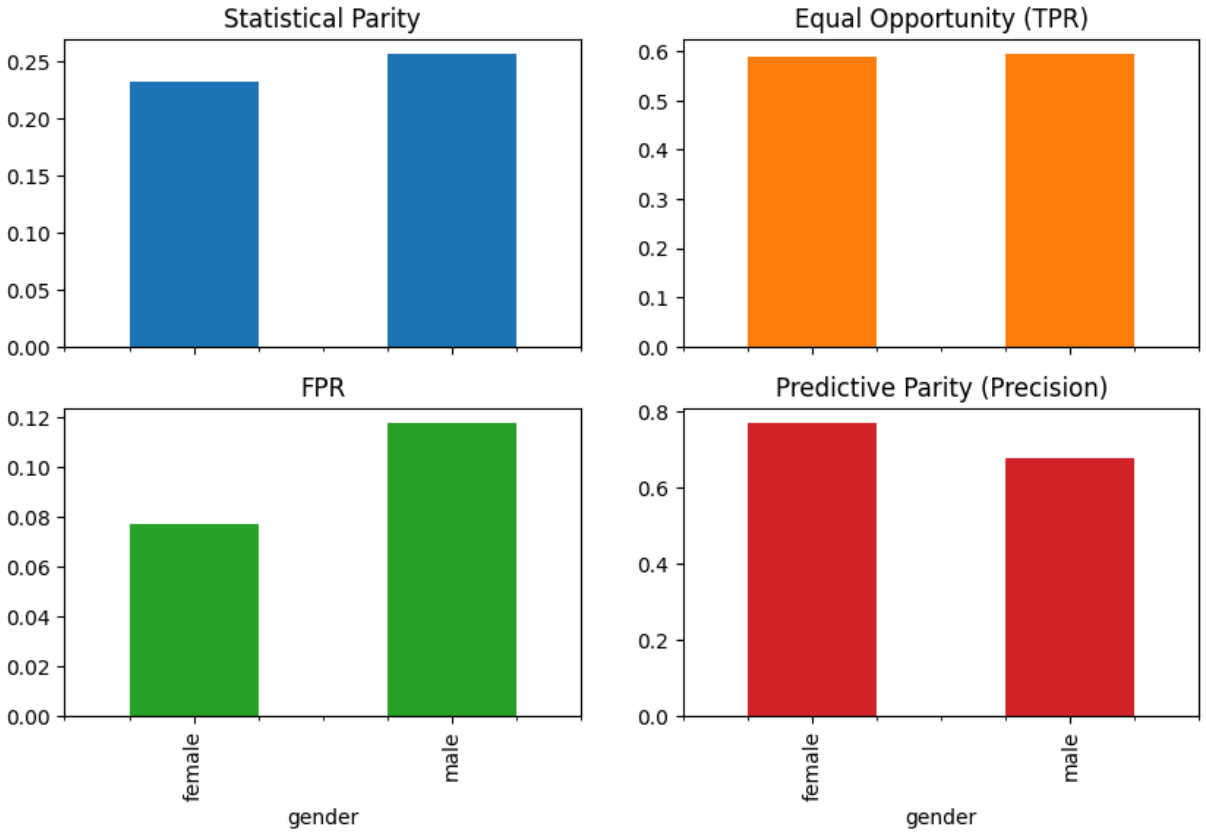
Con lo cual, se puede determinar la diferencia entre géneros para cada uno de los criterios ya descritos.

$$Statistical\ Parity\ Difference\ =\ 0.025$$

$$Equalized\ Odds\ Difference\ =\ 0.041$$

$$Equal\ Opportunity\ Difference\ (TPR)\ =\ 0.007$$

$$Predictive\ Parity\ Difference\ (Precision)\ =\ 0.094$$



3.4.3. Análisis de Resultados de Equidad

Teniendo en cuenta el umbral definido previamente, se puede concluir que este modelo es fair bajo los cuatro criterios, ya que ninguna de las diferencias es mayor a 0.1.

Sin embargo, observando el gráfico se puede identificar un FPR superior para el género masculino. Esta diferencia parece bastante grande pero, al tratarse de valores muy bajos, no supera el umbral de 0.1. Además, si bien la diferencia de la precisión no supera el umbral, está bastante cerca (0.094).

Desde el punto de vista del banco, para identificar un sesgo de género, el criterio más relevante es Equal opportunity. El hecho de que se cumpla este criterio significa que el modelo no favorece ni perjudica a ningún género a la hora de identificar a los malos pagadores. Además, el TPR tiene en cuenta a los FN, lo cual ya mencionamos que es relevante para el banco.

