Equidad en Aprendizaje Automático

Trabajo práctico final

Profesora:
Mariela Rajngewerc

Integrantes:
Gianni Bevilacqua
Gerardo Toboso
Javier Spina

Licenciatura en Ciencia de Datos - 1°C 2025

Introducción

En este trabajo desarrollamos un modelo de clasificación binaria para predecir si una persona debería recibir un crédito bancario, utilizando el conjunto de datos “German Credit Data”. Más allá del rendimiento tradicional del modelo (evaluado mediante métricas como accuracy, precision, recall y f1-score), abordamos específicamente el análisis de equidad (fairness), con especial foco en el tratamiento diferencial por el género de las personas que solicitan un préstamo. Finalmente, implementamos técnicas de mitigación de sesgos y comparamos los resultados con el modelo original para evaluar mejoras en términos de justicia algorítmica.

1. Conjunto de datos

Datasheets for Datasets

i) Motivación:

Según la fuente UC Irvine , el dataset original fue donado por el profesor Hans Hofmann. En los metadatos del dataset se indica que el área de interés es “Social Science”, es decir Ciencias Sociales, y por la fecha en la que se donó (1994); podemos asumir que está disponible para fines académicos y educativos, también podríamos asumir que la tarea específica claramente no era el estudio de fairness como lo estudiamos hoy por hoy y que la brecha entre géneros no era el foco.

ii) Composición:

Cada una de las observaciones del conjunto de datos es una persona que pidió un crédito, una serie de atributos personales (20) y una clasificación de riesgo crediticio (target).

iii) Proceso de recopilación:

Según el documento South German Credit Data: Correcting a Widely Used Data Set (sección 4), los datos fueron recopilados por una institución bancaria entre los años 1973 y 1975, en el sur de Alemania. Posteriormente fueron procesados como parte de una tesis doctoral en 1979 para luego ser procesadas y seleccionadas 1000 filas por Hofmann.

iv) Preprocesamiento/limpieza/etiquetado:

Los datos fueron anonimizados y también, según fuente, las etiquetas de algunas variables están mal utilizadas. También, hay variables como “personal_status_sex” que combina información de sexo y estado civil; no hay valores faltantes en el dataset y se encuentra ampliamente estandarizado. Por lo tanto, para nuestro casoextraemos la información de la variable “peronal_status_sex,” la resumimos y creamos una nueva feature “sex” que contiene unicamente la información del sexo de la persona para poder utilizarla posteriormente.

v) Usos:

Se ha utilizado ampliamente en la investigación de machine learning. En Kaggle se puede acceder a varias notebooks implementando modelos de solvencia crediticia, además de ser ampliamente referenciado en la academia. Su uso más novedoso está relacionado con el análisis de fairness.

El repositorio que enlaza a los documentos y sistemas se puede encontrar aquí: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data (incluye una versión descargable por medio de una librería de python). También en Kaggle: https://www.kaggle.com/datasets/uciml/german-credit/data. Tiene identificación DOI 10.24432/C5NC77.

Análisis exploratorio

Variables numéricas relevantes:

- La duración del crédito muestra una clara relación con el riesgo: los préstamos a largo plazo tienden a ser clasificados como de mayor riesgo.
- Solicitantes más jóvenes tienen mayor probabilidad de ser clasificados como de alto riesgo.
- Montos de crédito más altos también se asocian moderadamente con clasificaciones de mayor riesgo.

Variables categóricas importantes:

- El estado de la cuenta corriente es un fuerte predictor: saldos bajos se asocian con alto riesgo.
- Ahorros bajos (<100 DM) se asocian con clasificaciones de alto riesgo.
- Menor antigüedad laboral (menos de 1 año) está relacionada con clasificaciones de alto riesgo.
- Propietarios de vivienda tienen menor probabilidad de ser clasificados como de alto riesgo.

Consideraciones de equidad con respecto al género

- Existe una **disparidad notable** en las tasas de clasificación de alto riesgo entre géneros (35.2% para mujeres vs 27.7% para hombres).
- Esta **diferencia de 8 puntos porcentuales** plantea preocupaciones sobre posibles sesgos en la clasificación original.
- Las variables relacionadas con ingresos, empleo y propiedad podrían estar actuando como **proxies** que amplifican disparidades de género.

