UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

MITIGACIÓN DE SESGO DE GÉNERO EN UN MODELO DE CALIFICACIÓN CREDITICIA

Tesis sometida a la consideración de la Comisión del Programa de Posgrado en Computación e Informática para optar al grado y título de Maestría Académica en Computación e Informática

RICARDO CORRALES BARQUERO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2023

DEDICATORIA

A mis padres. Gracias por su apoyo.

AGRADECIMIENTO

Agradezco a todas las personas que me han ayudado y apoyado en el proceso. A mi directora de tesis, por siempre confiar en mí. A mis lectores por sus muy valiosos aportes. A mis padres por su apoyo. A Silvia por su ayuda con la tabulación de algunos datos. A la institución que me facilitó acceso a los datos, por su preocupación por mejorar en el ámbito social.

Esta tesis fue aceptada por la Comisión del Programa de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Computación e Informática.

EDGAR ENRIQUE CASASOLA MURILLO (FIRMA)
PERSONA FISICA, CPF-01-0729-0430.
Fecha declarada: 26/02/2024 09:02:15 AM
Esta es una representación gráfica únicamente, verifique la validez de la firma.

Dr. Edgar Casasola Murillo
Representante de la Decana
Sistema de Estudios de Posgrado

ELENA GABRIELA BARRANTES SLIESARIEVA (FIRMA)
PERSONA FISICA, CPF-01-0866-0165.
Fecha declarada: 19/02/2024 01:25:57 PM
Esta es una representación gráfica únicamente, verifique la validez de la firma.

Dra. Gabriela Barrantes Sliesarieva
Directora de Tesis

GABRIELA MARIN RAVENTOS (FIRMA)
Firmado digitalmente por GABRIELA MARIN RAVENTOS (FIRMA)
Fecha: 2024.02.19 16:04:20 -06'00'

Dra. Gabriela Marín Raventós
Asesora

ALVARO JOSE GUEVARA VILLALOBOS (FIRMA)
Firmado digitalmente por ALVARO JOSE GUEVARA VILLALOBOS (FIRMA)
Fecha: 2024.02.20 12:06:51 -06'00'

Dr. Álvaro Guevara Villalobos
Asesor

GUSTAVO LOPEZ HERRERA (FIRMA)
PERSONA FISICA, CPF-01-1415-0174.
Fecha declarada: 21/02/2024 10:31:26 AM
Esta es una representación gráfica únicamente, verifique la validez de la firma.

Dr. Gustavo López Herrera
Director
Programa de Posgrado en Computación e Informática

RICARDO CORRALES BARQUERO (FIRMA)
Digitally signed by RICARDO CORRALES BARQUERO (FIRMA)
Date: 2024.02.19 12:32:33 -06'00'

Ricardo Corrales Barquero
Candidato

ÍNDICE

RESUMEN EN ESPAÑOL

El presente documento expone un trabajo realizado sobre un conjunto de datos y un modelo matemático para apoyo de toma de decisiones en el proceso de crédito para clientes ya constituidos en un banco comercial de Costa Rica. El objetivo principal consistió en evaluar alternativas para mitigar los sesgos de género presentes en el modelo. Para lograrlo, se comenzó por identificar posibles fuentes de sesgo en el modelo, entre las cuales se identificaron posibles sesgos de tratamiento dispar, asociación, selección, sesgo malicioso y sesgo de automatización. Seguidamente se midieron dichos sesgos en más detalle, encontrando que son pequeños, excepto quizá por el sesgo de selección. En tercer lugar, se construyeron modelos alternativos que mitigaran estos sesgos, para finalmente, evaluar la diferencia tanto en las medidas de justicia que se utilizaron como en el rendimiento de los modelos alternativos respecto al original para determinar el que provee mayor valor al negocio. Aquí se encontró que las ganancias son menores y que lo que podría valer más la pena es mantener el modelo actual e investigar otros modelos de calificación crediticia utilizados en otras etapas del proceso de otorgamiento de crédito.

RESUMEN EN INGLÉS

This document presents a project carried out on a dataset and a mathematical model to support decision making in the credit process for established clients in a commercial bank in Costa Rica. The main objective was to evaluate alternatives to mitigate the gender biases present in the model. To achieve this, possible sources of bias in the model were identified, among which possible disparate treatment, association, selection, malicious, and automation biases were identified. These biases were then measured in more detail, finding that they are small, except perhaps for the selection bias. Thirdly, alternative models were built to mitigate these biases, to finally evaluate the difference both in the fairness measures that were used and in the performance of the alternative models compared to the original to determine the one that provides greater value to the business. Here, it was found that the gains are minor and that what could be more worthwhile is to maintain the current model and investigate other credit scoring models used in other stages of the credit granting process.

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ABREVIATURAS

AOD: *Average Odds Difference*

BEL: *Bayesian Extended Lift*

BUST: *Bottom-Up Stress Testing*

CAP: *Cumulative Accuracy Profile*

CPD: *Conditional Probability Distribution*

EOD: *Equal Odds Difference*

LIME: *Local Interpretable Model-agnostic Explanations*

SUGEF: Superintendencia General de Entidades Financieras

CAPÍTULO I. INTRODUCCIÓN

El acceso al crédito es de vital importancia en la sociedad moderna. Algunos estudios apuntan a que este juega un papel importante en la reducción de la pobreza [1]. Más aun, hay quienes argumentan que el acceso al crédito debería ser considerado un Derecho Humano, entre ellos, el ganador del Premio Nobel de la Paz de 2006, Muhammad Yunus [2]. Esta posición se basa en el principal argumento de que el acceso al crédito facilita el acceso a otros Derechos Humanos, como la comida, la vivienda, la educación y la salud. Si bien, esta posición es objeto de debate, no se puede negar que el acceso al crédito influye positivamente en los indicadores clásicos del desarrollo de las naciones [3]. Se puede argumentar entonces que garantizar un acceso equitativo al crédito tiene un impacto considerable en el bienestar de las personas. Sin embargo, dicho acceso se puede ver mermado debido a consideraciones que se realizan durante el proceso de otorgamiento del crédito.

El proceso moderno de otorgamiento del crédito se suele dividir, para su estudio, en varias etapas. Estas varían de autor a autor, pero, a modo ilustrativo, se explica a continuación el modelo encontrado en [4]. En una primera etapa, conocida como **promoción**, la entidad financiera anuncia sus productos de crédito a potenciales clientes por medio de campañas de publicidad. Seguidamente, en la segunda etapa, llamada **evaluación**, dichos clientes potenciales se acercan a la institución para solicitar sus productos. Esta última evalúa a quienes hacen la solicitud para determinar si es factible dentro de su modelo de negocio otorgarles un préstamo. El resultado de esta etapa es esa decisión de otorgar o no un crédito al cliente. En tercer lugar, durante la etapa de **aprobación**, la entidad determina las condiciones bajo las cuales se va a otorgar el crédito (por ejemplo, los periodos de pago, la tasa de interés, entre otras). Luego, en la etapa de **desembolso**, se le da el dinero al cliente. En último lugar, durante la etapa de **cobranza**, el acreedor realiza campañas para cobrar de vuelta el préstamo a sus deudores.

Durante todas las etapas del proceso de crédito, se hace necesario para la institución evaluar a sus clientes. Esto se hace con el fin de mitigar el llamado "riesgo de crédito". El término **riesgo de crédito** se refiere al riesgo financiero de que un cliente no pague de vuelta su crédito parcial o totalmente (esto se conoce como entrar en ***default***, o caer en **impago**) [5]. La evaluación de los clientes durante el proceso de crédito le permite a la entidad tomar distintas decisiones durante todas estas etapas. Por ejemplo, en la etapa de promoción se decide a quiénes dirigir cierto tipo de campañas publicitarias; durante la evaluación se decide a quién otorgarle un crédito y a quién no; durante la aprobación y el desembolso se toma decisiones sobre las

condiciones pactadas; y durante la cobranza se toma decisiones sobre cómo distribuir los recursos (humanos y materiales) para hacer el cobro a distintos clientes [6].

Históricamente, según explica Anderson [6], este tipo de decisiones se tomaban en base a la experiencia de la persona prestamista (cuando este era una persona) o del empleado de la institución que estaba encargado de cada etapa. No es hasta inicios de la Segunda Guerra Mundial cuando en Estados Unidos, debido a la migración de muchos hombres a la lucha en Europa (y debido a que aún no se reclutaban mujeres para este tipo de puestos bancarios), comienzan a dejarse por escrito las primeras reglas para la toma de decisiones en el proceso de crédito. Efectivamente, esto generó lo que hoy se podría considerar como una especie de "sistemas expertos" escritos en papel. En 1946, se considera que aparece por primera vez una técnica estadística para la evaluación de deudores según su riesgo de crédito. Esto se da cuando E. F. Wonderlic, presidente de la Corporación Household Finance, desarrolla el "*Credit Guide Score*". Sin embargo, no es hasta la década de 1960 cuando estas técnicas comienzan a tomar auge en la industria. Esto se debió a la expansión en el acceso a computadoras electrónicas, por lo que, por primera vez las entidades prestatarias fueron capaces de automatizar el manejo de la información de sus clientes. En la década de 1980 se da una expansión geográfica de este tipo de técnicas a países fuera de los Estados Unidos. Asimismo, se da también una expansión en cuanto al tipo de técnicas utilizadas, incluyendo algunas técnicas que hoy día se catalogan dentro de lo que conocemos como *aprendizaje de máquina*. Las técnicas estadísticas y de aprendizaje de máquina utilizadas para evaluar deudores es lo que hoy conocemos como **calificaciones de crédito**. Anderson [6] define las calificaciones de crédito como herramientas numéricas para la evaluación de los clientes según su riesgo de crédito durante el proceso de un préstamo.

Las calificaciones crediticias modernas hacen uso de muchos tipos de datos para evaluar a los deudores. Los datos específicos que se utilizan dependen del modelo y de los intereses de la institución acreedora; sin embargo, se puede mencionar que es común usar detalles financieros de los clientes, información de su comportamiento de pago (e.g. si el cliente ha tenido atrasos, si ha caído en impago en otras deudas, entre otros), datos sobre el empleo y salario del cliente, características sociodemográficas, entre otros [5], [6].

Ahora bien, el uso de las calificaciones de crédito no ha surgido sin problemas. Por ejemplo, se puede mencionar temas de falta de privacidad (se han observado modelos capaces de predecir información sensible, como la raza de la persona) [7], [8] y falta de explicabilidad (se han usado modelos en los cuáles es difícil determinar por qué generan ciertos resultados) [9], [10]. El problema con el que se tuvo la intención de trabajar en este proyecto fue el **sesgo**. Para ilustrar

Figura 1.1. Ilustración del concepto de sesgo.

este problema, note que el objetivo de una calificación de crédito es poder pronosticar cuáles deudores caerán en un impago y cuáles no. En este caso, existe una "verdad": hay deudores que pagarán de vuelta su crédito y otros que no. El objetivo de la calificación de crédito es entonces poder asignar de la manera más correcta posible una etiqueta de "alto riesgo" o "bajo riesgo" a cada cliente, de tal manera que aquellos clientes etiquetados como de alto riesgo tiendan a ser aquellos que caen en un impago, y aquellos clientes etiquetados como de bajo riesgo tiendan a ser quienes paguen de vuelta el préstamo sin problemas. Si se separa, como en la figura 1.1, a los clientes por quienes finalmente caen en impago y quienes no, y asumiendo que un modelo de calificación de crédito con alta exactitud ha etiquetado a estos clientes como "de alto riesgo" y "de bajo riesgo", es posible ver que existe una diferencia en la distribución de las etiquetas entre los grupos. Esta diferencia en la distribución de etiquetas es lo que se suele conocer como un sesgo [11], [12].

El sesgo no siempre es un problema. De hecho, en el ejemplo anterior, se podría catalogar a este como un sesgo deseable, puesto que el fin del modelo es justamente poder distinguir entre personas que caerán en impago y personas que no. Sin embargo, también existen sesgos indeseables. Estos aparecen cuando la diferencia en la distribución de etiquetas ocurre entre grupos cuyos atributos que los definen se consideran socialmente sensibles [11]. Por ejemplo, si a las mujeres se les asigna etiquetas de "alto riesgo" con más frecuencia que a los hombres, esto es considerado como una práctica discriminatoria, y por tanto este sería un sesgo indeseable.

Este tipo de sesgos se consideran discriminatorios pues tienen un impacto directo sobre las personas contra las cuales el sesgo opera. La obtención de una calificación de crédito poco favorable puede llevar a la institución acreedora a tomar distintas decisiones en cada etapa del crédito: por ejemplo, podría no dirigir campañas publicitarias a personas que considera, injustamente, como de alto riesgo. A un cliente con una calificación injustamente desfavorable

también podría negársele el acceso al crédito u otorgársele con condiciones más restrictivas, tales como tasas de interés más altas o plazos más cortos [5], [6].

Es importante recordar que detrás de estas decisiones hay un componente muy importante generado por un modelo automático (ya sea estadístico o de aprendizaje de máquina) [6]. Los sesgos indeseables descritos anteriormente se han visto ya en diversos casos. A modo de ejemplo, se puede mencionar el caso de una herramienta experimental de aprendizaje de máquina para contratación que se estuvo desarrollando en Amazon. Esta tuvo que ser descartada debido a que se encontró que estaba presentando sesgos injustos en contra de las mujeres [12], [13]. Un segundo caso muy estudiado ha sido el del software ProPublica en el sistema judicial de Estados Unidos. Esta es una herramienta que utiliza modelos estadísticos para generar una calificación cuyo objetivo es pronosticar la probabilidad de re-incidencia de personas que han cometido algún acto criminal. Ya se ha demostrado que ProPublica presenta un sesgo importante que desfavorece a las personas negras [14]. Un tercer caso es el de las herramientas de reconocimiento facial. Un estudio por Joy Buolamwini de MIT, ha demostrado que los errores en este tipo de herramientas aumentan hasta en un 35% cuando la persona que se intenta detectar es una mujer negra en comparación con hombres blancos [15].

Particularmente, en el contexto de las calificaciones de crédito, este tipo de discriminación se ha encontrado también. A modo de ejemplo, se puede citar el caso de Jamie Heinemeier Hansson, una mujer estadounidense que, a pesar de tener una mejor calificación de crédito nacional que su marido, y además compartir con él cuotas iguales en sus propiedades y presentar con él declaraciones de impuestos en conjunto, obtuvo peores condiciones que su esposo al solicitar una tarjeta de crédito de Apple. Cuando se investiga esta situación, se encuentra que hay un algoritmo de aprendizaje de máquina tomando este tipo de decisiones y ninguna persona responsable de este modelo sabe por qué se tomó una decisión que en este caso se consideró discriminatoria [16]. Otros estudios interesantes a considerar son, por ejemplo, el reporte de Goldman Sachs, donde concluyen que existe una brecha global de 287,000 millones de dólares estadounidenses en el acceso a crédito entre pymes propiedad de mujeres y pymes propiedad de hombres [17]. Es más, se concluye que la región que más contribuye a esta brecha es América Latina. El mismo reporte muestra que más del 70% de pymes propiedad de mujeres no cuentan con un acceso adecuado a servicios financieros. El reporte [18] del Banco Interamericano de Desarrollo, refuerza las conclusiones del reporte de Goldman Sachs y además atribuye estas brechas a las prácticas discriminatorias de las entidades financieras, entre otras razones.

En este punto, cabe preguntarse de dónde provienen estos sesgos. Durante la presente investigación, se han identificado tres posibles fuentes de sesgo para los modelos de calificación crediticia. En primer lugar, la fuente más comúnmente estudiada son los datos de entrenamiento. Por ejemplo, si se entrena un modelo de calificación crediticia sobre una base de datos históricos en los cuales se le han otorgado más créditos a hombres que a mujeres, el modelo puede aprender y reproducir dichos sesgos. Asimismo, datos erróneos pueden exacerbar este tipo de sesgos. En segundo lugar, existen prácticas que pueden generar sesgos durante el entrenamiento de los modelos. Se puede mencionar, por ejemplo, el no permitir que un modelo converja por completo. Esta práctica podría generar sesgos discriminatorios debido a que en las primeras iteraciones del proceso de entrenamiento, el modelo podría aprender solamente características muy generales de la población y no encontrar patrones más específicos que le permitan mitigar ese sesgo. Por último, está el sesgo generado durante el uso del modelo. Dado que estas herramientas se utilizan en contextos sociales, la persona usuaria del modelo puede introducir sus sesgos en la decisión cuando hace uso de este.

Para este proyecto se contó con acceso a un modelo de calificación crediticia de un banco comercial de Costa Rica, así como a los datos anonimizados que se usaron para entrenarlo y hacerle pruebas. Este modelo es utilizado solamente para evaluar clientes ya establecidos dentro de la institución a través de todas sus operaciones de crédito (este tipo de calificación crediticia se conoce con el nombre de **calificación de cliente** [6]).