Por otro lado, el criterio Predictive parity es relevante si se busca que el desempeño del modelo sea bueno y confiable para ambos géneros por igual. Una diferencia de precisión entre los grupos significa que el modelo tiene un desempeño desigual y podría estar favoreciendo a un género en términos de exactitud predictiva.

4. Mitigación de Sesgos

4.1 Selección de Técnicas de Mitigación

Para el pre-procesamiento del modelo, elegimos usar Reweighting por su capacidad para corregir desbalances demográficos directamente en los datos de entrenamiento. Como se observó previamente, el dataset presenta desbalance tanto entre clases (buenos y malos pagadores) como entre grupos sensibles (hombres y mujeres). Esta técnica consiste en ajustar los pesos de las instancias de entrenamiento en función de su etiqueta y atributo sensible, para garantizar una representación equilibrada entre los grupos. La idea es que cada subgrupo (por ejemplo, female-good credit, male-bad credit, etc.) tenga un peso determinado de manera que el clasificador no se vea sesgado hacia los grupos mayoritarios.

Por otra parte, elegimos aplicar Equalized Odds como técnica de mitigación en el post-procesamiento. Este método ajusta las predicciones finales para que el TPR y el FPR sean similares entre grupos sensibles. Esto se logra cambiando las predicciones positivas y negativas con mayor incertidumbre, de modo que la distribución de los errores sea similar entre grupos. En este caso, el objetivo sería igualar las tasas de detección de malos pagadores (TPR) y de buenos pagadores mal clasificados (FPR) entre los géneros.

De esta manera se podrían mejorar las métricas de Fairness del modelo inicial, entendiendo que posiblemente haya un deterioro en las métricas de *performance*.

4.2 Técnica 1: Preprocessing - Reweighting

El algoritmo de Reweighting se aplicó a través de la librería holisticai. Se calcularon los pesos para cada instancia teniendo en cuenta su género y su clase (good/bad credit). Estos pesos se utilizaron en el entrenamiento de un modelo de scikit-learn de regresión logística, igual al modelo inicial.

A continuación se detallan los resultados parciales:

Matriz de confusión:

	Predicho Negativo	Predicho Positivo
Real Negativo	123	18
Real Positivo	26	33

Métricas de performance para la Clase 1: Bad Credit:

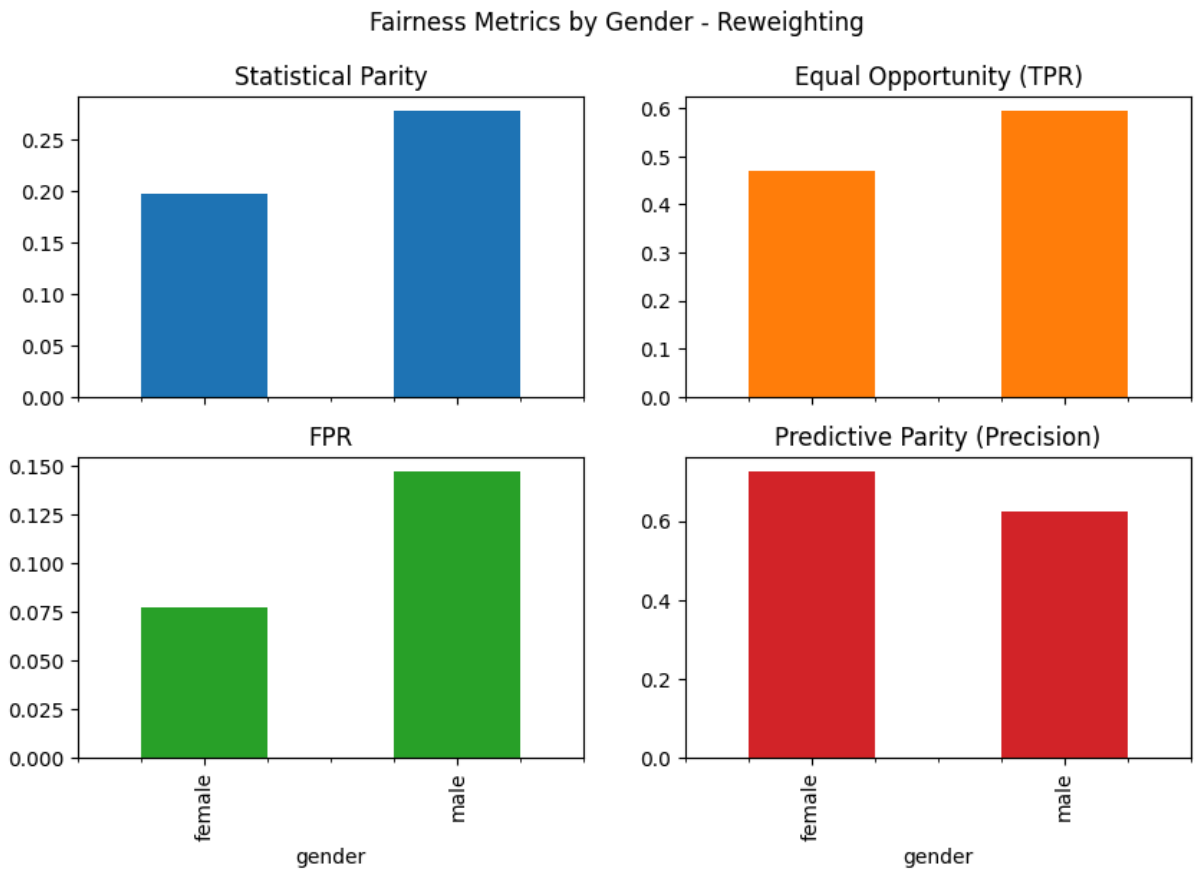
$$accuracy = 0.78$$

$$precisión = 0.65$$

$$recall = 0.56$$

$$f1 - score = 0.60$$

Métricas de equidad:



Género	Statistical parity	TPR (Equal opportunity)	FPR	Precisión (Predictive parity)
Femenino	0.196	0.471	0.077	0.727
Masculino	0.278	0.595	0.147	0.625

De la tabla se desprenden las siguientes diferencias absolutas:

$$Statistical\ Parity\ Difference = 0.081$$

$$Equalized\ Odds\ Difference = 0.125$$

$$Equal\ Opportunity\ Difference\ (TPR) = 0.125$$

$$Predictive\ Parity\ Difference\ (Precision) = 0.102$$

4.3 Técnica 2: Post Processing - Equalized Odds

Para complementar la mitigación, se aplicó Equalized Odds sobre las predicciones obtenidas del último modelo. Al igual que en el proceso anterior, acá también se utilizó la librería holisticai. Se configuraron los parámetros:

- *group_a*: male
- *group_b*: female
- *y_pred*: predicciones obtenidas luego de aplicar Reweighting

Los resultados luego de haber aplicado ambas técnicas de mitigación son los siguientes:

Matriz de confusión:

	Predicho Negativo	Predicho Positivo
Real Negativo	123	18
Real Positivo	30	29

Métricas de performance (tomando Clase 1: Bad Credit):

$$accuracy = 0.76$$

$$precisión = 0.62$$

$$recall = 0.49$$

$$f1 - score = 0.55$$

Métricas de equidad:

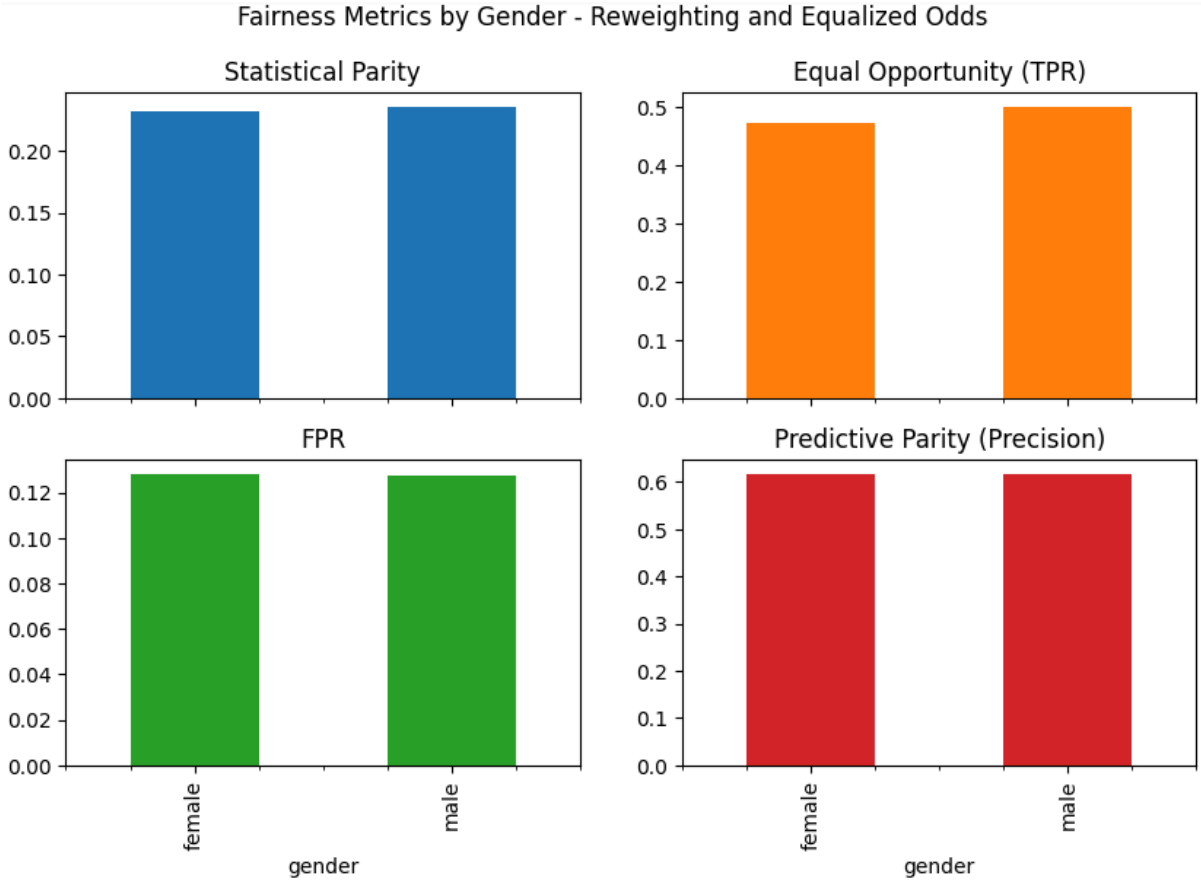
Género	Statistical parity	TPR (Equal opportunity)	FPR	Precisión (Predictive parity)
Femenino	0.232	0.471	0.128	0.615
Masculino	0.236	0.500	0.127	0.618