2. Creación de un modelo inicial

Preparación de los datos para el entrenamiento

Aplicamos “One-Hot Encoding” a nuestro conjunto de datos para convertir todas las variables categóricas a un formato numérico. Por último, realizamos la transformación de la variable objetivo y se asigna el valor “1” a las personas de bajo riesgo crediticio y “0” a las personas de alto riesgo crediticio siguiendo la convención habitual en problemas de clasificación binaria.

Luego con el objetivo de reducir el ruido y construir un modelo inicial más simple pero efectivo, aplicamos un criterio básico de **selección de variables basado en la correlación lineal** con la variable objetivo. En concreto, seleccionaremos aquellas variables que presenten una **correlación en valor absoluto mayor a 0.1** respecto a la misma. Este umbral fue elegido de manera pragmática: aunque no es alto, permite filtrar aquellas variables que tienen una **relación débil o inexistente con la variable objetivo**, sin ser demasiado restrictivo. De esta forma, eliminamos variables irrelevantes que podrían introducir ruido o redundancia en el modelo, al tiempo que conservamos aquellas que muestran al menos un **grado mínimo de asociación lineal** con el resultado que se desea predecir.

Evaluación de desempeño general

Una vez entrenado el modelo inicial obtuvimos los siguientes resultados en cuanto a métricas generales de clasificación se refiere:

- **Clase 1 (Bajo riesgo):** El modelo muestra un buen desempeño identificando solicitantes de bajo riesgo:

    - **Precisión:** 0.79 - El 79% de los solicitantes clasificados como bajo riesgo efectivamente lo son.
    - **Sensibilidad:** 0.91 - El modelo identifica correctamente el 91% de los solicitantes de bajo riesgo.
    - **F1-Score:** 0.84 - Buen balance entre precisión y sensibilidad para esta clase.

- **Clase 0 (Alto riesgo):** El desempeño es menor para identificar solicitantes de alto riesgo:

    - **Precisión:** 0.66 - El 66% de los solicitantes clasificados como alto riesgo realmente lo son.
    - **Sensibilidad:** 0.42 - El modelo solo identifica correctamente el 42% de los solicitantes de alto riesgo.
    - **F1-Score:** 0.52 - Desempeño bajo/moderado para esta clase.

- **Exactitud general:** 0.77 - El 77% de todas las predicciones son correctas.

Este desequilibrio en el desempeño entre clases podría atribuirse principalmente a la distribución desbalanceada de las clases en los datos, donde hay más solicitantes de bajo riesgo que de alto riesgo. Esta situación es común en problemas de crédito y requiere atención especial desde la perspectiva de equidad.

Interpretación de la matriz de confusión

Generamos una matriz de confusión que nos permite ver en detalle los errores y aciertos del modelo y entender su comportamiento en profundidad. A continuación realizamos una interpretación de los resultados:

- **Verdaderos positivos (TP):** 128 solicitantes fueron correctamente clasificados como de bajo riesgo.
- **Falsos negativos (FN):** 13 solicitantes de bajo riesgo fueron incorrectamente clasificados como de alto riesgo.
- **Falsos positivos (FP):** 34 solicitantes de alto riesgo fueron incorrectamente clasificados como de bajo riesgo.
- **Verdaderos negativos (TN):** 25 solicitantes fueron correctamente clasificados como de alto riesgo.

Implicaciones de los errores desde la perspectiva del banco

- **Falsos positivos (34):** Este tipo de error es particularmente problemático para instituciones financieras, ya que implica otorgar préstamos a personas que realmente son de alto riesgo, lo que podría resultar en pérdidas financieras y sería un grave error desde el punto de vista del banco que intenta maximizar sus ganancias.
- **Falsos negativos (13):** Aunque menos costoso desde la perspectiva financiera, este error implica negar préstamos a personas que realmente son de bajo riesgo, lo que representaría un problema de acceso de las personas a servicios financieros.

Con todo esto, y si consideramos la perspectiva del banco que otorga los préstamos, podemos afirmar que **un falso positivo es un error más grave** por los motivos ya explicados. La prioridad de una entidad como lo es un banco en la mayoría de los casos **es maximizar las personas que efectivamente van a pagar el préstamo** y un falso positivo representa alguien que fue clasificado como capaz de pagar (bajo riesgo) pero que realmente sería un solicitante con un alto riesgo.