1.1 Justificación

La mitigación de sesgos discriminatorios en modelos de calificación crediticia no solo beneficia a las personas históricamente marginadas, sino también al negocio dueño del modelo: el uso de modelos más justos puede ayudar a los acreedores a mejorar la calidad de su cartera, dado que se evalúa a los y las clientes de acuerdo a su verdadero nivel de riesgo y no a sesgos irracionales [19]. Asimismo, la mitigación de sesgos puede impactar positivamente el riesgo reputacional de la institución. Una tercera razón por la cual un intermediario financiero podría querer mitigar los sesgos en su modelo es la obligación de cumplir con leyes y regulaciones; por ejemplo, la ley conocida como "Fair Housing Act" en Estados Unidos o el artículo 14 de la Convención Europea en Derechos Humanos prohíben la discriminación en ciertos aspectos del otorgamiento de crédito [20].

El tratamiento del sesgo, sin embargo, no es un problema trivial. Los modelos utilizados en la actualidad no son perfectos. Es importante para el negocio conocer cuál es el grado de error esperable de sus modelos de calificación crediticia para así poder mantener un control sobre las

posibles pérdidas que se puedan dar debido a la caída en impago de clientes erróneamente clasificados como de bajo riesgo, así como para evitar incurrir en costos de oportunidad de negar el crédito a clientes de bajo riesgo. Al nivel de pérdidas que la institución está dispuesta a incurrir a cambio de las ganancias que espere obtener se le conoce con el nombre de **apetito de riesgo** [5]. Es importante notar que el apetito de riesgo y el rendimiento[1] del modelo de calificación crediticia que se usa están relacionados: es importante para el negocio mantener un nivel de rendimiento en el modelo que se ajuste a su apetito de riesgo. Cuando se hace un tratamiento del sesgo indeseable es necesario equilibrar la mitigación de este respecto a la exactitud del modelo utilizado. En un caso extremo, se podría eliminar todo rastro de sesgo indeseable asignando a todos los clientes la etiqueta de "bajo riesgo", lo cual tendría repercusiones completamente inaceptables en los resultados del modelo.

A continuación se presenta la pregunta de investigación y los objetivos de este trabajo, seguidos de la estructura del presente documento.

1.2 Pregunta de investigación

¿Cómo se puede mitigar el sesgo de género presente en un modelo de calificación crediticia, manteniendo resultados aceptables para el negocio?

1.3 Objetivos

El objetivo general de esta investigación fue:

Construir y evaluar un modelo de calificación de crédito que reduzca los sesgos de género, manteniendo resultados aceptables para el negocio.

Los objetivos específicos de la presente investigación incluyerpn:

1. Determinar posibles fuentes de sesgo de género en el modelo actual de calificación crediticia.
2. Medir el grado de sesgo de género en el modelo actual elegido.
3. Construir modelos de calificación crediticia alternativos con el fin de reducir el sesgo de género del modelo actual.

[1] Entiéndase por el momento "rendimiento" como una medida de qué tantos errores tiene el modelo a la hora de clasificar a los clientes como de alto o bajo riesgo. Más adelante, en el marco conceptual, se define una medida exacta del rendimiento de un modelo para uso en este estudio.

4. Evaluar los modelos alternativos construidos en cuanto a su grado de sesgo de género y rendimiento.

1.4 Estructura del documento

A continuación se detalla la estructura del presente documento. El capítulo II presenta el marco conceptual, donde se explican las principales nociones que se usaron a lo largo del estudio. El capítulo III explica los antecedentes encontrados en la literatura para este trabajo. El capítulo IV presenta la metodología que se siguió para cumplir con los objetivos propuestos. El capítulo V presenta los resultados obtenidos al aplicar cada paso explicado en la metodología y una breve discusión de los mismos. Finalmente, se dan las últimas observaciones y se explican las limitaciones del estudio y trabajo futuro en el capítulo VI.

CAPÍTULO II. MARCO CONCEPTUAL

En este capítulo se presentan los principales conceptos que se usaron durante la investigación. Se comienza hablando del entrenamiento y evaluación de modelos de clasificación, iniciando con los modelos de regresión logística, puesto que este es el tipo de modelo que está en uso actualmente y algunos de los análisis realizados dependen del funcionamiento específico de este tipo de modelos. Seguidamente, se explican las técnicas de regularización Ridge y Lasso que se pueden aplicar sobre este tipo de modelos y que también proveen detalles relevantes en algunas partes del estudio. Luego, se explican lo que son las redes bayesianas y los algoritmos que se utilizaron en este estudio para generarlas cuando fue necesario. En tercer lugar, se presenta la métrica de exactitud conocida con el nombre de Coeficiente de Gini, la cual se usó para evaluar el rendimiento de los modelos en estudio. Seguidamente, se explica el método *Bootstrapping* para obtener métricas sobre los modelos. Este método se utilizó en el estudio para dar mayor confianza estadística a los resultados finales. Luego, se explican los distintos tipos de sesgo según su fuente. Después, se habla sobre las distintas definiciones de justicia que se encuentran en la literatura y sus métricas asociadas, y se explica en detalle las métricas usadas durante este estudio (la justificación de por qué se usaron estas métricas se encuentra en los capítulos de metodología y resultados). En penúltimo lugar, se presentan los tipos de técnicas para mitigar el sesgo según su etapa de aplicación, así como las técnicas de mitigación de sesgo usadas en este estudio (cuya escogencia también se justifica en la metodología y resultados). Finalmente, se presentan varias pruebas estadísticas importantes que se utilizaron a lo largo del estudio.

2.1 Entrenamiento y evaluación de modelos de clasificación

En esta sección se explican los dos principales modelos de clasificación utilizados en este estudio; a saber, la regresión logística y las redes bayesianas.

2.1.1 Regresión logística

La regresión logística es un método estadístico usado para la clasificación binaria. El objetivo de la regresión logística es estimar la probabilidad de que cada observación pertenezca a una clase u otra [21]. En el presente caso de estudio, esto es la probabilidad de que un cliente caiga en impago. Este tipo de modelo hace uso de la siguiente función para estimar dicha probabilidad:

$$h(X) = P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

donde Y denota la variable a predecir (e.g. el impago), $X = (X_1, X_2, ..., X_p)$ denota el vector de variables predictoras y $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ son coeficientes (también llamados **parámetros** de la regresión) que se estiman a través de algún método particular [21]. No se explica aquí los métodos específicos que se usan para hacer la estimación de los coeficientes debido a que se salen del alcance del estudio; basta con mencionar que todos los métodos aquí utilizados son métodos iterativos que intentan converger a una solución que minimice la siguiente **función objetivo**:

$$logloss(D) := \sum_{(x,y)\in D} -y \cdot log(h(x)) - (1-y) \cdot log(y - h(x))$$

donde D denota el conjunto de datos de entrenamiento, el cual contiene valores de la forma (x, y) con x un vector de variables predictoras y y la variable a predecir [21].

A continuación, se explican tres de los métodos de **regularización** más utilizados en el entrenamiento de regresiones logísticas. El objetivo de este tipo de métodos es reducir la complejidad del modelo entrenado por medio de la modificación de la función objetivo, penalizando valores altos en los parámetros de la regresión. Muchas veces estas técnicas obtienen valores de cero para los parámetros, por lo que en cierta forma realizan una selección automática de variables.

2.1.1.1 Regularización *Ridge*

La regularización *Ridge* (también conocida como regularización L2) consiste en agregar a la función objetivo $L(D)$ un término de la siguiente forma:

$$L(D) = logloss(D) + \lambda \sum_{i=1}^{p} \beta_i^2$$

donde λ termina siendo un hiperparámetro del modelo que indica la importancia que se le dará a la regularización. Note que el efecto de minimizar esta nueva función objetivo es que no solo es importante obtener un buen ajuste a los datos, sino que también hay un cierto grado de importancia en mantener los parámetros del modelo con valores bajos [21].

2.1.1.2 Regularización *Lasso*

La regularización *Lasso* (también conocida como regularización L1) es muy similar a la regularización *Ridge*. La única diferencia es que en vez de tomar los cuadrados de los parámetros del modelo, usa los valores absolutos de estos:

$$L(D) = logloss(D) + \lambda \sum_{i=1}^{p} |\beta_i|$$

El principal efecto que se obtiene diferente a *Ridge* es que *Lasso* tiende a llevar a cero algunos de los coeficientes de variables predictoras correlacionadas (al azar entre cada grupo de

variables correlacionadas), mientras que Ridge mantiene todos los coeficientes a pesar de hacerlos pequeños [21].

2.1.1.3 *Elastic Net*

La técnica conocida como *Elastic Net* es una combinación del uso de las dos regularizaciones anteriores. Se agrega a la función objetivo el siguiente término:

$$L(D) = logloss(D) + \lambda \sum_{i=1}^{p} \left(\alpha \beta_i^2 + (1 - \alpha) \left| \beta_i \right| \right)$$

Note que entonces α es otro hiperparámetro en este caso, que denota la importancia relativa entre la regularización L1 y la regularización L2. En la práctica, los hiperparámetros α y λ suelen ser calibrados en un proceso de validación cruzada durante el entrenamiento del modelo [21].

A continuación se habla sobre redes bayesianas. Este es otro tipo de modelo estadístico que puede servir para hacer inferencia o predicción sobre datos tabulares y está especialmente diseñado para modelar correlaciones y relaciones de causalidad entre variables. Este tipo de modelo es utilizado en muchas métricas de justicia e incluso en algunas técnicas de mitigación del sesgo, por lo que vale la pena entender su funcionamiento particular. También se explican algunas técnicas importantes para entrenar este tipo de modelos, debido a que algunas técnicas de mitigación de sesgo, explicadas más adelante, hacen uso de estas.

2.1.2 Redes bayesianas

Una red bayesiana es un modelo estadístico que pretende modelar tanto las correlaciones entre variables como las relaciones causales entre estas. Se dice que una variable aleatoria X está **relacionada causalmente** a un conjunto de variables aleatorias Y_1, Y_2, ..., Y_n si X es una función estocástica de Y_1, Y_2, ..., Y_n, es decir la función de densidad de probabilidad de X está determinada para cada posible combinación de valores de Y_1, Y_2, ..., Y_n [22].

Las redes bayesianas modelan estas relaciones por medio del uso de un grafo acíclico dirigido (conocido como **grafo de causalidad**) donde cada nodo representa una variable en estudio y cada arista del grafo representa la existencia de una correlación significativa entre las variables que une. Además, la dirección de las aristas se define de tal manera que cada variable en el grafo esté relacionada causalmente a sus padres (y solamente a sus padres). Las redes bayesianas incluyen una función llamada función de distribución de probabilidad condicional (**CPD** por sus siglas en inglés) para cada vértice del grafo que describe la relación causal con sus padres [22].

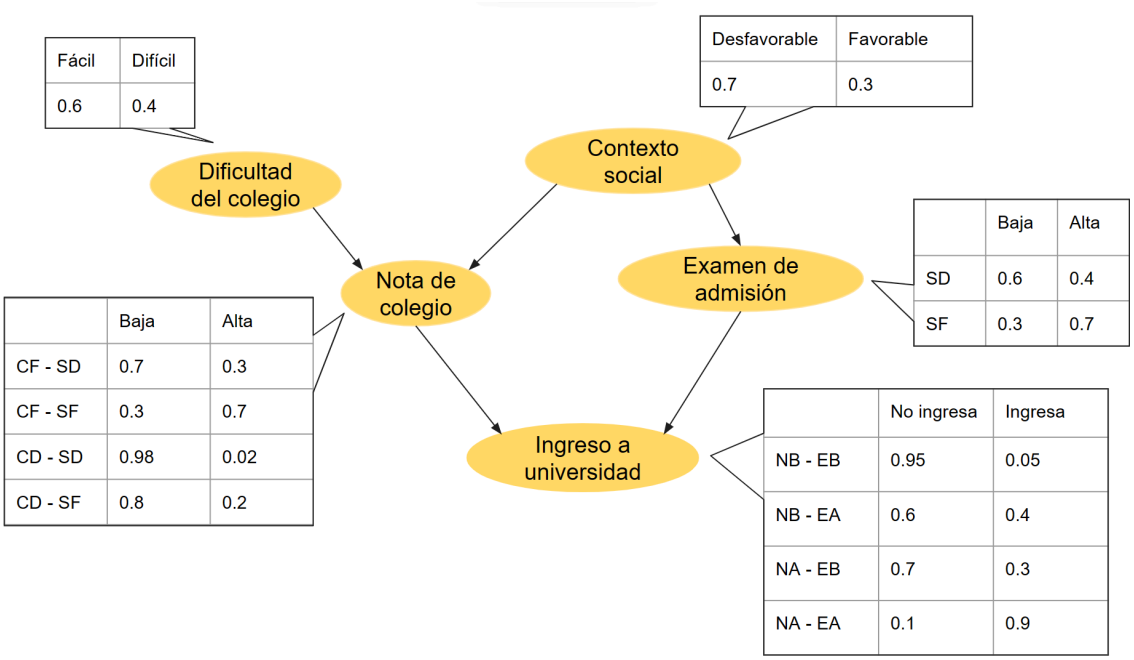

Figura 2.1. Ejemplo de una red bayesiana (elaboración propia).

Para efectos de este estudio, se trabajó solamente con variables categóricas cuando se hizo uso de las redes bayesianas, por lo que, en este documento, las CPDs siempre se pueden describir por medio de una tabla en la cual se especifica la probabilidad de que la variable en estudio tome sus posibles valores dadas las posibles combinaciones de valores de sus padres.

La figura 2.1 muestra un ejemplo de una red bayesiana que modela información relacionada a los estudiantes de una universidad. Se incluyen como variables aleatorias: la dificultad del colegio en el cual estudió el estudiante, la nota de presentación que obtuvo en el colegio, el contexto social del cual proviene el estudiante (si es favorable o desfavorable), la nota obtenida en el examen de admisión y si este ingresa o no a la universidad. Las relaciones causales codificadas en el grafo muestran que es posible asignar una probabilidad, por ejemplo, a la nota de presentación que obtiene un estudiante en el colegio en base a la dificultad del colegio y a su contexto social. Por ejemplo, si un estudiante viene de un colegio difícil y un contexto social desfavorable, tendrá una probabilidad de 0.98 de obtener una nota de presentación del colegio baja y un 0.02 de obtener una nota alta, mientras que si viene de un colegio fácil y un contexto social favorable, tendrá una probabilidad de 0.3 de obtener una nota baja y de 0.7 de obtener una nota alta. Esta información se ve codificada en la tabla (CPD) asociada al nodo "Nota de colegio".

Dada una red bayesiana, es posible hacer **inferencia** sobre esta, es decir, dados valores para un subconjunto de las variables en la red, es posible determinar la probabilidad de obtener

valores dados para cualquier otro subconjunto de valores en la red. De esta manera, instanciando valores para todas las variables excepto una, se puede usar una red bayesiana para predecir el valor de la variable no instanciada [22]. En el presente caso de estudio, esto se traduce en predecir la probabilidad de impago de un cliente dados valores para las variables predictoras seleccionadas.

Existen métodos para obtener una red bayesiana a partir de un conjunto de datos (i.e. entrenar una red bayesiana). Estos métodos se basan, en una primera etapa, en obtener una estructura para el grafo de causalidad que se ajuste de alguna forma a los datos (i.e. **aprendizaje de estructura**), y en una segunda etapa, en obtener valores para las CPDs que se ajusten también a los datos de entrenamiento, dada la estructura aprendida (i.e. **aprendizaje de parámetros**) [22]. A continuación, se explica la métrica K2, que se usó para evaluar la bondad de ajuste de la estructura de una red bayesiana al conjunto de datos de entrenamiento, y luego se explican las técnicas de aprendizaje de estructura y de parámetros usadas en este estudio.

2.1.2.1 Métrica K2

La métrica K2 fue planteada originalmente en [23] para evaluar la bondad de ajuste de una red bayesiana a un conjunto de datos de entrenamiento. No se provee aquí la fórmula exacta, debido a que su derivación es compleja y se sale del alcance de este estudio. Tan solo se dirá que la métrica mide la probabilidad conjunta de la estructura de una red bayesiana y un conjunto de datos de entrenamiento con las mismas variables, asumiendo que:

1. Todas las variables son discretas.
2. Todas las observaciones en el conjunto de datos de entrenamiento son independientes.
3. La probabilidad de tener cualquier combinación de valores en los CPDs de la red bayesiana, dada una estructura específica es uniforme. Es decir, que se considera que existe la misma probabilidad de obtener cualquier combinación de CPDs dada la estructura presentada.

2.1.2.2 Algoritmo *Hill Climb Search* para aprendizaje de estructura

Dado un conjunto de datos de entrenamiento, el objetivo del aprendizaje de estructura es obtener un grafo de causalidad (sin CPDs aún) que modele las relaciones causales entre los atributos del conjunto de datos. Una forma de realizar esto es usando una métrica que evalúe qué tan bien modela un grafo de causalidad dado al conjunto de datos dado, por ejemplo, la métrica K2. Debido a que realizar una búsqueda exhaustiva en el espacio de todas las posibles estructuras dado un conjunto fijo de nodos (i.e. los atributos del conjunto de datos) es un problema altamente complejo (NP-*hard*), se han planteado otros algoritmos que funcionan aproximadamente bien [22].