$$Statistical\ Parity\ Difference = 0.004$$

$$Equalized\ Odds\ Difference = 0.029$$

$$Equal\ Opportunity\ Difference\ (TPR) = 0.029$$

$$Predictive\ Parity\ Difference\ (Precision) = 0.002$$


Fairness Metrics by Gender - Reweighting and Equalized Odds
Statistical Parity
Equal Opportunity (TPR)
FPR
Predictive Parity (Precision)
0.00
0.05
0.10
0.15
0.20
0.0
0.1
0.2
0.3
0.4
0.5
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
0.1
0.2
0.3
0.4
0.5
0.6
female
male
gender
female
male
gender

5. Análisis Comparativo y Resultados

A continuación se realizará una comparación entre los tres modelos elaborados a lo largo del trabajo.

5.1 Comparación de Performance

Para resumir los resultados obtenidos en cada uno de los modelos analizados, presentamos la siguiente tabla comparativa de performance:

Modelo	Accuracy	Precisión (clase 1)	Recall (clase 1)	F1-Score (clase 1)	Costo total
Inicial	0.81	0.70	0.59	0.64	135
Reweighting	0.78	0.65	0.56	0.60	148
Reweighting + EO	0.76	0.62	0.49	0.55	168

Se puede observar un empeoramiento claro en la capacidad predictiva conforme se aplican técnicas de mitigación, lo cuál es esperable considerando que la búsqueda al usarlas no es tener mejor rendimiento, sino un modelo más justo.

5.2 Comparación de Equidad

La siguiente tabla comparativa, muestra la evolución de las métricas de *fairness* en los modelos desarrollados. El valor que se expresa es la diferencia absoluta entre hombres y mujeres.

Modelo	Statistical Parity	Equal Opportunity	Predictive Parity	Equalized Odds
Inicial	0.025	0.007	0.094	0.041
Reweighting	0.081	0.125	0.102	0.125
Reweighting + EO	0.004	0.029	0.002	0.029

Como se podría esperar, hay una mejoría entre el modelo inicial y al cual se le aplicaron ambas técnicas de mitigación discutidas anteriormente. Ambos modelos son equitativos bajo las definiciones de Statistical Parity, Equal Opportunity, Predictive Parity y Equalized Odds.

Contrario a lo que podíamos esperar, el modelo inicial termina siendo *fair* bajo más definiciones que aquel al que se le aplicó Reweighting. Algunas posibles causas de este comportamiento poco intuitivo son la presencia de variables *proxy* que distorsionan el efecto del Reweighting, y la posibilidad de que el algoritmo usado para entrenar el modelo causara *over-fitting* a patrones específicos de los subgrupos (reduce la capacidad de generalización equitativa).