3. Evaluación de equidad del modelo inicial

Matriz de confusión y métricas de rendimiento por género

Para profundizar en el análisis del modelo inicial, separamos la matriz de confusión por género con el objetivo de identificar diferencias en el comportamiento del clasificador según el grupo sensible. A continuación, describimos los resultados para cada caso:

- **Categoría masculina:**

    - **Verdaderos positivos (TP)**: 94 hombres clasificados originalmente como de bajo riesgo fueron correctamente clasificados como tales.
    - **Falsos negativos (FN)**: 8 hombres clasificados originalmente como de bajo riesgo fueron erróneamente clasificados como de alto riesgo.
    - **Falsos positivos (FP)**: 26 hombres clasificados originalmente como de alto riesgo fueron incorrectamente clasificados como de bajo riesgo.
    - **Verdaderos negativos (TN)**: 16 hombres clasificados originalmente como de alto riesgo fueron correctamente identificados.

- **Categoría femenina:**

    - **Verdaderos positivos (TP):** 34 mujeres clasificadas originalmente como de bajo riesgo fueron correctamente clasificadas como tal.
    - **Falsos negativos (FN):** 5 mujeres clasificadas originalmente como de bajo riesgo fueron erróneamente clasificadas como de alto riesgo.
    - **Falsos positivos (FP):** 8 mujeres clasificadas originalmente como de alto riesgo fueron clasificadas incorrectamente como de bajo riesgo.
    - **Verdaderos negativos (TN):** 9 mujeres clasificadas originalmente como de alto riesgo fueron correctamente identificadas.

Podemos observar que en ambos grupos, **el modelo logró clasificar correctamente a la mayoría de las personas consideradas como de bajo riesgo** (TP). Sin embargo, en valores absolutos, el número de verdaderos positivos fue mayor en hombres (94) que en mujeres (34), lo cual es consistente con la mayor proporción de hombres en el conjunto de datos.

La **tasa de falsos positivos** fue notablemente mayor en hombres: 26 errores sobre un total de 42 personas realmente de alto riesgo, mientras que en mujeres se observaron solo 8 falsos positivos sobre 17 casos de alto riesgo. Esto sugiere que el modelo fue más propenso a subestimar el riesgo crediticio en hombres.

Group Fairness y su pertinencia en nuestro análisis

En el contexto en el que nos encontramos, donde desarrollamos un modelo de aprendizaje automático que ayude en la asignación de créditos bancarios, es fundamental garantizar que las decisiones del sistema no estén sesgadas en función de atributos sensibles como el género. Un enfoque que nos ayudará a evaluar cuestiones relacionadas al sesgo por grupos es el **Group Fairness (equidad grupal)**, el cual permite evaluar si el modelo trata de manera equitativa a distintos grupos poblacionales definidos por atributos sensibles.

Este enfoque se basa en comparar métricas de desempeño del modelo **—como tasas de verdaderos positivos, falsos positivos, precisión o tasa de aprobación—** entre los distintos grupos. Su relevancia radica en que ayuda a identificar **disparidades sistemáticas que podrían derivar en discriminación algorítmica**, afectando la transparencia, la justicia y, potencialmente, el cumplimiento normativo de la organización.

Interpretación de los diferentes criterios de Group Fairness

Entre las principales definiciones de equidad según el enfoque de **Group Fairness**, se identifican **cuatro métricas clave**, que en nuestro caso pueden interpretarse de la siguiente manera:

Statistical Parity: En nuestro contexto, la clase positiva (1) representa a un solicitante con bajo riesgo crediticio. Este criterio busca que la proporción de personas clasificadas por el modelo como de bajo riesgo sea la misma en ambos grupos (géneros). Esto apunta a garantizar un acceso equitativo a decisiones favorables sin importar el grupo al que se pertenezca.

Equalized Odds: Este criterio exige que el modelo tenga la misma tasa de verdaderos positivos (clasificar correctamente como bajo riesgo) y de falsos positivos (clasificar como de bajo riesgo a quienes en realidad no lo son) para ambos grupos.