Tabla 2.1. Algoritmo *Hill Climb Search* para aprendizaje de estructuras de redes bayesianas. Tomado de [23].

```
hill_climb_search(estructura_inicial, score)
% Algoritmo de búsqueda ávido para entrenamiento de estructuras de redes bayesianas.
%
% Entradas:
%     - estructura_inicial: una estructura inicial del grafo
%     - score: función de puntuación usada para evaluar la bondad de ajuste del modelo

mejor_score <- -infinito
mejor_estructura <- estructura_inicial

repetir:
    mejor_score_vecindario <- score(mejor_estructura)
    mejor_estructura_vecindario <- mejor_estructura

    estructura_actual = mejor_estructura
    para cada v1, v2 en estructura_actual.vertices:
        si (v1, v2) o (v2, v1) está en estructura_actual.aristas:
            estructura <- estructura_actual.borrar_arista(v1, v2)
            si score(estructura) > score(mejor_estructura_vecindario):
                mejor_score_vecindario <- score(estructura)
                mejor_estructura_vecindario <- estructura

            estructura <- estructura_actual.reversar_arista(v1, v2)
            si score(estructura) > score(mejor_estructura_vecindario):
                mejor_score_vecindario <- score(estructura)
                mejor_estructura_vecindario <- estructura

        si no:
            estructura <- estructura_actual.agregar_arista(v1, v2)
            si score(estructura) > score(mejor_estructura_vecindario):
                mejor_score_vecindario <- score(estructura)
                mejor_estructura_vecindario <- estructura

    si mejor_estructura_vecindario = estructura_actual:
        salir
    si no:
        mejor_estructura <- mejor_estructura_vecindario
        mejor_score <- mejor_score_vecindario

retornar mejor_estructura
```

El algoritmo *Hill Climb Search* se muestra en la tabla 2.1. Este es un algoritmo ávido (*greedy*) que comienza con una estructura inicial para el grafo e iterativamente busca grafos que mejoren la métrica (K2 en este caso) en el vecindario inmediato del grafo [22]. Este vecindario está definido como el conjunto de todos los posibles grafos que se pueden derivar del actual usando una de las tres siguientes operaciones básicas:

- agregar una arista
- eliminar una arista
- cambiar el sentido de una arista.

2.1.2.3 Algoritmo *Maximum Likelihood* para aprendizaje de parámetros

Finalmente, el último paso para entrenamiento de una red bayesiana es la estimación o aprendizaje de los parámetros de esta. Con "parámetros", en este caso, hacemos referencia a las entradas de las tablas que definen las CPDs.

El algoritmo de probabilidad máxima (*Maximum Likelihood*) suele ser considerado el más sencillo y consiste en asumir que las frecuencias conjuntas observadas en el conjunto de datos corresponden a las probabilidades conjuntas correspondientes [22]. En el ejemplo de la figura 2.1, se habrían obtenido las entradas de la tabla del nodo "Examen de admisión" simplemente calculando la frecuencia con la que un estudiante obtiene una nota alta o baja en el examen de admisión para cada posible valor de la variable "Contexto social".

Ahora bien, como se mencionó en la introducción de este capítulo, es debido presentar la métrica que se usó para medir el rendimiento de los modelos, a saber, el coeficiente de Gini. La escogencia de esta métrica se debe a que esta es la que ya se utiliza en la institución en estudio para seleccionar los modelos actualmente, y suele ser un estándar en la literatura.

2.1.3 Evaluación de modelos de clasificación con el coeficiente de Gini

El coeficiente de Gini es una medida que se utiliza para evaluar el rendimiento de modelos de clasificación. Este constituye una medida de qué tan cercano es un modelo de clasificación a un modelo perfecto (i.e. que clasifica el 100% de las muestras correctamente), y qué tan lejano es a un modelo aleatorio (i.e. que clasifica aleatoriamente las muestras). Existen distintas formas equivalentes de definir el coeficiente de Gini. Presentamos aquí la forma explicada en [24], la cual utiliza la llamada curva CAP (*Cummulative Accuracy Profile*) para definir el coeficiente.

Observe la figura 2.2. Se considera el caso de un clasificador binario. El eje horizontal corresponde al total de la población de un conjunto de prueba. En el presente caso de estudio, esta población correspondería a todos los posibles deudores del banco. Estos se deben colocar en orden según la calificación otorgada por el modelo en consideración, es decir, primero aquellos clientes que el modelo estima tienen mayor probabilidad de caer en impago y por último los que el modelo estima que tienen menor probabilidad de caer en impago. El eje vertical representa solo a aquellos clientes que efectivamente caen en impago. Note que ambos ejes han sido normalizados para mostrar porcentajes: en el eje horizontal, 100% representa al 100% de las observaciones del

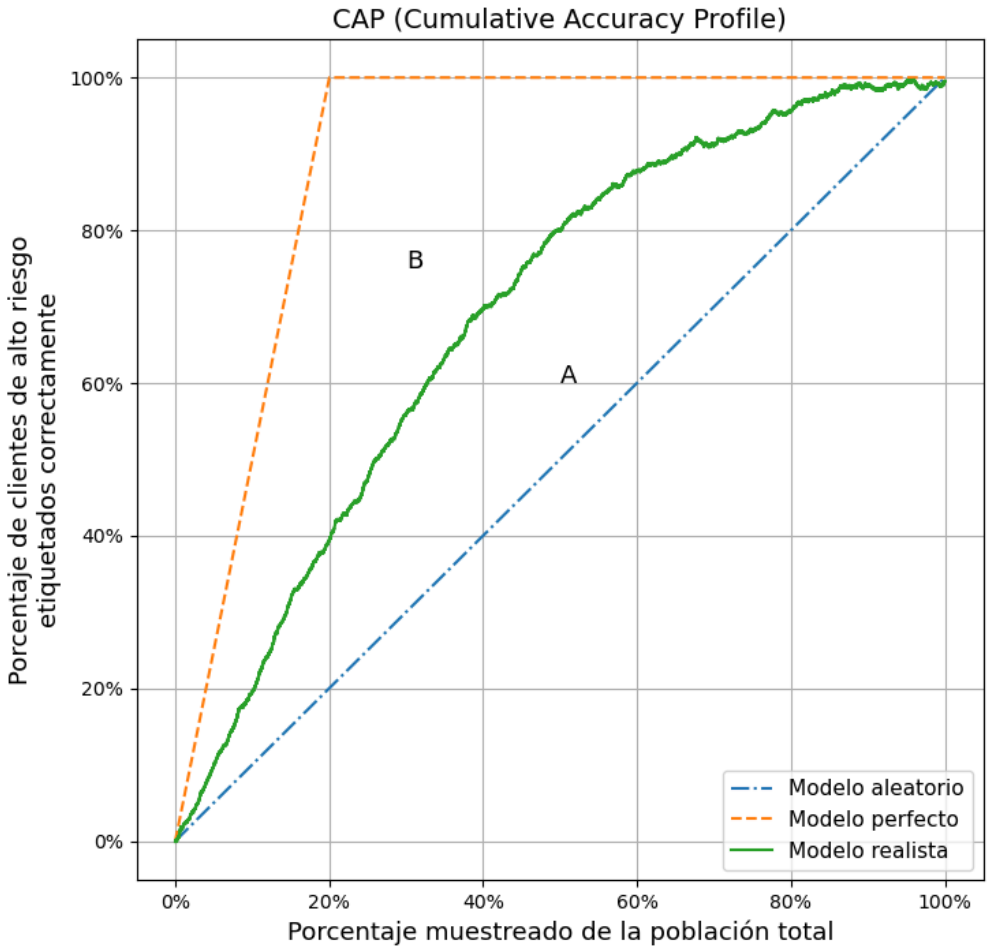

Figura 2.2. Curva CAP.

conjunto, mientras que en el eje vertical, 100% representa al total de los clientes en el conjunto de prueba que caen en impago.

La curva con puntos y rayas representa un modelo aleatorio. Note que en este caso la curva corresponde a la recta identidad, puesto que es esperable que para cualquier subconjunto del conjunto de prueba haya una cantidad de personas en impago proporcional a la cantidad de personas que caen en impago del conjunto total. O sea, cuando se ha muestreado un 30% de la población, con un clasificador aleatorio, se espera haber muestreado a un 30% de aquellos que caen en impago. Cuando se muestrea un 50% de la población, se espera haber encontrado a un 50% de aquellos que caen en impago.

La curva rayada representa un modelo perfecto. Dado que se tiene a los clientes ordenados de tal forma que aparecen primero aquellos que el modelo estima que tienen mayor probabilidad

de impago, es esperable que un modelo perfecto encuentre en primer lugar a aquellos que caen en impago, por lo que el 100% en el eje vertical se alcanza rápidamente. Una vez que se han muestreado todos los clientes que caen en impago, la curva no puede subir más y continúa plana en 100%.

La curva continua muestra un ejemplo de cómo se puede ver un modelo típico. Es claro entonces que entre más se acerque la curva del modelo a la curva naranja, y entre más se aleje de la curva azul, mejor será el modelo. Las letras A y B en la figura denotan las áreas respectivas entre la curva verde y la azul, y entre la curva naranja y la verde. El coeficiente de Gini [24], se calcula entonces con la fórmula:

$$Coeficiente\ de\ Gini\ =\ \frac{A}{A+B}$$

Note que con esta fórmula se obtiene entonces un número entre 0 y 1. Entre más cercano sea el coeficiente a 1, el comportamiento del modelo será más cercano al de un modelo perfecto, mientras que, entre más cercano sea a 0, el comportamiento será más cercano a un modelo aleatorio.

Finalmente, en el contexto de entrenamiento y evaluación de modelos de clasificación, presentamos el método *Bootstrapping*, el cual nos ayuda a obtener más información acerca de los posibles errores estadísticos asociados a mediciones particulares, como la medición del coeficiente de Gini.

2.1.4 Método *Bootstrapping* para estimación de incertidumbre

El método *Bootstrapping* es una técnica estadística muy comúnmente utilizada para cuantificar la incertidumbre asociada con una medición estadística. Funciona de la siguiente manera: dado un conjunto de datos de prueba, se desea simular la obtención de muchas muestras que se comporten de manera similar a los datos de este conjunto de prueba, hacer la medición de interés en cada muestra y de ahí estimar la incertidumbre de esta medición calculando estadísticos como el promedio, desviación estándar o intervalos de confianza. Para realizar dicha simulación, se toma una cantidad n de muestras *con reemplazo* (i.e. se pueden repetir elementos) del conjunto de datos de entrenamiento. Sobre estas muestras es que se hace la medición y se calculan sus estadísticos asociados [21].

Se usó esta técnica para generar intervalos de confianza al 95% para el coeficiente de Gini y las métricas de justicia que se usaron en el estudio, los cuales se obtuvieron calculando los

Figura 2.3. Tipos de sesgo en modelos de clasificación.

percentiles 0.025 y 0.975 de las mediciones sobre las muestras de *Bootstrapping*. Es decir, el 95% de las mediciones se encontraron en el rango reportado.

Habiendo explicado los temas básicos de modelos de clasificación, entramos ahora a temas más específicos del presente caso de estudio, comenzando por las posibles fuentes de sesgo que se pueden encontrar en un modelo de este tipo.

2.2 Sesgos en modelos de clasificación

El sesgo se ha definido anteriormente como un fenómeno en el cual una persona o un modelo asigna distintas distribuciones de etiquetas a elementos de clases distintas [11]. Los tipos de sesgo que se presentan durante el uso de modelos de aprendizaje de máquina son distintos de los tipos de sesgo que presentan las personas. Durante la revisión de literatura hecha para este estudio (explicada en más detalle en el capítulo de antecedentes), se encontró algunos intentos de clasificar los tipos de sesgo a los que se enfrentan los modelos de aprendizaje de máquina. Sin embargo, ninguna clasificación fue satisfactoria, por lo que se propone aquí una nueva clasificación, la cual se muestra en la figura 2.3 y se explica a continuación. Sin embargo, antes de entrar en detalles con dicha clasificación, es necesario aclarar algunos términos.

En la introducción de este documento se habla de sesgos aceptables y sesgos discriminatorios. Cuando se trata de sesgos discriminatorios, normalmente se trabaja con datos de personas. Se conoce como **atributo protegido** o **atributo sensible** a aquel atributo de los datos que constituye la variable de interés sobre la cual históricamente se ha dado discriminación en el contexto en estudio [11]. En este caso de estudio, por ejemplo, el atributo sensible es el género de

las personas. Otros ejemplos podrían ser la edad o la etnia. Al grupo de personas históricamente discriminadas en base a este atributo se les conoce como el **grupo protegido** o **grupo no-privilegiado** [11], [19], [25].

Entonces, volviendo a la clasificación de tipos de sesgo, en un primer nivel, se puede clasificar los sesgos según su fuente. Se tienen tres categorías: sesgos en los datos de entrenamiento, sesgos introducidos por el modelo y sesgos introducidos durante el uso del modelo.

Entre los sesgos en los datos de entrenamiento, se tiene, en primer lugar, el **tratamiento dispar**. Este tipo de sesgo ocurre cuando los datos para entrenar el modelo presentan resultados que están directamente sesgados. Por ejemplo, en el caso del sesgo de género en calificaciones de crédito, si se tiene un conjunto de datos en el cual hay créditos que no se otorgaron a mujeres solo por el hecho de ser mujeres, esto es tratamiento dispar [19].

En segundo lugar, el **sesgo de asociación** se refiere a un tipo de sesgo en el cual, incluso si la variable sensible (e.g. el género) no se encuentra en el conjunto de datos, existen otras variables no-sensibles altamente correlacionadas con esta primera [12]. Un ejemplo típico en Estados Unidos es el uso del código postal de los clientes, el cual está altamente correlacionado con su raza debido a razones históricas. A este tipo de variables asociadas con una variable sensible, se les conoce con el nombre de variables ***redlining*** [19].

El **sesgo de selección** ocurre cuando un grupo de observaciones se encuentra subrepresentado en el conjunto de datos [12]. Por ejemplo, si un banco utiliza como datos de entrenamiento información solamente de los y las clientes a quienes les ha otorgado créditos históricamente, y si históricamente se le ha otorgado más créditos a hombres que a mujeres, se tiene un sesgo de selección operando en contra de las mujeres.

Un cuarto tipo de sesgo que puede estar presente en los datos de entrenamiento es el **sesgo intencional**. Mancuhan y Clifton [20] mencionan como un ejemplo de este tipo de sesgo el Acta de Igualdad de Oportunidades Crediticias de Estados Unidos, en la cual se hace una excepción explícita que permite el uso de la edad en calificaciones de crédito, a pesar de que este atributo se considera sensible.

Dentro de los sesgos introducidos por el modelo, se tiene el **uso directo de atributos sensibles** y el **sesgo intencional** [20], [25]. Estos sesgos son similares al tratamiento dispar y el sesgo intencional en el conjunto de datos; simplemente se hace la distinción de que estos tipos de

sesgo pueden ser introducidos en la etapa de entrenamiento del modelo aun con un conjunto de datos sin sesgos.

Dos tipos adicionales de sesgos introducidos por el modelo son el **sesgo de subestimación**, el cual puede ocurrir cuando se utiliza un modelo que no ha convergido por completo [26], y el **sesgo malicioso**, en el cual se lleva a cabo una manipulación intencional de un modelo para producir resultados sesgados [27].

Finalmente, durante el uso del modelo, se ha identificado el **sesgo de automatización**. Este tipo de sesgo puede ocurrir de manera positiva, cuando se favorece los resultados de un modelo de aprendizaje de máquina por encima del criterio humano, independientemente del grado de exactitud de cada método; o de manera negativa, cuando más bien se favorece el criterio humano [11].

Para poder determinar si un modelo de aprendizaje de máquina es justo o no, o inclusive para determinar si un modelo es más justo que otro, es necesario tener una definición clara de lo que es la justicia y cómo medirla. En la siguiente subsección se trata este problema.

2.3 Definiciones de justicia

Definir la justicia es un problema filosófico sin una respuesta única. Esto se ve reflejado en la literatura, pues existen diferentes definiciones de métricas que intentan definir cuándo un modelo es más justo que otro. Se presenta aquí una categorización basada en los estudios de Balayn, Lofi y Houben [11], y de Verma y Rubin [25]. Es importante recalcar que, como bien se menciona en estos mismos estudios, se ha demostrado que estas definiciones son matemáticamente incompatibles, en el sentido de que no es posible tener un modelo que sea justo bajo todas las definiciones aquí presentadas simultáneamente.

2.3.1 Clasificación de métricas de justicia

A un nivel amplio, las definiciones de justicia se clasifican en tres grandes grupos: justicia de grupo, justicia individual y justicia causal [11], [25].

La justicia de grupo define como "justo" a un modelo que trata de "igual manera" a los miembros de un grupo protegido (e.g. en el presente caso, las mujeres) respecto a los demás grupos o al total de las observaciones. Entre las métricas que se basan en esta definición de justicia, se encuentran aquellas **basadas en las etiquetas predichas**, las cuales usan solamente la distribución de etiquetas asignadas por el modelo a los distintos grupos para medir qué tan justo

es el modelo. También están las métricas **basadas en etiquetas predichas y etiquetas reales**, que, como su nombre lo indica, usan también las etiquetas "reales" provenientes del conjunto de datos de entrenamiento (este tipo de métricas asume que tales etiquetas "reales" existen en primera instancia). En tercer lugar, existen también métricas de justicia de grupo **basadas en probabilidades de predicción y etiquetas reales**, las cuales usan las probabilidades de que las etiquetas predichas sean correctas, según las asigne el modelo [11], [25].