Más allá de la observación del caso del modelo con Reweighting, se puede comprobar que utilizar técnicas de mitigación de sesgo, efectivamente hizo que las diferencias entre grupos para las distintas definiciones de *fairness* fueran más bajas.

5.3 Análisis del Trade-Off y Comportamiento Contraintuitivo

La narrativa tradicional del trade-off lineal entre accuracy y fairness se presentó al comparar el modelo inicial con el modelo al que se le aplicó Reweighting y Equalized Odds: ante la implementación de medidas que buscan mejorar el *fairness*, el rendimiento del modelo empeoró.

De todas formas, esto resultó insuficiente para capturar la complejidad observada en el paso intermedio. En su lugar, identificamos un trade-off multidimensional donde:

- No todas las técnicas de mitigación son igualmente efectivas
- La aplicación secuencial puede ser necesaria (Reweighting + Equalized Odds)
- El contexto específico del dataset determina la efectividad de cada método

El hallazgo tiene implicaciones muy relevantes para la toma de decisiones en contextos en los que se busca producir modelos equitativos. En particular, refuerza la necesidad de:

1. Evaluación experimental rigurosa de cada técnica de mitigación
2. Monitoreo continuo de múltiples dimensiones de equidad
3. Enfoques combinados en lugar de soluciones únicas
4. Comprensión profunda de las interacciones entre sesgos en los datos

6. Conclusiones

A continuación, detallaremos las conclusiones extraídas del análisis de nuestros modelos.

6.1 Resumen de Hallazgos

A lo largo del trabajo se abordó de manera integral la problemática de la equidad en un modelo de asignación de créditos. Partiendo del análisis exploratorio del conjunto German Credit Data, se identificaron fuentes de sesgo histórico y de representación que podrían afectar la equidad de un modelo de aprendizaje automático.

El modelo baseline, una regresión logística, presentó niveles aceptables de accuracy, pero con algunas diferencias en métricas de equidad entre hombres y mujeres. De todas maneras, el modelo se puede considerar justo bajo las definiciones de equidad empleadas y el umbral utilizado.

La incorporación de Reweighting como técnica de preprocesamiento no resultó como se esperaba, siendo que generó un deterioro en las métricas de equidad. Esto sumado a que empeoró ligeramente la performance del modelo, nos hizo cuestionar la metodología y la importancia de la evaluación de los sesgos posibles en conjunto.

Por su parte, la aplicación de Equalized Odds en el post-procesamiento mejoró la equidad, estabilizando las diferencias entre grupos. Sin embargo, mantuvo la pérdida de performance en niveles manejables.

En definitiva, se cuantificó un trade-off claro: el modelo más equitativo (Reweighting + EO) redujo el accuracy en 5 puntos porcentuales (baja de 81% a 76%) pero mejoró la equidad en todas las métricas.

6.2 Limitaciones y Trabajo Futuro

Si bien se obtuvieron resultados apropiados, este trabajo presenta algunas limitaciones. Por ejemplo, solo se trabajó con el género como atributo sensible, pero atributos como la edad o la nacionalidad podrían también introducir sesgos. Además, el hecho de que sea un dataset con datos de 1994 limita un poco la generalización para el sistema crediticio actual.

Por otro lado, consideramos que el análisis unidimensional del sesgo es una posible fuente de distorsión de los resultados obtenidos, especialmente en el modelo intermedio. Para trabajo futuro, se propone:

1. Implementar un análisis de correlaciones exhaustivo para identificar y controlar variables proxy

2. Explorar técnicas de mitigación multidimensional que consideren más features

3. Investigar approaches de in-processing como adversarial debiasing que podrían ofrecer mejores trade-offs

4. Validar los hallazgos en datasets crediticios más recientes y diversos

6.3 Reflexiones Finales

Este trabajo expone que la equidad en aprendizaje automático debe ser tenida en cuenta en todas las etapas del ciclo de vida de un modelo, y que va más allá de una simple implementación de técnicas de mitigación.

En contextos como el de los créditos bancarios, las decisiones de los algoritmos pueden tener impacto social y económico de manera directa, estos hallazgos enfatizan la necesidad de evaluar continua y multidimensionalmente la equidad, y ser transparente metodológicamente. Es por esto que es importante darle importancia tanto a la detección de sesgos como a su mitigación, aplicando las técnicas adecuadas para cada caso, e intentando además de alcanzar un balance entre precisión y equidad.

Anexos

Repositorio de GitHub del proyecto: https://github.com/matias-raina/TP_Fairness_ML

Datasheet desarrollado:
https://github.com/matias-raina/TP_Fairness_ML/blob/main/German%20Credit%20Data%20-%20Datasheet.pdf