Equal Opportunity: Es una versión más flexible de Equalized Odds. Se enfoca únicamente en la tasa de verdaderos positivos, buscando que el modelo etiquete como de bajo riesgo a las personas que realmente lo son, con la misma probabilidad, sin importar su grupo.

Predictive Parity: Este criterio se centra en la precisión del modelo. En este caso, exige que, entre los solicitantes identificados por el modelo como de bajo riesgo, la proporción que realmente lo son, sea igual en ambos grupos. Esto asegura que la confianza que se puede tener en una predicción favorable sea independiente del grupo sensible.

Evaluación de disparidad

Para evaluar la equidad del modelo en términos de **Group Fairness**, decidimos adoptar como criterio de disparidad una diferencia máxima tolerable de **0.1** (en valor absoluto) entre las métricas de equidad de ambos géneros. Este umbral se fundamenta en dos consideraciones:

- **Interpretabilidad práctica**: diferencias mayores al 10% en métricas como la tasa de verdaderos positivos o la tasa de aprobación pueden implicar efectos sustanciales y sistemáticos sobre el acceso a beneficios (como créditos, becas o servicios), y por lo tanto, pueden considerarse potencialmente injustas o discriminatorias desde una perspectiva de impacto social.
- **Balance entre equidad y performance**: un umbral de 0.1 nos permite mantener un buen compromiso entre reducir el sesgo y no afectar excesivamente la capacidad predictiva del modelo. Umbrales más estrictos podrían forzar al modelo a sacrificar demasiado rendimiento, mientras que valores más permisivos podrían ignorar desigualdades sustanciales.

En este análisis, entonces, **consideramos como aceptable toda diferencia en módulo menor o igual a 0.1**. Si alguna métrica clave supera dicho umbral, interpretamos que hay indicios de sesgo que deben ser luego mitigados. Los resultados luego de la evaluación fueron los siguientes:

- **Demographic Parity (pprev)**:
  La diferencia absoluta en la proporción de personas clasificadas como de bajo riesgo entre hombres y mujeres fue de **0.0833**, lo cual está por debajo del umbral. Por lo tanto, **cumple con este criterio**.
- **Predictive Parity (precision)**:
  La diferencia en la precisión entre ambos grupos fue de **0.0261**, también por debajo del umbral. Por lo tanto, **cumple con este criterio**.
- **Equalized Odds**:
  - **TPR**: La diferencia fue de **0.0497**, por lo tanto **cumple con Equal Opportunity**.
  - **FPR**: La diferencia fue de **0.1485**, superando el umbral establecido. Por lo tanto, **no se cumple Equalized Odds**.

Elección de criterio de Fairness

A pesar de que el modelo cumple con varios criterios de equidad, como **Demographic Parity** y **Predictive Parity**, **no cumple con Equalized Odds** debido a una diferencia considerable en la tasa de falsos positivos entre hombres y mujeres. Esto indica que el modelo tiende a cometer más errores (en este caso, aprobar a personas de alto riesgo) en uno de los grupos sensibles.

Desde la perspectiva del banco, se remarca la importancia de tener un modelo justo en métricas de **True Positive Rate** y **False Positive Rate**, ya que se requiere clasificar justamente a la población que solicita y puede pagar un préstamo, y hacer que el modelo tenga una tasa similar de errores relacionados a las personas a las que se les asigna un préstamo teniendo un alto riesgo crediticio, sin importar su género. De esa manera, el riesgo que tomará el banco al otorgar créditos a personas con un alto riesgo crediticio será similar tanto para hombres como para mujeres. Por lo tanto, el criterio elegido es **Equalized Odds**.

4. Mitigación de sesgos

Uso de dos técnicas de mitigación y evaluación de desempeño

Dados los resultados que observamos en la sección anterior, para que nuestro modelo final cumpla con el criterio de equidad que establecimos debemos utilizar **técnicas de mitigación**. Para esta parte decidimos utilizar la librería de Python Holistic AI la cual ofrece una amplia gama de estas técnicas que nos ayudarán en nuestro objetivo.