La justicia individual pretende que un modelo justo trate de "igual manera" a cada individuo, independientemente de si este pertenece a un grupo protegido o no. La **justicia por desconocimiento** (*fairness through unawareness*) es la manera más inocente en que se pretende hacer esto, solamente haciendo que el modelo no tenga acceso directo a la información protegida de un individuo. La **justicia por contraste** (*counterfactual fairness*) se basa en la obtención del mismo resultado para dos individuos cuyos atributos son iguales exceptuando quizá sus atributos protegidos. La **justicia por conocimiento** (*fairness through awareness*) se basa en la idea de que dos individuos "similares" (bajo alguna métrica) deberían recibir etiquetas "similares" (bajo alguna métrica) [11], [25].

Las métricas de justicia causal se basan en el uso de grafos de causalidad. No para todos los modelos es posible construir un grafo de este tipo. Entre este tipo de métricas se encuentra la **discriminación causal**, la cual establece que un modelo es justo si el resultado dado por este no depende en su grafo de causalidad de un descendiente del atributo protegido [25].

Una segunda versión de justicia causal es la **falta de discriminación irresuelta**. Esta se basa en el concepto de *atributos de resolución*, los cuales son atributos en el grafo de causalidad que se ven influenciados por el atributo protegido de una manera que no se considera discriminatoria. Por ejemplo, el género de una persona puede influenciar su salario, sin embargo, no se considera discriminatorio hacer uso del salario de una persona para considerar las condiciones que se le otorgan en un crédito. La falta de discriminación irresuelta ocurre cuando no existe ningún camino en el grafo de causalidad desde el atributo protegido hasta la etiqueta predicha, excepto a través de un atributo de resolución [25].

En tercer lugar, se encuentra la **falta de discriminación por proxy**. Un *atributo proxy* es un atributo cuyo valor puede ser derivado por medio del uso de otro atributo. Por ejemplo, muchas veces es posible determinar el género de una persona a través de su profesión. Un modelo tiene una falta de discriminación por proxy si en su grafo de causalidad no existe ningún camino desde el

atributo protegido hasta la etiqueta predicha que pase por un atributo proxy del atributo protegido [25].

El cuarto y último tipo identificado de justicia causal es la **inferencia justa**. Para hacer uso de esta definición es necesario identificar todos los caminos del atributo protegido a la etiqueta predicha en el grafo de inferencia de un modelo como legítimos o ilegítimos. Un modelo justo es entonces aquel en el cual no hay ningún camino ilegítimo entre el atributo protegido y la etiqueta predicha en su grafo de causalidad [25].

Finalmente, durante la revisión de literatura se encontró una referencia sobre la **justicia en el proceso**, la cual no se ha podido ajustar en ninguna categoría de las anteriormente mencionadas. Esta definición se basa de manera un poco más abstracta en la búsqueda de un tratamiento justo durante el proceso que lleva a la predicción, tomando en cuenta y documentando atributos de entrada que son usados por el modelo [9].

En las siguientes sub-secciones se detallan algunas métricas de justicia específicas que fueron usadas en este estudio. La justificación de por qué se usan estas y no otras se puede ver con detalle en el capítulo de metodología y el de resultados.

2.3.2 Porcentaje de puntos que fallan un test situacional

Un test situacional es un tipo de prueba básica que se puede realizar sobre un modelo de clasificación. El test consiste en cambiar el atributo protegido de un individuo, sin cambiar los demás atributos y alimentar con estos datos al modelo para verificar que el resultado no cambie. Si el resultado del modelo cambia al cambiar el atributo protegido, esto muestra evidencia de la existencia de un tratamiento dispar del modelo. Este tipo de test puede ser utilizado como métrica de la equidad en el modelo midiendo la proporción de individuos que fallan el test en un conjunto de prueba [28]. Note que este es un tipo de métrica de justicia individual por contraste.

2.3.3 Porcentaje de personas en el grupo no-privilegiado a las que el modelo asigna un resultado negativo

Un siguiente paso para medir la justicia de un modelo está dada por el porcentaje de personas en la clase no-privilegiada a las que el modelo asigna un resultado negativo (medido sobre un conjunto de prueba) [29]. Este es un tipo de métrica de justicia de grupo basada en las etiquetas predichas. Por supuesto, esta es una métrica muy limitada, puesto que no considera la diferencia que hay respecto al grupo privilegiado.

Tabla 2.2. Definiciones en matriz de confusión para cálculo del EOD y AOD.

	Caen en impago	No caen en impago
Modelo predice impago	TN (verdaderos negativos)	FN (falsos negativos)
Modelo predice no-impago	FP (falsos positivos)	TP (verdaderos positivos)

2.3.4 *Equal Odds Difference* (EOD)

La métrica llamada EOD (*Equal Odds Difference*) pretende medir la diferencia entre el grupo protegido y el grupo privilegiado en la probabilidad del modelo de asignar un resultado positivo a un cliente dado que este cliente realmente hace sus pagos a tiempo [28]. Es un tipo de métrica de justicia de grupo basada en etiquetas predichas y etiquetas reales. En particular, en este caso de estudio, se mide la diferencia entre dos cantidades: la probabilidad de que los hombres con bajo riesgo de crédito sean evaluados correctamente en el modelo y la misma probabilidad para las mujeres. Intuitivamente, esta métrica evalúa al modelo como "justo" si el género no influye en la decisión del modelo de asignar un resultado positivo a un cliente; es decir, el modelo no favorece a nadie en base a su género.

Formalmente, el EOD se define de la siguiente manera sobre un conjunto de prueba:

$$EOD = TPR_P - TPR_U$$

donde TPR_P se refiere a la tasa de verdaderos positivos de la clase privilegiada y TPR_U se refiere a la tasa de verdaderos positivos de la clase no-privilegiada. La tasa de verdaderos positivos de cada grupo se define a su vez como:

$$TPR = \frac{TP}{TP+FN}$$

donde TP es la cantidad de verdaderos positivos asignados por el modelo, es decir, la cantidad de individuos que no caen en impago a los que el modelo les asigna un resultado positivo[2]. El término FN hace referencia a los falsos negativos, es decir, aquellos individuos en el conjunto de prueba a los que el modelo asigna un resultado negativo a pesar de que no caen en impago [28]. La tabla 2.2 resume estos términos.

[2] Convencionalmente, en modelos de calificación crediticia, se suele tomar como valor "positivo" el impago, por lo que la tasa de verdaderos positivos haría referencia a la probabilidad del modelo de predecir impago para un cliente que realmente impaga. Sin embargo, en las métricas EOD y AOD se suele equiparar la definición de "positivo" con los resultados que suelen considerarse positivos según el contexto. En este caso, lo positivo para el cliente es predecir no-impago, puesto que implica beneficios para este.

Una limitación importante de esta métrica es que asume que los datos de prueba son correctos, es decir, que no existen sesgos provenientes de estos. También cabe observar que, según lo reportado por [25], los autores fueron capaces de entrenar un modelo con EOD igual a 0.00% en el conjunto de datos de *German Credit* [30]. Por sí solo este parece ser un buen resultado, sin embargo, tomando en cuenta otras métricas, se observa las limitaciones del EOD, como se explica en la siguiente sub-sección.2.3.5 *Average Odds Difference* (AOD)

Una de las principales limitaciones del EOD es el hecho de que solo considera a las personas en el conjunto de prueba con verdaderos resultados positivos, es decir, en este caso solo se considera a aquellas personas que no caen en impago. Se podría decir que el EOD solo se preocupa por un trato justo para aquellas personas que no caen en impago. Podría ocurrir que el modelo obtenga un EOD de cero y sin embargo, tenga un trato que se considere injusto para las personas que sí caen en impago, asignando mejores resultados (erróneamente) al grupo privilegiado, por ejemplo.

El AOD (*Average Odds Difference*) pretende corregir esta limitación del EOD considerando también a las personas que tienen un verdadero resultado negativo. Formalmente, se define como:

$$AOD = \frac{\left(FPR_P - FPR_U\right) + \left(TPR_P - TPR_U\right)}{2}$$

donde TPR_P y TPR_U tienen el mismo significado que en el EOD y FPR_P y FPR_U representan la tasa de falsos positivos para la clase privilegiada y la clase no privilegiada, respectivamente. El FPR de cada clase se define a su vez como:

$$FPR = \frac{FP}{FP + TN}$$

donde FP es la cantidad de individuos en el conjunto de prueba para los que el modelo predice un resultado positivo a pesar de que realmente no lo tienen, y TN es la cantidad de individuos en el conjunto de prueba con un resultado negativo y para los cuales el modelo predice correctamente un resultado negativo [28]. Nuevamente, la tabla 2.2 resume estos términos. Note que, al igual que el EOD, esta es una métrica de justicia de grupo basada en etiquetas predichas y etiquetas reales.

2.3.6 Porcentaje de individuos discriminados según métrica BEL (*Bayesian Extended Lift*)

Mancuhan y Clifton [20] presentan el uso de la métrica *Bayesian Extended Lift* (BEL) para hacer una medición de la discriminación en el conjunto de datos. A diferencia de las métricas

anteriores, esta no se preocupa por los resultados de un modelo, sino solamente por aquellos resultados que se suelen tomar como "reales" en el conjunto de datos de entrenamiento.

Para definir el BEL, se dividen los atributos del conjunto de datos en los conjuntos $A = \{a_1, a_2, \ldots, a_l\}$, $B = \{b_1, b_2, \ldots, b_m\}$ y $R = \{r_1, r_2, \ldots, r_n\}$, donde A representa a los atributos protegidos, B representa a los atributos no-protegidos y R representa a los atributos con efecto *redlining*, es decir, aquellos atributos que no son protegidos pero que están correlacionados con los de A. Además, se define $C = \{-, +\}$ como las posibles clases en las que se clasifica cada punto en el conjunto de datos (en nuestro caso, si el cliente cae en impago o no). Es decir, una instancia x en el conjunto de datos es de la forma:

$$x = (x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n, c)$$

donde $x_1 \in dom(a_1)$, $x_2 \in dom(a_2)$, ..., $x_l \in dom(a_l)$, $y_1 \in dom(b_1)$, $y_2 \in dom(b_2)$, ..., $y_m \in dom(b_m)$, $z_1 \in dom(r_1)$, $z_2 \in dom(r_2)$, ..., $z_n \in dom(r_n)$, $c \in C$, con $dom(w)$ denotando el dominio de w. Entonces, el BEL se define como:

$$BEL(x) = \frac{P(c|x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n)}{P(c|y_1, y_2, \ldots, y_m)}$$

de tal forma que $P(c|x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n) > t > P(c|y_1, y_2, \ldots, y_m)$, donde t denota el límite de decisión de un modelo entre $c = -$ y $c = +$.

Note que esta definición asume la existencia de un modelo de clasificación binaria entre las clases en C, de tal forma que el BEL está definido para una instancia del conjunto de datos cuando, al tomar en cuenta todas las variables de esta, el modelo asigna una probabilidad alta de que la instancia pertenezca a una de las clases en C, mientras que cuando se toma en cuenta solo los atributos no-protegidos de la instancia, este asigna una probabilidad baja (donde el límite entre "alto" y "bajo" está dado por t). Entonces, la métrica BEL, mide, para cada instancia en la que hay un cambio en la decisión del modelo según si se usan o no los atributos protegidos y *redlining*, la proporción de cambio en la probabilidad asignada por el modelo. Sin embargo, como se mencionó anteriormente, esta métrica no mide la discriminación sobre un modelo específico, sino que construye un modelo basado en redes bayesianas para hacer esta medición sobre este modelo. Por esta razón, se puede clasificar esta métrica entre las métricas de justicia causales, específicamente como un tipo de métrica basada en la falta de discriminación por proxy.

Tabla 2.3. Algoritmo para discretizar datos continuos en preparación para obtener la métrica de Mancuhan y Clifton [19] (elaboración propia).

```
discretizar_cuantiles(datos, max_cortes)
% Toma un conjunto de datos continuos en forma de tabla y los discretiza con tres objetivos:
%     - Mantener los cuantiles tan cercanos como sea posible a los originales
%     - En cada columna no crear más cortes que los especificados en max_cortes
%     - Mantener la correlación entre variables tan cercana como sea posible a la original

para cada columna en datos:
    mejor_distancia = infinito

    para n_cortes=2..max_cortes:
        columna_discretizada <- discretizar columna en base a n_cortes cuantiles
        si no es posible lo anterior, discretizar en base a n_cortes cortes homogéneos

        % Se usa la correlación y p-values dados por la r de Pearson en ambos casos
        calcular vectores de correlaciones y p-values entre columna (continua) y las otras columnas
        calcular vectores de correlaciones y p-values entre columna (discreta) y las otras columnas

        dist <- distancia euclidiana ponderada entre vectores de correlaciones (pesos según p-values)

        si dist < mejor_distancia:
            mejor_distancia <- dist

retornar datos_discretizados
```

Para estudiar la discriminación en el conjunto de datos, entonces, se realiza lo siguiente, según proponen los autores originales de la métrica:

1. En primer lugar, se hace necesario discretizar todos los datos que se utilizan. Esto no es explícitamente mencionado en [20], puesto que los autores comienzan usando datos puramente categóricos; sin embargo, los algoritmos utilizados por los autores para la construcción de redes bayesianas asumen que todos los datos utilizados son categóricos. El algoritmo de discretización utilizado para este estudio (de elaboración propia) se describe en la tabla 2.3. A continuación, se señalan algunos aspectos notables de este algoritmo:

    a. Se intenta discretizar cada columna preferiblemente usando los cuantiles de los datos provistos en la columna. Esto se debe a que muchos de los datos con los que se trabajó en este caso de estudio eran altamente *skewed* (se usa la palabra en inglés para distinguir este aspecto del sesgo de género que se está trabajando), por lo que era preferible usar este tipo de discretización [21]. Sin embargo, existen casos con extremo *skewness* en los que aun discretizando según los cuantiles, se obtiene un solo intervalo de valores, por lo que en dichos casos se opta por realizar los cortes de discretización de manera uniforme entre el valor mínimo y el valor máximo encontrados en la columna.

b. Se agrega el parámetro `max_cortes` debido a limitaciones de poder de procesamiento. En todos los experimentos del presente estudio, se usó un valor de 5 para este parámetro.

c. Note que este algoritmo se realiza de tal forma que se mantenga la correlación entre las variables lo más cercana a la correlación original entre las variables continuas, según el estadístico r de Pearson. El algoritmo minimiza la distancia euclidiana entre el vector dado por los valores de correlación de cada columna con las demás y el vector dado por los valores de correlación de la versión discretizada de la columna con las demás. Asimismo, se agregan pesos a la función de distancia para tomar en cuenta los *p-values* arrojados por la prueba estadística, de tal forma que se le da mayor peso a los estadísticos de correlación con diferencias en los *p-values* que sean bajas. Los pesos se normalizan para obtener valores entre 0 y 1 usando la siguiente función:

$$f(p_1, p_2) = \frac{1}{1+e^{-(|p_1-p_2|)/1.5}}$$

donde p_1 y p_2 denotan los *p-values* mencionados anteriormente.

2. Usando el conjunto de datos discretizado, se entrena una red bayesiana que permite hacer inferencia sobre la variable a predecir. Esta red se construye inicialmente con la suposición de *Naive Bayes* (es decir, se asume que existe una relación de causalidad entre todas las variables predictoras y la variable a predecir) y luego se aprenden las demás relaciones causales de forma automática, así como las tablas de probabilidad (CPDs) entre cada par de variables relacionadas por la red. Los algoritmos usados para entrenar esta red bayesiana son *Hill Climb Search* (acompañado de la métrica K2) para aprender la estructura, y *Maximum Likelihood* para aprender los parámetros de la red. En el presente caso de estudio, se usó la implementación de estos algoritmos implementada en el módulo `bnlearn` de Python [31].

3. A partir de la red anterior, se eliminan los atributos protegidos (en este caso, solo el género) y los atributos *redlining*, que, según definen los autores originales para este método, corresponden con aquellos atributos que tienen una conexión causal directa en la red bayesiana, exceptuando el atributo a predecir. Esta nueva red debe pasar por el

Tabla 2.4. Algoritmo para eliminar atributos protegidos y *redlining* de la red bayesiana. Tomado de Mancuhan y Clifton [19].

```
eliminar_atributos_protegidos_redlining(red, atributos_protegidos)
% Genera una red bayesiana sin atributos protegidos ni redlining según la definición de
% Mancuhan y Clifton
%
% Entradas:
%     - red: red bayesiana cuyos nodos corresponden a atributos de los datos
%     - atributos_protegidos: lista de atributos protegidos

red_relativa <- red

para cada atributo en atributos_protegidos:
    padres <- encontrar_padres(red, atributo)
    hijos <- encontrar_hijos(red, atributo)
    red_relativa.borrar_nodos(atributo, padres, hijos)

retornar red_relativa
```

proceso de aprendizaje de estructura nuevamente para actualizar las tablas de probabilidad asociadas. Este paso se resume en la tabla 2.4.