Primera técnica de mitigación

Como primera técnica se aplica una que entra en la categoría de técnicas **preprocessing**, esta es **"Correlation Remover"**. Esta técnica tiene como objetivo principal eliminar la correlación lineal entre las variables de entrada y una o más variables sensibles, sin alterar excesivamente la estructura general de los datos. Es decir, busca preservar la información útil para la predicción, pero eliminando dependencias que podrían inducir sesgos.

Dado que nuestro modelo base es una regresión logística, que es particularmente sensible a la multicolinealidad y a la presencia de variables fuertemente correlacionadas, resulta pertinente aplicar un método como **Correlation Remover** antes del entrenamiento. Esto permite que el modelo no herede patrones de discriminación presentes en los datos originales, sin comprometer completamente la capacidad predictiva (que es lo que estamos buscando).

Entrenamiento de modelo usando primera técnica de mitigación

Según la documentación oficial de la librería, el parámetro alpha utilizado para instanciar el modelo que aplica este método **tiene un valor por defecto de 1**, lo cual implica, en el caso de esta técnica, una **eliminación completa** de la correlación lineal entre la variable sensible y el resto de las variables predictoras.

Tomamos la decisión de dejarlo en su valor por defecto ya que al tratarse de un mitigador de pre-procesamiento, el objetivo es prevenir que el modelo aprenda patrones discriminatorios desde el inicio del proceso de entrenamiento. Por tanto, aplicar la transformación con su intensidad máxima es coherente con una **postura conservadora y preventiva**, especialmente en dominios sensibles como el crédito, donde el impacto de decisiones injustas puede ser grave.

Evaluación de equidad para primera técnica de mitigación

En esta evaluación, nos centramos particularmente en dos métricas de equidad: **Equality of Opportunity Difference** (basada en la tasa de verdaderos positivos) y **False Positive Rate Difference**. Observamos que, al aplicar el mitigador, el modelo **empeoró su desempeño en ambas métricas** en comparación con el modelo original. En especial, la **diferencia en la tasa de falsos positivos** entre grupos se amplifica notablemente, alcanzando un valor absoluto cercano a **0.2**, lo que indica un sesgo aún más acentuado.

Esto sugiere que, si bien el objetivo del mitigador era reducir disparidades, en este caso puntual terminó generando **una mayor desigualdad en el tratamiento de los errores entre los grupos sensibles**, lo cual refuerza la importancia de evaluar críticamente el impacto real de cada técnica antes de adoptarla como solución definitiva.

Evaluación de clasificación para primera técnica de mitigación

En términos de métricas clásicas de clasificación, el modelo entrenado con la técnica de mitigación **Correlation Remover** mostró un desempeño muy similar al del modelo original, sin diferencias significativas en su capacidad de clasificación general. Esta estabilidad también se refleja en la matriz de confusión, donde los valores de verdaderos positivos (TP = 128) y falsos negativos (FN = 13) se mantuvieron idénticos respecto al modelo inicial. El único cambio notable fue un pequeño ajuste en la cantidad de falsos positivos, que disminuyó levemente de 34 a 33, mientras que los verdaderos negativos pasaron de 25 a 26, lo que sugiere una ligera mejora en la detección de personas consideradas como de alto riesgo. Sin embargo, este cambio es marginal y no representa un impacto significativo en el rendimiento global del modelo.

Segunda técnica de mitigación

Como segunda técnica de mitigación, se utilizó “**Prejudice Remover”**, un método perteneciente a la categoría de técnicas **in-processing**, ya que actúa durante el proceso de entrenamiento del modelo. Esta técnica introduce una penalización adicional en la función de pérdida, con el objetivo de desalentar la dependencia entre las predicciones del modelo y la variable sensible, que en nuestro caso es el género. En otras palabras, busca que el modelo aprenda a realizar buenas predicciones **sin apoyarse en información que pueda estar injustamente correlacionada con atributos protegidos**.

Prejudice Remover representa una forma directa y eficaz de incorporar criterios de justicia algorítmica sin alterar significativamente la estructura del modelo **ni comprometer su desempeño predictivo**, lo cual sigue siendo uno de nuestros objetivos principales.

Entrenamiento de modelo usando segunda técnica de mitigación

Según la documentación de la librería, el valor por defecto del parámetro tau utilizado para instanciar el modelo que aplica este método **es 1**. Este parámetro controla la **intensidad de la penalización** aplicada durante el entrenamiento cuando las predicciones del modelo presentan dependencia con la variable sensible. Es decir, a valores más altos de este parámetro, mayor es la penalización que recibe el modelo por aprender patrones discriminatorios relacionados con atributos protegidos.

Optar por este valor por defecto implica aplicar una penalización **moderadamente estricta**, lo suficientemente fuerte como para influir en el comportamiento del modelo, pero sin llegar a comprometer su capacidad predictiva. Esta elección resulta especialmente pertinente en nuestro caso, ya que el modelo base es una **regresión logística**, conocida por su sensibilidad a correlaciones lineales en los datos. En este contexto, dicho valor funciona como un mecanismo de **regularización ética**, orientado a evitar que el modelo utilice, de manera directa o indirecta, información vinculada al género para determinar la solvencia crediticia de una persona.

Evaluación de equidad para segunda técnica de mitigación

A diferencia de lo observado con otros enfoques, los resultados del modelo entrenado con **Prejudice Remover** muestran mejoras claras en las métricas de equidad que nos interesan particularmente. En primer lugar, se logró una **reducción significativa en la métrica False Positive Rate Difference**, que en el modelo original se encontraba fuera del umbral de equidad (0.1) y que ahora se reduce hasta quedar dentro del límite aceptable. Además, mantiene un buen desempeño en **Equality of Opportunity Difference**, conservando su valor por debajo del umbral definido.

Estos nuevos valores permiten afirmar que **el mitigador logró su cometido**, ya que ambas métricas se encuentran por debajo del umbral que consideramos nosotros. De esta manera, el modelo mitigado no solo mantiene un rendimiento similar en términos de clasificación, sino que también **corrige las desigualdades observadas en el modelo original**, alineándose con el criterio de fairness previamente seleccionado, **Equalized Odds**.

Evaluación de clasificación para segunda técnica de mitigación

En cuanto a las métricas clásicas de clasificación, el modelo entrenado con la técnica de mitigación **Prejudice Remover** mostró un rendimiento **muy similar al modelo original**, tanto a nivel global como por clase. La precisión, recall y f1-score son prácticamente equivalentes, lo cual indica que la introducción de esta técnica **no afectó negativamente el desempeño predictivo del modelo**.

Además, al observar la matriz de confusión, se confirma que el desempeño se mantiene estable: el número de **verdaderos positivos (TP)** aumentó levemente de 128 a 129, y los **falsos negativos (FN)** disminuyeron de 13 a 12. Por otro lado, los **falsos positivos (FP)** y **verdaderos negativos (TN)** se mantuvieron constantes (34 y 25, respectivamente). Este resultado refuerza que el modelo mitigado logra **preservar su capacidad predictiva**, al tiempo que mejora su comportamiento en términos de equidad, cumpliendo con el criterio de equidad previamente definido como objetivo.

Análisis y comparación de modelos mitigados junto con el inicial

Se realiza una recolección de las métricas de equidad y se exponen los resultados del modelo original junto con los resultados de los dos modelos mitigados en una misma tabla para poder compararlos.

Se observa que:

- El modelo que implementa **Prejudice Remover** es significativamente “más justo” en la mayoría de las métricas de equidad que el modelo inicial y el otro que implementa **Correlation Remover**.
- La métrica **Equality of Opportunity Difference**, resulta similar al modelo original y se mantiene muy por debajo del umbral elegido.
- La métrica **False Positive Rate Difference** en el modelo que implementa **Prejudice Remover** baja **de 0.1484 a 0.016**, quedando significativamente por debajo del umbral, dando como resultado un modelo que cumple con el criterio de **Equalized Odds.**

Esto hace que este proceso de entrenamiento, análisis y selección de mitigadores haya sido **exitoso** y podamos contar con un modelo que cumpla con el criterio de equidad que elegimos para este caso.