4. Usando estas dos redes bayesianas, es posible calcular $P(c|x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n)$ (con la primera) y $P(c|y_1, y_2, \ldots, y_m)$ (con la segunda), por lo que se puede calcular el BEL para cada individuo en el conjunto de entrenamiento. Luego, la métrica final consiste en determinar el porcentaje de individuos en el conjunto de entrenamiento que tienen una métrica BEL mayor a un umbral dado[3]. Al igual que los autores, se usó en este estudio como umbral el valor 1, el cual indica que la probabilidad de asignar a un cliente un resultado negativo es mayor en cualquier medida al usar los atributos protegidos y *redlining*, que si estos atributos no se consideran. Este último paso se resume en la tabla 2.5.

En la siguiente sección de este capítulo, se hace un repaso breve de la clasificación más utilizada para las técnicas de mitigación del sesgo, seguido de una explicación detallada de las técnicas utilizadas en este estudio. Al igual que en caso de las métricas, la justificación de por qué se usan estas técnicas y no otras se detalla en las secciones de metodología y resultados.

[3] Note que la métrica BEL solo está definida para aquellos individuos para los cuales existe un cambio en la predicción del modelo al usar o no usar los atributos protegidos y *redlining*. Para efectos de calcular el porcentaje, se toma el BEL como -1 para aquellos individuos para los cuales la métrica no está definida, de tal forma que se considera que estos individuos no sufren una discriminación por parte del modelo.

Tabla 2.5. Algoritmo para descubrir instancias discriminadas. Tomado de Mancuhan y Clifton [19].

```
descubrir_discriminacion(atributos_protegidos, c, umbral, t, datos)
% Determina las instancias discriminadas y no discriminadas del conjunto de datos, así como la
% red bayesiana que describe el conjunto de datos
%
% Entradas:
%     - atributos_protegidos: lista de atributos protegidos
%     - c: valor negativo de variable predictora (e.g. impago)
%     - umbral: umbral sobre el cual se considera que una instancia es discriminada
%     - t: límite de decisión para la red bayesiana
%     - datos: conjunto de datos de entrenamiento

red <- construir_red_bayesiana(datos)
red_relativa <- eliminar_atributos_protegidos_redlining(red, atributos_protegidos)
instancias_discriminadas <- conjunto_vacio
instancias_no_discriminadas <- conjunto_vacio

para cada instancia en datos:
    prob_clase <- estimar_prob(instancia, c, red)
    prob_relativa_clase <- estimar_prob(instancia, c, red_relativa)
    belift <- calcular_belift(prob_clase, prob_relativa_clase, t)

    si belift >= umbral:
        instancias_discriminadas.agregar(instancia)
    si no:
        instancias_no_discriminadas.agregar(instancia)

retornar instancias_discriminadas, instancias_no_discriminadas, red
```

2.4 Técnicas de mitigación del sesgo

En la literatura (ver, por ejemplo, [11]) es usual clasificar las técnicas de mitigación del sesgo según su etapa de aplicación. En primer lugar, tenemos las técnicas de **pre-procesamiento**, las cuales se basan en la modificación de los datos de entrenamiento para reducir el sesgo en estos y por tanto el sesgo que un modelo de aprendizaje de máquina aprende a partir de los datos. Las técnicas de **procesamiento** se basan en modificar el proceso de aprendizaje del modelo para mitigar el sesgo, por ejemplo, mediante cambios en la función objetivo. Finalmente, las técnicas de **pos-procesamiento** son aplicadas sobre modelos ya entrenados para reducir el sesgo presente en sus resultados. Por ejemplo, existen técnicas que modifican los datos que se alimentan a un modelo entrenado para tratar de mitigar el sesgo que el modelo pueda presentar. En la práctica, sin embargo, se ha encontrado que muchas técnicas tienen componentes que pueden clasificarse en más de una de estas categorías. A continuación, se explican en detalle las técnicas de mitigación de sesgo utilizadas en este estudio, empezando por la llamada Fairway.

2.4.1 Fairway

La técnica Fairway presentada por Chakraborty, Majumder, Yu y Menzies [28] consiste en realidad en dos técnicas que podrían manejarse de forma separada: un paso de pre-procesamiento

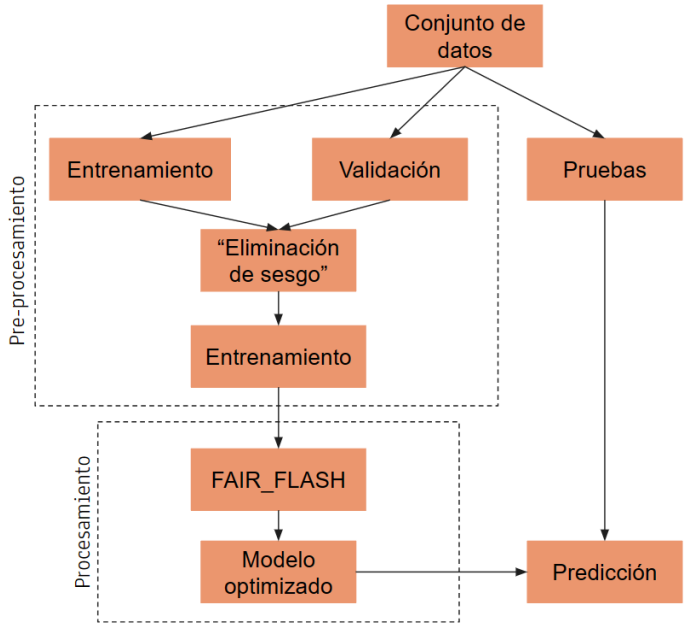

Figura 2.4. Diagrama explicativo de la metodología Fairway. Tomado de [30].

de los datos y un paso de procesamiento. El diagrama de la figura 2.4 resume los pasos a seguir en esta técnica. Note que el conjunto de datos se divide en datos de entrenamiento, validación y pruebas, como suele ser usual.

En la etapa del pre-procesamiento, los autores introducen un paso para lo que ellos llaman "eliminación del sesgo". Este paso consiste en entrenar dos modelos predictivos usando los datos de entrenamiento y validación. Un modelo se entrena solo con observaciones de la clase protegida y el otro solo con observaciones de la clase privilegiada. Luego, todas las observaciones son evaluadas por ambos modelos y se genera un nuevo conjunto de entrenamiento/validación que excluye a todas aquellas instancias en las que los dos modelos difieran en su predicción [28]. La tabla 2.6 resume los pasos anteriores.

El otro procedimiento que los autores presentan es el uso de una función multi-objetivo que sustituye a la función objetivo tradicional que se usa en el entrenamiento. Esta técnica es llamada por los autores FAIR_FLASH. Basta con definir las funciones de interés que se desea que el modelo optimice de tal forma que todas deban minimizarse (por ejemplo, se usa $1 - G$ en vez de G, donde G representa el coeficiente de Gini del modelo). Luego, se suman estas funciones y se optimiza el modelo sobre la suma de ellas. Finalmente, se evalúa el modelo optimizado usando el conjunto de datos de prueba, como es usual [28].

A continuación se presenta la segunda metodología usada en este estudio: LimeOut.

Tabla 2.6. Pre-procesamiento de técnica Fairway. Tomado de [30].

```
fairway_preprocesamiento(datos)
% "Eliminación de sesgo" para técnica Fairway
%
% Entradas:
%     - datos: conjunto de datos de entrenamiento

datos_hombres <- datos.seleccionar(hombres)
datos_mujeres <- datos.seleccionar(mujeres)

modelo_hombres <- entrenar_modelo(datos_hombres)
modelo_mujeres <- entrenar_modelo(datos_mujeres)

para cada instancia en datos:
    y_modelo_hombres <- modelo_hombres.predecir(instancia)
    y_modelo_mujeres <- modelo_mujeres.predecir(instancia)

    si y_modelo_hombres ≠ y_modelo_mujeres:
        datos.eliminar(instancia)

retornar datos
```

2.4.2 LimeOut

Bhargava, Couceiro y Napoli [9] presentan la técnica LimeOut. Esta se divide en dos etapas que se aplican ambas durante el entrenamiento del modelo. A estas etapas les llaman $LIME_{Global}$ y $ENSEMBLE_{Out}$.

2.4.2.1 $LIME_{Global}$

El objetivo principal de la primera etapa de LimeOut, llamada $LIME_{Global}$, es determinar las características más importantes para el modelo de clasificación actual. Estas características son una entrada para la siguiente etapa, donde se busca mitigar el efecto de aquellas que sean protegidas o tengan el efecto *redlining*. Los autores del método recomiendan determinar cuáles son las 10 características más importantes del modelo (o algún número manejable para un humano) [9]. En el presente caso, el modelo utilizaba solamente seis características para hacer las predicciones, por lo que esta etapa se pudo obviar y fue posible aplicar directamente $ENSEMBLE_{Out}$ con todos los atributos disponibles. Por completitud, se describe acá brevemente esta etapa y se desarrolla en más detalle en el Anexo A.

Otro aspecto importante a considerar fue que LIMEOut asume que ya existe un modelo entrenado sobre el conjunto de datos al comenzar. En el presente caso de estudio, esto no representa ningún problema, puesto que así es; sin embargo, en un caso en el que no exista un primer modelo, bastaría con entrenar un modelo primeramente sin usar ninguna técnica de mitigación de sesgos.

$LIME_{Global}$ se basa en la metodología LIME (*Local Interpretable Model-agnostic Explanations*), propuesta por Ribeiro, Singh y Guestrin en 2016 [32]. Esta es una metodología usada para generar explicaciones locales para cualquier tipo de modelo, especialmente para aquellos que suelen ser más difíciles de interpretar. LIME consiste en generar, para cualquier observación dada, un modelo simple que aproxime localmente al modelo complejo que se pretende explicar. Así, por ejemplo, se puede usar un modelo lineal para aproximar localmente el resultado de un modelo basado en redes neuronales, de tal forma que sea más intuitivo para un usuario entender qué características de los datos tienen importancia para el modelo a la hora de hacer la predicción, al menos para elementos relativamente similares al elemento en estudio.

Los autores también propusieron una metodología para seleccionar aquellas instancias en un conjunto de datos que pudiesen proveer más información a un usuario humano acerca del comportamiento del modelo a explicar, es decir, un conjunto de instancias de los datos que sean diversas y no redundantes. A esta metodología le llamaron escogencia submodular (*submodular pick*) [32].

Originalmente, el método se planteó solamente para modelos de procesamiento de lenguaje natural o de visión por computadora. En 2020, Garreau y von Luxburg [33], [34] extendieron el método LIME para incluir modelos basados en datos tabulares. Esta extensión consiste en discretizar los datos usando la distribución empírica de los mismos encontrada en un conjunto de datos de entrenamiento.

Bhargava, Couceiro y Napoli [9] utilizan esta nueva versión de LIME para generar, en la etapa $LIME_{Global}$ de LimeOut, una explicación global del modelo en uso. Esto se realiza combinando los resultados de las explicaciones locales, en última instancia generando una puntuación de cada uno de los atributos de los datos según su importancia para el modelo. Como se mencionó anteriormente, los autores luego plantean hacer una escogencia de las variables más importantes según se obtienen de acá. Estas variables serán insumo de la siguiente etapa, $ENSEMBLE_{Out}$.

2.4.2.2 $ENSEMBLE_{Out}$

$ENSEMBLE_{Out}$ toma como insumo las variables más importantes para el modelo según se identifican en la etapa anterior. De estas variables, se debe identificar aquellas que son protegidas o con efecto *redlining*. Seguidamente, se entrena un conjunto de modelos excluyendo, para cada modelo, un subconjunto de estas variables de las entradas del modelo. Los autores recomiendan, para evitar un crecimiento exponencial en la cantidad de modelos entrenados, entrenar solamente una cantidad $N + 1$ de modelos, donde N es la cantidad de atributos protegidos o *redlining*

identificados en los insumos de esta etapa. Así, los primeros N modelos serán modelos en los que no se considera una de las N variables protegidas o *redlining*, y el último modelo será uno en el que no se considera ninguna. Para obtener el modelo final, se toma como resultado el promedio de las probabilidades asignadas por cada uno de los modelos del conjunto [9].

Finalmente, se presenta la tercera técnica de mitigación de sesgo usada en este estudio.

2.4.3 Técnica de Mancuhan y Clifton basada en redes bayesianas

La tercera técnica de mitigación de sesgo que se trabajó en este caso de estudio fue la técnica presentada por Mancuhan y Clifton basada en redes bayesianas [20]. Esta técnica incluye una etapa de pre-procesamiento basada en el uso de la métrica que ellos mismos plantean y una etapa de procesamiento que no es independiente del tipo de modelo, sino que obliga al usuario de la técnica a entrenar un modelo basado en redes bayesianas. A continuación se explica cada etapa.

La etapa de pre-procesamiento utiliza los resultados obtenidos de la medición de la métrica de justicia basada en la métrica BEL que se presentó en la sección 2.3.6. Note que el penúltimo producto que se obtiene al medir el sesgo en el conjunto de datos de entrenamiento usando la técnica ahí explicada es un indicador de cuáles individuos en los datos de entrenamiento se considera que han sido discriminados y cuáles no. Usando este insumo, el paso de pre-procesamiento para la técnica de mitigación, consiste en cambiar la variable a predecir para aquellos individuos en el conjunto de entrenamiento que se consideran discriminados [20]. En este caso, esto quiere decir que, si un individuo se detecta como discriminado según la métrica BEL, entonces, como paso de pre-procesamiento para mitigar el sesgo, se cambia su indicador de impago de verdadero a falso o viceversa.

En la etapa de procesamiento, según el método descrito, se debe tomar la red bayesiana generada durante la medición de la métrica de justicia. A esta red bayesiana se le debe eliminar los nodos correspondientes a las variables protegidas y re-entrenar el modelo para re-calcular las tablas de probabilidad asociadas a cada nodo restante (aprendizaje de parámetros). No se eliminan las variables *redlining* puesto que los autores consideran que aún pueden contener información relevante para el modelo; la mitigación del sesgo en estas variables se realizó en el paso de pre-procesamiento. Finalmente, la red bayesiana obtenida sin los atributos protegidos y con las nuevas tablas de probabilidad es el modelo final que se utiliza para hacer inferencia sobre nuevas instancias [20]. La tabla 2.7 resume los pasos explicados anteriormente.

Finalmente, en la siguiente sección se definen algunas pruebas estadísticas que fueron utilizadas a lo largo del estudio.

Tabla 2.7. Procesamiento de técnica de Mancuhan y Clifton [19].

```
clasificacion_no_discriminatoria(instancias_discriminadas, instancias_no_discriminadas, red,
                                                          atributos_protegidos)
% Generación de modelo "no discriminatorio" para técnica de Mancuhan y Clifton
%
% Entradas:
%     - instancias_discriminadas: conjunto de instancias discriminadas según se obtienen del
%                                 algoritmo 6.2
%     - instancias_no_discriminadas: conjunto de instancias no discriminadas según se obtienen
%                                    del algoritmo 6.2
%     - red: red bayesiana obtenida del algoritmo 6.2
%     - atributos_protegidos: lista de atributos protegidos

red_no_discriminatoria <- red
red_no_discriminatoria.eliminar(atributos_protegidos)

instancias_discriminadas_corregidas <- cambiar_etiquetas_clase(instancias_discriminadas)
datos_corregidos <- concat(instancias_discriminadas_corregidas, instancias_no_discriminadas)

red_no_discriminatoria.aprender_parámetros(datos_corregidos)

retornar red_no_discriminatoria
```

2.5 Pruebas estadísticas

Se presentan aquí las pruebas en un orden lógico según su complejidad. Se comienza por la prueba de Chi Cuadrado, que fue usada para medir la correlación entre el género y el impago (ver sección 4.1.1). Luego, se sigue con las pruebas t de Student para dos muestras (independientes y pareadas) y la prueba H de Kruskal-Wallis. Estas pruebas miden la correlación entre una variable categórica y una continua. Las pruebas de t de Student fueron usadas en combinación con el método *Bootstrapping* para verificar si existe una diferencia estadísticamente significativa en los resultados de las mediciones de justicia y rendimiento sobre distintos modelos de clasificación (ver secciones 4.8, 5.7 y 5.8). La prueba de Kruskal-Wallis fue utilizada para verificar si había correlación entre el género y otras variables, es decir, para detectar variables *redlining* (ver sección 4.1.2). Luego, se presenta la prueba r de Pearson, la cual mide la correlación entre variables continuas y fue usada en el algoritmo de la tabla 2.3 para discretizar datos. Finalmente, presentamos el índice de asociación de Kendall ajustado por empates, el cual fue mencionado en la introducción. El resultado de la aplicación de este índice para cada cliente constituye una de las variables predictoras del modelo en estudio.

2.5.1 Prueba de Chi Cuadrado

La prueba de Chi Cuadrado para dos variables categóricas permite determinar con cierto grado de certeza la existencia de una correlación entre dichas variables. Esta prueba se basa en el uso de una tabla, llamada tabla de contingencia, en la que se contabilizan las observaciones de

Tabla 2.8. Ejemplo de tabla de contingencia para prueba de Chi Cuadrado.

	Cae en impago	No cae en impago	**Total**
Hombre	25	215	**240**
Mujer	21	245	**266**
Total	**46**	**460**	**506**

cada posible combinación de valores de la primera y la segunda variable (ver ejemplo en tabla 2.8: se contabilizan las ocurrencias de hombres y mujeres que caen o no en impago para determinar la correlación entre el género y el estado del cliente). En base a esta tabla, se calculan los valores esperados de cada entrada, asumiendo la independencia entre las variables. Si se denota por X a la primera variable (como variable aleatoria), y por Y a la segunda, y si a y b son posibles valores para cada una de estas variables aleatorias, respectivamente, entonces estos valores esperados se calculan de la siguiente manera:

$$e_a^b = P(X = a)P(Y = b)T$$

donde e_a^b denota el valor esperado de la cantidad de observaciones que toman el valor a para X y b para Y, y T denota el total de observaciones en la muestra. En este caso, las probabilidades $P(X = a)$ y $P(Y = b)$ se pueden estimar a partir de la tabla de contingencia usando las cantidades:

$$P(X = a) = \frac{total\ de\ observaciones\ en\ las\ que\ X=a}{T}$$

$$P(Y = b) = \frac{total\ de\ observaciones\ en\ las\ que\ Y=b}{T}$$

Con esto, es posible calcular la estadística de chi cuadrado, dada por:

$$\chi^2 = \sum_{a,b} \frac{\left(v_a^b - e_a^b\right)}{e_a^b}$$

donde v_a^b denota el valor realmente observado para la combinación de valores a y b (el valor en la tabla de contingencia) y la suma se realiza sobre todas las combinaciones de valores a y b que pueden tomar las variables X y Y [35].

Finalmente, usando los grados de libertad que se pueden determinar a partir de la tabla de contingencia, es posible calcular un *p-value* para la estadística obtenida, el cual indica la probabilidad de obtener los valores observados asumiendo que existe independencia entre las variables. Un *p-value* pequeño (en este estudio se usó siempre 0.05 como límite para cada *p-value* utilizado), indicaría entonces que hay muy pocas probabilidades de obtener los valores observados si hay independencia, y por tanto es más probable que las variables sean dependientes entre sí [35].

2.5.2 Prueba t de Student para muestras independientes

El objetivo de la prueba t de Student para muestras independientes es determinar si existe una diferencia significativa entre la media de dos muestras independientes. Esta asume que todas las observaciones son independientes entre sí y que las dos muestras también son independientes. Asimismo, asume que las observaciones se toman de una distribución aproximadamente normal. Estas suposiciones no suelen ser tan fuertes en casos reales [35].

Matemáticamente, el estadístico se define de la siguiente manera:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{1}{n} \cdot \frac{SS_1 + SS_2}{2n-2}}}$$

donde $\overline{X_1}$ y $\overline{X_2}$ denotan las medias de las dos muestras, SS_1 y SS_2 denotan las sumas de cuadrados de las respectivas muestras y n representa la cantidad de observaciones en cada muestra[4] [35].

Como se mencionó anteriormente, este estadístico fue usado en conjunto con la técnica de *Bootstrapping* para asegurarse de que las diferencias observadas en métricas de justicia y rendimiento entre distintos modelos fueran estadísticamente significativas. Esto se logró haciendo la respectiva medición sobre una cantidad n de muestras de *Boostrapping* para cada modelo y aplicando el estadístico sobre estas medidas. Por tanto, al igual que en el caso del estadístico Chi Cuadrado, el interés en este estudio es en el *p-value* asociado, el cual indica la probabilidad de que las muestras sean dependientes.

[4] Se asume que ambas muestras tienen la misma cantidad de observaciones. También es posible calcular el estadístico sin asumir esto, pero no fue necesario usarlo en este estudio.

Cabe anotar que, en general, fue preferible usar la versión de la prueba para muestras pareadas explicada a continuación. Sin embargo, como se explica en la sección 4.8, hubo un caso en el que fue necesario aplicar la versión para muestras independientes debido a la naturaleza del método de mitigación de sesgo utilizado.

2.5.3 Prueba t de Student para muestras relacionadas

La prueba t de Student para muestras relacionadas pretende determinar si existe una diferencia significativa en mediciones sobre muestras pareadas. Por ejemplo, en el caso de un tratamiento médico, se podría obtener un vector X con mediciones de la presión para pacientes antes de aplicarles un tratamiento y luego un vector Y con mediciones de la presión para los mismos pacientes después de aplicarles el tratamiento. La prueba pretende determinar si hay una diferencia estadísticamente significativa en las mediciones de presión antes y después del tratamiento. Note que en este caso, cada entrada de X está relacionada (o pareada) con una entrada en Y, es decir, las muestras no son independientes [35].

En este caso de estudio, se utilizó esta prueba en combinación con la técnica *Boostrapping* para determinar si hubo diferencias estadísticamente significativas en mediciones de rendimiento y justicia entre el modelo original y los modelos alternativos. Por tanto, en este caso, el vector X correspondería a una serie de mediciones de rendimiento o justicia sobre una serie de muestras de *Bootstrapping* obtenidas de un conjunto de prueba con probabilidades (o predicciones) de impago dadas por el modelo en estudio y el vector Y correspondería a las mismas mediciones sobre la misma muestra variando solamente la probabilidad (o predicción) de impago de cada observación, obteniéndola de un modelo distinto.

En este caso, el estadístico se calcula por medio de la siguiente fórmula:

$$t = \frac{\overline{D}}{\sqrt{\frac{SS_D}{n(n-1)}}}$$

donde $\overline{D}$ corresponde al valor promedio del vector $D = X - Y$, SS_D corresponde a la suma de cuadrados de los valores en D y n es la cantidad de observaciones (i.e. el tamaño de X o Y) [35]. Nuevamente, para efectos de este estudio, el interés estuvo especialmente en el *p-value* asociado.

2.5.4 Prueba H de Kruskal-Wallis

La prueba H de Kruskal-Wallis es una prueba estadística no paramétrica que se usa para determinar si existen diferencias estadísticamente significativas en la mediana de una variable cuantitativa entre distintos grupos (i.e. clasificando según otra variable categórica). En otras palabras, esta es una prueba que indica una posible correlación entre una variable categórica y una variable continua. Se usa en vez de las pruebas t de Student cuando los supuestos de estas pruebas no se cumplen (especialmente la normalidad de las muestras) [35].

Para calcular esta estadística, es necesario asignar un "rango" a cada punto en el conjunto de datos, de tal forma que el punto con el valor más bajo tiene rango 1, el siguiente valor más bajo tiene rango 2 y así sucesivamente. Si en algún momento se encuentran varios puntos con valores idénticos, se debe asignar a todos estos puntos el promedio del rango que se les asignaría si tuvieran un orden (por ejemplo, si hay 3 puntos con el tercer valor más pequeño, se les asigna un a todos un rango de 4, ya que si no fueran iguales, tendrían rangos 3, 4 y 5, y estos rangos tienen valor promedio 4). Una vez se le ha dado un rango a todos los puntos, la estadística se calcula con la siguiente fórmula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N + 1)$$

donde:

- N es la cantidad de observaciones en la muestra
- k es la cantidad de grupos (dados por la variable categórica)
- R_i es la suma de los rangos en el grupo i
- n_i es la cantidad de observaciones en el grupo i [35]

Al igual que con las pruebas anteriores, usando el valor de la estadística es posible determinar un *p-value*, el cual, como es usual en estadística, indica la probabilidad de encontrar este valor suponiendo la hipótesis nula (i.e. no hay diferencia entre las medianas de los grupos determinados por la variable categórica). Un *p-value* bajo indica una alta probabilidad de una correlación entre la variable categórica y la variable numérica [35].

2.5.5 Prueba r de Pearson

El llamado estadístico r de Pearson es comúnmente utilizado para medir la correlación lineal entre dos variables, tomando valores cercanos a 1 si las variables tienden a estar correlacionadas linealmente con una pendiente positiva, cercanos a -1 si la pendiente es negativa y cercanos a 0 si no hay una correlación lineal entre las variables. Dada una muestra de puntos $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, la fórmula de cálculo para el estadístico es la siguiente:

$$r = \frac{\Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\Sigma\left(x_i - \bar{x}\right)^2 \left(y_i - \bar{y}\right)^2}}$$

donde las sumatorias se aplican variando i entre 1 y n y $\bar{x}$ y $\bar{y}$ denotan, respectivamente, los valores promedio de la primera y la segunda variable [35].

2.5.6 Índice de asociación de Kendall (ajustado por empates)

El índice de asociación de Kendall es otra prueba estadística no paramétrica que toma valores entre -1 y 1. Este mide la asociación ordinal entre dos cantidades, es decir, el índice muestra valores más altos cuando las observaciones dadas en la muestra tienden a mantener el mismo orden si se ordenan por cualquiera de los dos atributos cuantitativos, y valores cercanos a cero cuando no hay relación entre el orden en base a una variable o la otra. Si el ordenamiento tiende a ser opuesto con un atributo respecto al otro, entonces el estadístico toma valores cercanos a -1 [35].

En este estudio no se usó de manera directa este índice, pero constituye una de las variables de entrada del modelo original, como se mencionó en la introducción, para medir la "tendencia en saldo", es decir, para medir si el saldo adeudado de un cliente ha ido en aumento o en disminución, en general, en su historial de crédito.

La definición del estadístico se realiza de la siguiente manera. Suponga que las observaciones dadas son de la forma (x, y) donde x y y son las variables cualitativas en estudio. Dadas dos observaciones $(x_1, y_1), (x_2, y_2)$, se dice que estas son *concordantes* si $x_1 > x_2$ y $y_1 > y_2$, o bien si $x_1 < x_2$ y $y_1 < y_2$. O sea, dos observaciones son *concordantes* si al ordenarlas por x o por y, el orden se mantiene. Si $x_1 = x_2$ o $y_1 = y_2$, se dice que hay un *empate* entre las observaciones. En cualquier otro caso, se dice que las observaciones son *discordantes*. Entonces, asumiendo que se cuenta con n observaciones, el estadístico τ_b se define como:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

donde:

- $n_0 = n(n - 1)/2$
- $n_1 = \sum_i t_i(t_i - 1)/2$
- $n_2 = \sum_j u_j(u_j - 1)/2$
- n_c denota a la cantidad de pares concordantes
- n_d denota a la cantidad de pares discordantes
- t_i denota la cantidad de observaciones con empates en el i-ésimo grupo de empates en x
- u_j denota la cantidad de observaciones con empates en el j-ésimo grupo de empates en y [35]

Habiendo aclarado estos términos y clasificaciones, se pasa a continuación al capítulo de antecedentes, donde se presenta el estado del arte con base en los conceptos anteriormente expuestos.

CAPÍTULO III. ANTECEDENTES

Con el fin de ilustrar el estado del arte, a continuación se presentan los resultados más importantes de una investigación bibliográfica realizada por el autor para entender y catalogar las distintas técnicas de medición y mitigación de sesgo que se han utilizado específicamente en el contexto de sesgos de género en calificaciones de crédito [36]. Las preguntas más relevantes para este trabajo acerca de los artículos estudiados fueron:

- En el caso de los artículos que solo medían el sesgo:
    1. ¿Qué tipos de sesgo son tratados?
    2. ¿Cuáles métricas de justicia se utilizan?
    3. ¿Qué criterios se usan para seleccionar las métricas utilizadas?
- En el caso de los artículos que medían y mitigaban el sesgo, además de las preguntas anteriores, se tuvo:
    5. ¿Qué tipos de técnicas se usan para mitigar el sesgo?
    6. ¿Qué criterios se usan para seleccionar las técnicas de mitigación de sesgo utilizadas?
    7. ¿Qué requerimientos tienen estas técnicas para poder ser aplicadas?

Se presentan a continuación otras revisiones de literatura que fueron de utilidad para contestar a las preguntas planteadas, seguidas de estudios primarios en los que se presentaban experimentos o casos de estudio que pretendieran medir y quizá también mitigar sesgos de género en el contexto de calificaciones de crédito.

3.1 Revisiones de literatura

Durante el estudio, se encontraron tres revisiones de literatura relacionadas al tema. En primer lugar está la investigación de Verma y Rubin [25] publicada en 2018. En su estudio, los investigadores compilaron una lista de las métricas de justicia más utilizadas en artículos publicados en conferencias y revistas de aprendizaje de máquina y justicia durante el periodo 2012-2017. También demostraron el uso de cada métrica en el conjunto de datos de Crédito Alemán [30]. El estudio de estos autores fue de alta utilidad para el presente trabajo para sondear los distintos

tipos de métricas (ya presentadas en el marco conceptual) que fueron encontradas en los estudios primarios.

En 2020, se publicó la investigación de Favaretto, De Clercq y Elger [19], la cual consistió en una revisión sistemática de literatura enfocada en el tema de la discriminación en el campo de analítica de *Big Data*. En el estudio se exploraron estudios de seis bases de datos publicados entre 2010 y 2017. Se incluyeron estudios de disciplinas variadas, por ejemplo, ciencias sociales, ciencias de la computación, leyes, bioética y filosofía. Los objetivos del estudio eran "(1) entender las causas y consecuencias de la discriminación en la minería de datos, (2) identificar barreras para una minería de datos justa y (3) explorar potenciales soluciones al problema". Este último punto incluía soluciones legales y humanas. De particular interés para el presente trabajo fue la clasificación que realizaron los autores de los tipos de sesgo que se pueden presentar en modelos de aprendizaje de máquina, y en la cual se basa la presentación presentada anteriormente en el marco conceptual.

En 2021, Balayn, Lofi y Houben [11] presentaron una revisión de metodologías para medir y reducir el sesgo en sistemas analíticos y de administración de datos (*data management*). El estudio realizó una revisión de los campos de aprendizaje de máquina, minería de datos, visión por computadora, procesamiento del lenguaje natural, sistemas de recomendación, interacción humano-computador, computación humana, ingeniería de software, *data management* y FAT (*Fairness, Accountability, Transparency*, por sus siglas en inglés). En este artículo se identifican brechas en el conocimiento que incluyen la falta de guía respecto a qué métrica elegir en un caso dado y la dificultad de aplicación de las técnicas a casos reales. Cabe recalcar que el presente trabajo pretende dar un paso hacia la resolución de estos problemas, por medio de la catalogación de métricas y técnicas según el tipo de sesgo que pretenden medir o mitigar y otros criterios que se puedan usar para seleccionarlas para su uso en un caso particular.

3.2 Estudios primarios

Durante la investigación se encontraron veinte artículos que cumplían con los criterios de aplicar una técnica de medición o mitigación del sesgo en un experimento o caso de estudio. Se incluyeron artículos que aplicaran las técnicas en el contexto de un modelo de calificación de crédito y en el caso de sesgos de género, aunque no siempre al mismo tiempo. Es decir, se incluyeron artículos en los cuales se presentaba más de una aplicación de las técnicas presentadas, entre las cuales hubiera al menos una aplicación sobre un modelo de calificación de crédito y una aplicación que midiera o mitigara sesgos de género, pero estas dos aplicaciones no son necesariamente la misma.

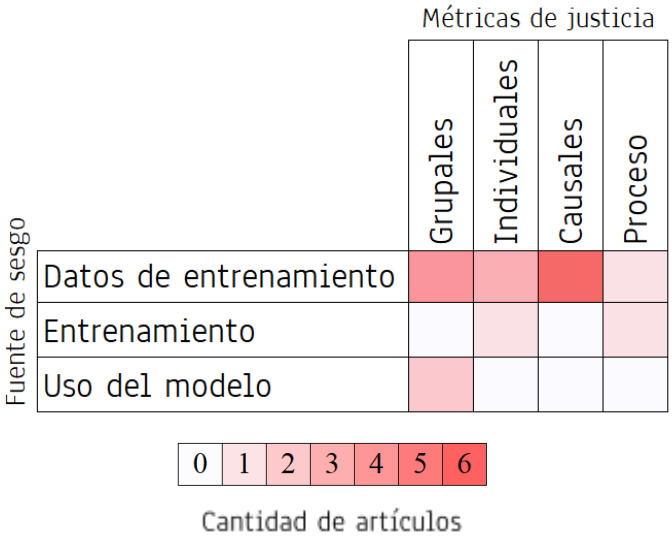

Figura 3.1. Tipos de métricas de justicia utilizadas para medir distintos tipos de sesgo.

De esos veinte artículos, diecinueve ([7], [8], [9], [10], [20], [25], [28], [29], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]) correspondieron a experimentos y uno ([47]) a un caso de estudio. Cabe destacar que no se encontró ningún caso de estudio que aplicara técnicas de mitigación de sesgo; el caso de estudio encontrado se dedicaba solo a medirlo.

Trece artículos revisados presentaban técnicas de reducción de sesgo ([7], [8], [9], [10], [20], [28], [29], [38], [39], [40], [43], [44], [45]), mientras que los otros siete solo lo medían ([25], [37], [41], [42], [46], [47], [48]). Siete artículos no trataban el sesgo de género en el mismo experimento que se trataba el caso de calificaciones de crédito, sin embargo, fue fácil observar que en todo caso las técnicas eran aplicables en un caso combinado.