5. Comparación final

Comparación del modelo mitigado y el original

Se realizan las matrices de confusión para el modelo inicial y para el modelo que implementa **Prejudice Remover**, esta vez separadas por los grupos de interés (solicitantes de ambos géneros). Las métricas de rendimiento son prácticamente iguales entre los dos modelos. Sin embargo hay ligeros cambios en cuanto a la cantidad de personas que fueron catalogadas como de alto o de bajo riesgo. A continuación observamos los resultados:

Género femenino

En el caso de las mujeres, el modelo mitigado muestra una mejora en la **sensibilidad**: se incrementa el número de verdaderos positivos (de 34 a 37) y se reduce la cantidad de falsos negativos (de 5 a 2). Esto indica una mayor capacidad del modelo para identificar correctamente a las mujeres clasificadas como de bajo riesgo, lo que podría reflejar una corrección de sesgos previos que llevaban a clasificarlas injustamente como de alto riesgo.

Sin embargo, también se observa un aumento leve en los falsos positivos (de 8 a 10) y una disminución en los verdaderos negativos (de 9 a 7), lo cual sugiere un **ligero deterioro en la precisión** del modelo para este grupo. Este comportamiento más "indulgente" puede ser aceptable en contextos donde el objetivo es reducir la discriminación injustificada hacia personas que, históricamente, han sido desfavorecidas.

Género masculino

En el caso de los hombres, el modelo mitigado reduce la cantidad de falsos positivos (de 26 a 24) y aumenta los verdaderos negativos (de 16 a 18), lo que representa una **mejora en la precisión** y una mejor capacidad para identificar correctamente a personas con alto riesgo crediticio.

Por otro lado, se observa una leve pérdida de verdaderos positivos (de 94 a 92) y un aumento en los falsos negativos (de 8 a 10), lo que indica que el modelo ahora **se vuelve un poco más conservador** al clasificar a hombres como de bajo riesgo, permitiendo que algunos casos de bajo riesgo pasen desapercibidos.

Mejoras en Fairness halladas

Los cambios que se muestran en las matrices de confusión previamente citadas hacen notar que el modelo mitigado mejoró la equidad de sus clasificaciones, haciendo que el grupo que no era privilegiado pueda tener una clasificación más justa. De ésta manera, utilizando el modelo mitigado, **3 mujeres más** (el 1.5% del conjunto de pruebas) habrían obtenido un préstamo y **3 mujeres menos** habrían sido erróneamente clasificadas como personas con un alto riesgo.

Balance final

¿Qué ganamos?

- **Reducción** clara de disparidades entre géneros, especialmente en el tratamiento de mujeres.
- **Corrección** de sesgos que afectan más negativamente a mujeres (menos FN).
- **Mejora** en la precisión para hombres (menos FP).
- Un modelo más justo y éticamente alineado, con menor dependencia del atributo sensible.

¿Qué perdimos?

- Ligera **pérdida** de precisión para mujeres (más FP).
- **Disminución** marginal de sensibilidad en hombres (más FN, menos TP).

Por ende, se obtuvo un mejor modelo en términos de equidad, **sin sacrificar sustancialmente el desempeño general**.

Conclusión

El desarrollo de modelos de aprendizaje automático aplicados a decisiones financieras, como la aprobación de créditos, tiene un impacto directo y profundo en la vida de las personas. Un modelo injusto puede reforzar desigualdades existentes y limitar el acceso de ciertos grupos a oportunidades económicas.

En este trabajo, evidenciamos cómo la equidad puede evaluarse de forma rigurosa y cómo herramientas específicas permiten mitigar sesgos detectados. El uso del criterio **Equalized Odds** permitió alinear los objetivos del banco (minimizar riesgo financiero) con principios de equidad social (evitar discriminación por género). Implementar modelos como el obtenido con **Prejudice Remover** en la práctica podría contribuir a una distribución más justa del crédito y a una mayor inclusión financiera.

Desde la perspectiva del banco, este enfoque también tiene un valor estratégico. Al lograr que las tasas de error sean similares entre ambos géneros, el banco se arriesga de forma equitativa al otorgar créditos a personas consideradas como de alto riesgo crediticio, sin que su género influya en dicha probabilidad. Esto implica una gestión del riesgo más justa y transparente, alineada con principios éticos, y puede fortalecer la reputación institucional frente a clientes y organismos de control.

Además, este trabajo destaca la importancia de considerar la equidad como un componente central en el desarrollo de soluciones de **ciencia de datos**, más allá de las métricas tradicionales de rendimiento. De este modo, se promueve un uso ético, responsable y sostenible de la inteligencia artificial en sectores críticos como el financiero.