En esta línea, es de especial importancia mencionar que ningún artículo hacía un tratamiento particularmente especial del sesgo de género. Todos los artículos estudiados lo presentaban solamente como un ejemplo de un tipo de sesgo. No es claro a partir de la revisión de literatura si los sesgos de género ameritan un tratamiento diferenciado de otros tipos de sesgo, como por ejemplo, sesgos por edad o sesgos étnicos.

En la figura 3.1, se observa la correlación existente en los artículos revisados entre los tipos de sesgo que explicamos en el marco conceptual y las métricas de justicia usadas para medir dichos tipos de sesgo. En primer lugar, es importante observar que la gran mayoría de artículos trabajan con sesgos en los datos de entrenamiento. Particularmente, solo se encontraron dos artículos en los que se trabajaba con sesgos introducidos durante el entrenamiento ([7], [9]) y dos que trabajaban con sesgos introducidos durante el uso del modelo ([40], [42]). En el caso de los

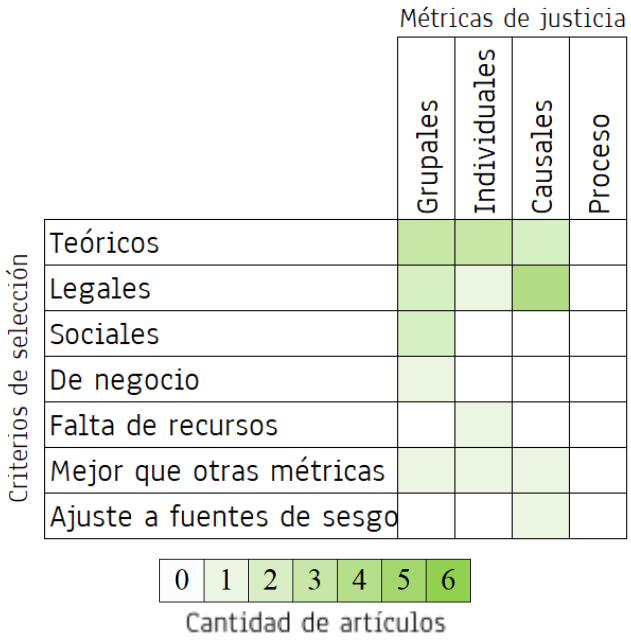

Figura 3.2. Criterios de selección usados para elegir distintos tipos de métricas de justicia.

artículos que pretendían medir el sesgo introducido durante el entrenamiento, uno de ellos [7] usaba una métrica de justicia individual, mientras que el otro [9] usaba una métrica de justicia en el proceso. Por otra parte, los dos artículos que medían los sesgos introducidos durante el uso del modelo usaban métricas de justicia grupal [40], [42].

El resto de artículos encontrados ([8], [10], [20], [25], [28], [29], [37], [38], [39], [41], [43], [44], [45], [46]) pretendía medir el sesgo introducido solamente por los datos de entrenamiento. Para ello, estos artículos usaban amplia variedad de métricas, entre las cuales destacan las causales por su cantidad. Se observó, particularmente, que muchos de estos artículos no consideraban la posibilidad de que los sesgos pudiesen provenir de otras fuentes.

Se puede observar también en la figura 3.1 que hay algunas casillas con cero artículos, los cuales representan vacíos en el estado del arte. Como se menciona más adelante en las conclusiones, valdría la pena realizar investigaciones para llenar estos vacíos.

En la figura 3.2 se observan otros criterios que los autores de cada artículo mencionan para seleccionar cada métrica. Se han clasificado aquí, para simplificar, los criterios en las siguientes categorías:

- Teóricos: Incluye criterios que tienen un interés más académico que práctico. Por ejemplo, en [25] se hace uso de varias métricas sobre el conjunto de datos de Crédito Alemán [30] con el fin de compararlas entre sí.

- Legales: Criterios que pretenden cumplir con alguna ley específica. Por ejemplo, en [38] se usan métricas para cumplir con lo estipulado en el Acta de Igualdad de Oportunidades en el Empleo de los Estados Unidos.

- Sociales: Criterios que pretenden cumplir con algún mandato social, sin referirse a ninguna ley en particular. Por ejemplo, en [29] se usan métricas que pretenden cumplir con un mandato social de "inclusión financiera".

- De negocio: Criterios que pretenden cumplir con una necesidad del negocio. Por ejemplo, en [29] se usan métricas que, aparte de medir la justicia, también pretenden medir indirectamente la reducción en el riesgo financiero de la institución prestamista.

- Falta de recursos: Incluye artículos que citan la falta de algún recurso como criterio para la selección de la métrica utilizada. Por ejemplo, en [45] se explica que hacía falta información para poder hacer uso de métricas de grupo, por lo que se opta por el uso de métricas individuales.

- Mejor que otras métricas: Algunos artículos presentan métricas cuyo fin es mejorar vacíos que han tenido métricas anteriores. Por ejemplo, los autores en [38] explican que la métrica EOD previene la aparición de llamados "individuos *token*"[5], a diferencia de la métrica llamada Paridad Demográfica.

- Ajuste a fuentes de sesgo: En el artículo [20], se cita explícitamente que la métrica usada fue seleccionada específicamente porque permite medir sesgos de tratamiento dispar y sesgos de asociación de manera simultánea.

Entonces, en la figura 3.2 se observa que los criterios presentados por los autores en los artículos revisados para escoger cada métrica son muy variados. Todos los cuadrantes presentan entre cero y cuatro artículos, con solo dos de ellos presentando tres artículos y solo uno presentando cuatro. El cuadrante que presenta cuatro artículos es el que muestra la correlación entre el uso de métricas de justicia causales (que, como se muestra en la figura 3.1, han sido las

[5] Un "individuo *token*" es un individuo al que se considera que se le otorga un buen resultado a pesar de "no merecerlo" solamente para mejorar el resultado de una métrica de justicia [38].

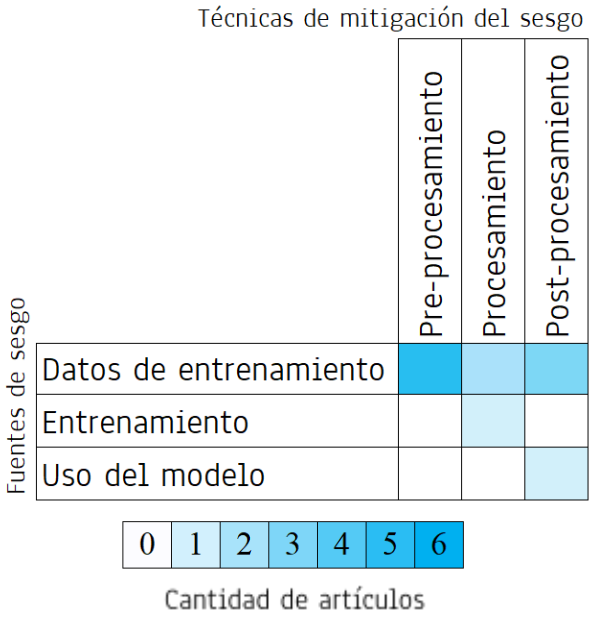

Figura 3.3. Tipos de técnicas de mitigación de sesgo utilizadas según el tipo de sesgo encontrado.

más usadas) y criterios de selección legales. Esto podría deberse a que los criterios legales suelen ser los que se encuentran mejor definidos y por tanto suelen ser más fáciles de implementar.

Una observación interesante es que los autores que usaron métricas de justicia en el proceso no citaron ningún criterio de selección particular para este tipo de métricas.

El principal aporte de las figuras 3.1 y 3.2 es facilitar la selección de una métrica de justicia a partir del tipo de sesgo que se vaya a tratar y algunas otras consideraciones que se puedan tener. Por ejemplo, se observa que el único tipo de métrica que ha sido utilizado anteriormente para tratar el sesgo en el uso del modelo son las métricas de justicia de grupo. Si, por otra parte, se quisiera tratar un caso en el que se tienen sesgos en el conjunto de datos provenientes del conjunto de datos, y si además se tiene alguna limitación de recursos, como falta de acceso al atributo protegido, se podría escoger el uso de una métrica de justicia individual.

Cabe mencionar también que no se encontró ningún artículo durante la revisión de literatura que tratara el sesgo de subestimación o el sesgo malicioso.

Unas figuras similares (figura 3.3 y figura 3.4) muestran la relación entre los tipos de sesgo, las técnicas para reducción de este y los criterios de selección de dichas técnicas.

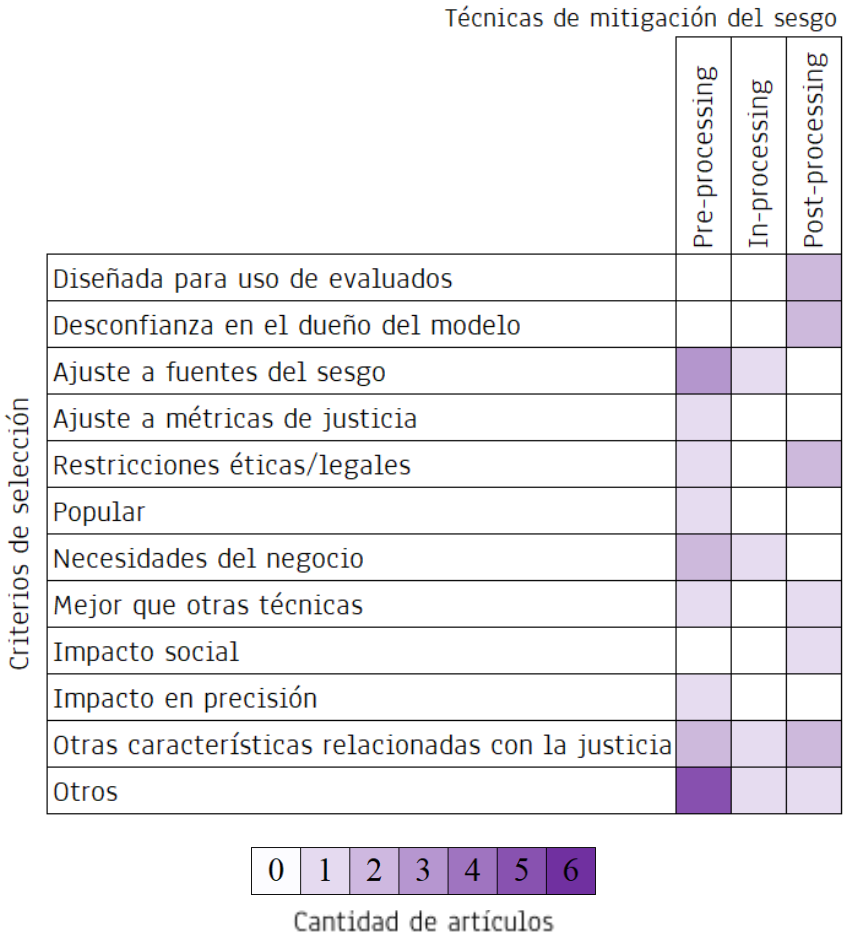

Figura 3.4. Criterios de selección para elegir la técnica de mitigación de sesgo utilizada.

En la figura 3.3 es interesante observar la diagonal principal, donde se nota que es usual mitigar los sesgos en los datos de entrenamiento usando técnicas de pre-procesamiento, los sesgos en el modelo usando técnicas de procesamiento y los sesgos en el uso del modelo usando técnicas de pos-procesamiento. Esto podría ir de acuerdo con una intuición general.

Sin embargo, también se observa el uso de técnicas de pre-procesamiento y pos-procesamiento para mitigar el sesgo en los datos de entrenamiento. Parece ser razonable también hacer uso de técnicas de mitigación en etapas posteriores a la etapa del proceso de entrenamiento en la cual se introduce un sesgo. Tomando en cuenta esto, y según la figura, podría haber un vacío en la investigación en el uso de técnicas de mitigación aplicadas durante el pos-procesamiento para mitigar sesgos introducidos durante el entrenamiento.

De manera similar a la figura 3.2, en la figura 3.4 se han agrupado otros criterios de selección que pueden ser usados para escoger una técnica de mitigación de sesgo. Se explican a continuación:

- Diseñada para uso de evaluados: Existen algunas técnicas de mitigación que han sido diseñadas para el uso de las personas evaluadas. Por ejemplo, los autores de [7] proponen que las personas evaluadas por un modelo (por ejemplo, un modelo de clasificación de crédito) cambien algunos atributos presentados al modelo con el fin de explotar debilidades del mismo para obtener un buen resultado (por ejemplo, la aprobación del crédito). La decisión de cuáles características cambiar y cómo se basa en el uso de técnicas de ataque observadas en ciberseguridad.

- Desconfianza en el dueño del modelo: Algunos artículos citan no confiar en la buena voluntad del dueño del modelo como un criterio para seleccionar la técnica de mitigación de sesgo. Este es el caso de los artículos [7], [40].

- Ajuste a fuentes del sesgo: Algunos artículos citan explícitamente el tipo de sesgo que se pretende mitigar como una razón para utilizar la técnica de mitigación presentada. Por ejemplo, en [44], los autores explican que seleccionaron la técnica utilizada debido a que esta pretende mitigar sesgos en los datos de entrenamiento y el fin de la investigación es obtener conjuntos de datos libres de sesgos discriminatorios.

- Ajuste a métricas de justicia: En [43], los autores mencionan explícitamente que se ha escogido el uso de la técnica que presentan debido a que esta mejora la métrica de justicia de grupo utilizada.

- Restricciones éticas/legales: Aquí se incluyen criterios que citan restricciones dadas por la legalidad o eticidad de lo que se puede hacer para mitigar el sesgo. Por ejemplo, en [7] se especifica que los cambios a las características usadas por el modelo, en los cuales se basa la técnica de mitigación de sesgo presentada, deben ser legales.

- Popular: Algunos estudios, como [43], citan la popularidad de la técnica como una razón de escogencia.

- Necesidades del negocio: Criterios relacionados con necesidades específicas del caso de negocio entre manos. Por ejemplo, en [20], se explica que la técnica fue escogida debido a su habilidad para manejar más de un atributo protegido.

- Mejor que otras técnicas: De manera similar al caso de las métricas, hay autores que presentan técnicas de mitigación de sesgo que pretenden mejorar defectos de otras técnicas anteriores. Por ejemplo, en [10], los autores realizan una comparación de su método con métodos anteriores para mostrar la mejoría de este primero.

- Impacto social: En [40] se menciona que la técnica presentada se ha escogido con el fin de "mitigar otros daños [sociales] que no pueden ser capturados por el concepto de 'justicia'".

- Impacto en precisión: Algunos artículos mencionan explícitamente el bajo impacto que tiene la técnica en la precisión del modelo como un criterio para seleccionarla.

- Otras características relacionadas con la justicia: Aquí se incluyen artículos que citan otras características distintas a la justicia que la técnica pretende mejorar, como la explicabilidad [9], interpretabilidad [10], calidad de predicción [10], generalizabilidad [10], privacidad [7], [8], monotonicidad [39] y utilidades del prestamista [39].

- Otros: Aquí se incluyen otros criterios que no se agruparon en las categorías anteriores. Por ejemplo, en [43], se menciona que la técnica utilizada no requiere de hiperparámetros y que esta es una de las razones por las que se escoge para su uso.

La distribución de los artículos en la figura 3.4 es muy variada. La mayoría de cuadrantes no sobrepasa los dos artículos, siendo la única excepción el cuadrante de abajo a la izquierda, el cual denota la categoría de "otros" en los criterios de selección. Se podría concluir entonces que los criterios de selección de técnicas de mitigación muestran pocos patrones y que dependen fuertemente del caso de estudio particular.

De manera similar a las figuras 3.1 y 3.2, las figuras 3.3 y 3.4 permiten seleccionar un tipo de técnica de mitigación del sesgo a partir de los tipos de sesgo presentes y otras características del problema en estudio. Es interesante notar que no se encontraron artículos que mitigaran el sesgo para todos los sub-tipos de sesgo identificados.

El estudio aquí presentado se basó en la aplicación de las técnicas presentadas anteriormente a un caso de estudio, como se detalla en la siguiente sección. Es por esto que son de vital importancia los resultados de las figuras anteriores. En el siguiente capítulo se explica la metodología que se siguió en este trabajo para alcanzar los objetivos propuestos.

CAPÍTULO IV. METODOLOGÍA

Como se menciona en la introducción, para este trabajo se contó con acceso a un modelo de calificación de crédito (específicamente un modelo de calificación de cliente) de un banco comercial de Costa Rica, así como acceso al conjunto de datos (anonimizados) usado para entrenar dicho modelo. El principal objetivo fue mitigar los sesgos de género presentes en este modelo. Esto se logró con base en una metodología soportada por un enfoque pragmatista.

Los datos de entrenamiento y pruebas del modelo están constituidos por información relacionada a 146.181 clientes del banco con al menos una operación de crédito activa entre enero de 2017 y diciembre de 2021[6]. Para cada uno de estos clientes se contó con los siguientes atributos, donde "actual" hace referencia a diciembre de 2020:

- Atraso actual: máxima cantidad de días de atraso en el pago mínimo de las operaciones de crédito del cliente al mes actual. Este es un valor entero entre 0 y 90 (se excluyen los clientes con un valor mayor a 90 puesto que en dicho caso al banco no le interesa conocer si entrarán en impago en un futuro; el cliente *ya* se encuentra en impago). El 96.56% de las personas en el conjunto de datos tenían exactamente 0 días de atraso.

- Edad en crédito: cantidad de meses (de los últimos 48) en los que el cliente tiene al menos una operación de crédito activa. Se excluyen aquellos clientes con un valor menor a 6 en esta variable, por lo que esta se encuentra entre 6 y 48. Un 65.67% de las personas en el conjunto de datos tienen exactamente una edad de 48 meses en crédito.

- Saldo actual: logaritmo de la suma de saldos pendientes de pago del cliente sobre todas sus operaciones de crédito vigentes. Se usa escala logarítmica para mitigar problemas de escala en la variable.

- Tendencia en saldo: índice de asociación de Kendall ajustado por empates entre el saldo actual del cliente y el mes en el histórico, para los últimos 48 meses. Este es un indicador estadístico con valores entre -1 y 1 que señala si el saldo en general ha tendido a subir o bajar durante la historia del cliente.

- Tiempo desde el último atraso: cantidad de meses que han transcurrido desde la última vez que la persona tuvo un atraso en el pago de una operación. Es un valor entero entre 0 y

[6] De estos datos, 12 entradas fueron eliminadas debido a que se encontraron inconsistencias en la base de datos en el género asignado a los clientes respectivos.

48, tomando el valor de 48 si la persona no ha tenido atrasos en la historia usada. El 41.45% de los clientes en el conjunto de datos toma un valor de 48 para esta variable.

- Atraso promedio: promedio de días de atraso en los que ha incurrido el cliente en sus pagos mensuales en los últimos 48 meses, ponderado con una curva logística para dar mayor énfasis a meses más recientes. Un 59.03% de los clientes toma un valor de 0 para esta variable.

- Género: género del cliente (masculino o femenino). Un 38.90% de los clientes disponibles en el histórico son mujeres y el resto son hombres[7].

- Estado: estado del cliente 12 meses después de la observación (i.e. a diciembre 2021). Es decir, si se considera que el cliente ha entrado en un impago o no. La definición de impago utilizada se basa en la definición de "crédito dudoso" dada por la Guía Metodológica para pruebas BUST (*Bottom-Up Stress Testing*) de la SUGEF (Superintendencia General de Entidades Financieras) [49], la cual indica que se debe considerar como un "crédito dudoso" aquel con un atraso de más de 90 días o que se encuentre en cobro judicial. Si un cliente tiene al menos un crédito en estado dudoso, se considera que ha entrado en impago. Apenas un 4.22% de los clientes en el conjunto de datos tienen un estado de impago.

- Entrenamiento: un indicador de si el dato fue usado para entrenamiento o pruebas del modelo. Un 25% del conjunto de datos original fue usado para pruebas.

El modelo de clasificación actual que fue entrenado con estos datos es una regresión logística que pretende determinar la probabilidad de impago de un cliente en base a sus demás variables asociadas[8]. Todos los detalles anteriormente explicados se especifican en el reporte generado para la última recalibración del modelo en 2022 [50].

La figura 4.1 muestra los pasos que se siguieron, así como las dependencias entre los mismos. Se numeran los pasos solamente para facilidad de referencia; sin embargo, el orden en que debían efectuarse estuvo dado solamente por sus dependencias. También se muestran los

[7] El conjunto de datos no provee información de personas con géneros distintos al binario tradicional hombre-mujer.
[8] Originalmente, el modelo fue entrenado usando información tanto de clientes personales como de clientes corporativos (i.e. personas jurídicas). El estudio del sesgo para el caso de personas jurídicas se sale del alcance de este trabajo, por lo que se eliminaron del conjunto de datos. Se observó que el coeficiente de Gini no variaba mucho al usar solo los datos de personas físicas (había una diferencia de alrededor de 1×10^{-5}).

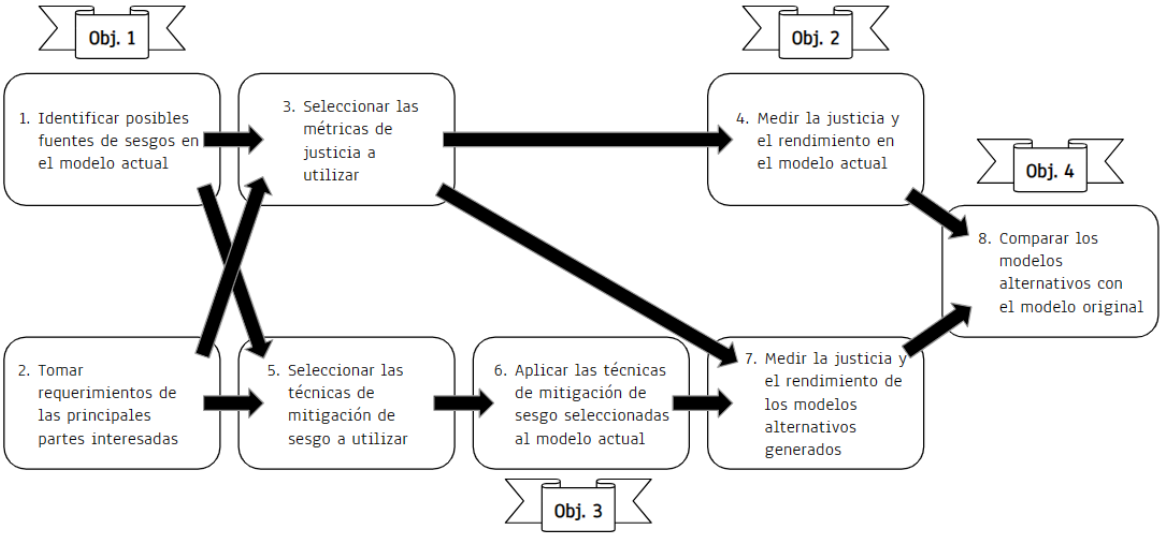

Figura 4.1. Diagrama de flujo que ilustra la metodología seguida.

puntos en los que se cumple con cada objetivo específico. A continuación se explica en detalle cada paso.

4.1 Identificar posibles fuentes de sesgo en el modelo actual

En el paso 1, se identificaron las posibles fuentes de sesgo en el modelo actual. Esto ayudó en pasos posteriores del desarrollo. Para identificar estas fuentes, se usó de manera directa las definiciones presentadas en el marco conceptual. A continuación, se explica en detalle la manera de descartar cada posible fuente de sesgo.

4.1.1 Sesgos provenientes de los datos de entrenamiento

Los sesgos provenientes de los datos de entrenamiento, como se explicó anteriormente, aparecen en los datos de entrenamiento del modelo, por lo que los análisis para descartar sesgos provenientes de esta fuente se realizaron sobre este conjunto de datos.

En primer lugar está el tratamiento dispar. Según lo expuesto en el marco conceptual, este ocurre cuando hay discriminación directa. Si bien a nivel del conjunto de datos no es posible determinar con certeza si hubo un tratamiento dispar para los clientes, sí es posible descartar esta posibilidad con alta probabilidad si se determina que no existe una correlación entre el atributo "género" y el atributo "estado". En términos estadísticos, no es posible determinar con 100% de seguridad si tal correlación existe o no, sin embargo, es posible determinarlo con un cierto grado de certeza. Para hacer esta determinación, se aplicó la prueba de Chi Cuadrado entre la variable

género y la variable estado, la cual da una medida de correlación entre estas dos variables. En base al *p-value* arrojado por esta prueba se determina entonces si hay una alta correlación entre las variables o no.

En segundo lugar, está el sesgo de asociación, el cual, según se explicó anteriormente, ocurre cuando existe una correlación entre alguna de las variables independientes que usa el modelo y el atributo protegido (en este caso el género). Para determinar si es posible descartar tal correlación, se aplicó la prueba H de Kruskal-Wallis entre cada una de las variables predictoras del modelo (continuas) y la variable género (discreta). El *p-value* arrojado por esta prueba indica si existe una correlación entre el género y cada una de las variables predictoras del modelo de regresión logística actual.

Luego está el sesgo de selección. Según lo que se explicó en el marco conceptual, por definición, este tipo de sesgo ocurre cuando se seleccionan los individuos de la muestra que constituye los datos de entrenamiento de tal forma que estos no son representativos de la población general. En este caso, se logró verificar si existía un sesgo que hacía que ciertas personas aparezcan en el conjunto de entrenamiento con mayor probabilidad que otras a través de las entrevistas que se explican en mayor detalle en la sección 4.2.

Finalmente está el sesgo intencional, el cual, según Mancuhan y Clifton [20], ocurre cuando se utilizan atributos que se consideran discriminatorios de manera intencional, por lo que bastó con analizar si se usaron este tipo de atributos.

4.1.2 Sesgos provenientes del entrenamiento del modelo

Los sesgos que se introducen durante el entrenamiento del modelo dependen del proceso de entrenamiento seguido, así como el tipo de modelo utilizado. En esta subsección se explica lo que se realizó para tratar de descartar la presencia de este tipo de sesgos en el modelo actual de calificación de cliente en estudio.

Para el caso del uso directo de atributos sensibles, bastó con observar si el género, único atributo sensible en este estudio, fue usado o no para entrenar el modelo.

Para determinar si había un sesgo intencional en el entrenamiento del modelo, de manera similar al caso del sesgo intencional en los sesgos provenientes de los datos de entrenamiento, bastó con analizar cualitativamente si se hizo uso de alguna metodología o se tomó alguna decisión durante el entrenamiento con el fin de favorecer a un grupo sobre otro injustamente [50].

En tercer lugar, está el sesgo de subestimación. Como se explicó anteriormente, este puede ocurrir cuando no se le permite al proceso de entrenamiento de un modelo converger por completo. Es difícil descartar la presencia de este tipo de sesgo debido a que no es posible saber con completa certeza si un modelo converge a un mínimo absoluto, por lo que no se tomó ninguna acción para descartar este tipo de sesgo.

Por último, está el sesgo malicioso. Tampoco es posible descartar la presencia de este tipo de sesgo, debido a que, por su naturaleza, se esperaría que, si existe una manipulación maliciosa del modelo, el responsable de esta mantenga dicha manipulación oculta. Por tanto, tampoco se realizó ningún tipo de acción para descartar la presencia de este tipo de sesgo.

4.1.3 Sesgos provenientes del uso del modelo

Para investigar este tipo de sesgos hace falta observar el proceso de toma de decisiones que se sigue posteriormente al despliegue del modelo. El sesgo de automatización, que es el que puede suceder durante el uso del modelo, ocurre cuando las personas usuarias del modelo le dan a los resultados de este un uso que no es acorde con el rendimiento medido en el modelo. En este caso, se logró obtener información suficiente del proceso de toma de decisiones a partir de la entrevista con la Directora de Riesgos de Crédito de la institución para poder determinar si existía o no un posible sesgo de automatización (ver sección 4.2).

4.2 Tomar requerimientos de las principales partes interesadas

El paso 2 consistió en una recolección de requerimientos, con el fin de determinar las métricas de justicia a usar durante el trabajo y los métodos para reducir el sesgo. Estos requerimientos coincidieron con los criterios de selección presentados en los antecedentes. Es decir, el objetivo de este paso fue determinar los criterios de selección presentados en dichas figuras que aplicaban al caso de estudio. Para poder alcanzar dicho objetivo, se realizaron entrevistas semi-estructuradas a las principales partes interesadas en el banco, a saber, la persona directora de la oficina de riesgo y a la oficina jurídica de la institución. La persona directora de la oficina de riesgo se considera de principal relevancia debido a que los colaboradores de esta oficina son los principales usuarios y usuarias del modelo en estudio. La oficina jurídica del banco en estudio se consideró importante de entrevistar para refinar aspectos jurídicos que pudiesen influir en la consideración que pudiere hacer el negocio de un modelo exitoso. A continuación se explica en más detalle el tipo de preguntas realizadas durante dichas entrevistas.

4.2.1 Identificar criterios bajo los cuales un modelo se considera sesgado

Uno de los principales objetivos de las entrevistas realizadas fue identificar criterios bajo los cuales el negocio consideraba que un modelo de calificación de crédito estaba sesgado, con el fin de identificar posibles métricas que se pudiesen usar para medir el grado en el que estos criterios estaban presentes en el modelo actual y en los modelos alternativos que se desarrollaron. En particular, para ese objetivo, se contó con 10 preguntas, que se muestran a continuación, junto con un nombre corto para identificarlas. Estas preguntas fueron planteadas en base a los criterios de selección de las métricas explicados en los antecedentes, de tal manera que la respuesta a cada pregunta permitiese al investigador descartar una o más métricas de justicia de las encontradas durante la investigación bibliográfica explicada en el capítulo de antecedentes.

1. **Conocer fuente de sesgo**: ¿Es importante para el negocio conocer la fuente de sesgo de género (datos/modelo/uso del modelo)?

2. **Mitigar sesgos de origen externo a la institución**: ¿Quisiera la institución mitigar sesgos que provengan de un origen externo a la institución (e.g. instituciones sociales)?

3. **Hacer análisis estático del modelo**: ¿Se requiere realizar un análisis estático del modelo, es decir, se requiere poder analizar si existe sesgo en el modelo sin hacer uso de un conjunto de datos de prueba para hacer dicha evaluación?

4. **Mitigar aparición de individuos *token***: ¿Es importante para el banco mitigar la aparición de "individuos *token*" al aplicar una técnica de mitigación del sesgo? Un "individuo *token*" es un individuo al que se considera que se le otorga una buena calificación de crédito a pesar de "no merecerla" solamente para mejorar el resultado de una métrica de justicia [38].

5. **Necesidad de dar explicaciones generales del modelo**: ¿Es necesario para la institución dar explicaciones generales del funcionamiento del modelo? Es decir, independientemente de que se requiera dar una explicación del resultado de un cliente particular, ¿existe alguna persona o entidad que requiera tener una explicación de los resultados generales que se obtienen del modelo?

6. **El modelo debe pasar un test situacional**: ¿Es necesario que el modelo pase un test situacional para ser considerado justo? Es decir, dado un resultado para un cliente particular, si se cambia su género, ¿es necesario que el modelo arroje el mismo resultado siempre?

7. **Confidencialidad**: ¿Es necesario que el modelo desconozca el género de cada persona a la que evalúa?

8. **Reducción del riesgo**: ¿Es importante para el banco mantener buenos resultados en alguna métrica de rendimiento del modelo que se traduzca en la reducción del riesgo de crédito?

9. **Criterios legales**: ¿Existen leyes o normativas a las que la institución se adhiera que le fuercen a utilizar definiciones particulares de métricas, umbrales o características o grupos protegidos?

10. **Mitigación de efecto *redlining***: ¿Es importante para el banco mitigar el efecto *redlining*?

4.2.2 Identificar criterios bajo los cuales una técnica de mitigación de sesgo es adecuada

El segundo objetivo de las entrevistas fue identificar criterios bajo los cuales la institución consideraba que una técnica de mitigación de sesgo era adecuada. Esto con el fin último de seleccionar las técnicas de mitigación a utilizar. Se muestran a continuación las preguntas relevantes para este objetivo junto con un nombre corto para su identificación. De igual manera que con las preguntas anteriores, estas se formularon en base a los criterios de selección de las técnicas de mitigación presentados en el capítulo de antecedentes, con el fin de descartar posibles técnicas de mitigación de sesgo a utilizar en este trabajo.

11. **Permite hacer recomendaciones**: ¿Es necesario que el banco pueda hacer recomendaciones a sus clientes de acciones legales que puedan realizar para mejorar su calificación de crédito?

12. **Interpretabilidad**: ¿Es deseable para el banco poder dar explicaciones de resultados particulares arrojados por el modelo de calificación de comportamiento?

13. **Privacidad**: ¿Es deseable para el banco que el modelo desconozca el género de las personas que evalúa?

14. **Generalizabilidad**: ¿Es necesario que la técnica de mitigación se mantenga si hay cambios en el tipo de modelo utilizado?

15. **Monotonicidad**: ¿Es importante para la institución que, si un cliente A tiene una mejor calificación que un cliente B en el modelo original, entonces el cliente A mantenga una

mejor calificación que el cliente B en un modelo alternativo que mitigue el sesgo de género?

16. **Bien conocida**: ¿Es deseable para el banco usar técnicas de mitigación bien conocidas?
17. **Dependencia de hiperparámetros**: ¿Sería aceptable para el banco tener que calibrar uno o más hiperparámetros de una técnica de mitigación del sesgo seleccionada?
18. **Exactitud**: ¿Debe un modelo de calificación de cliente mantener una buena exactitud siempre?
19. **Mitigación de sesgos en datos**: ¿Es deseable para el negocio mitigar los sesgos no solo en el modelo, sino también en los datos de entrenamiento?

Una vez completadas las entrevistas, se procedió a analizar las respuestas de manera sistemática. En este análisis, se identificaron patrones recurrentes y temas comunes en las respuestas de las personas entrevistadas. Con base en estos hallazgos, se elaboró una interpretación de las respuestas en términos de sí/no, lo que permitió categorizar y resumir los resultados de manera más concisa y fácilmente comprensible.

Este enfoque de sí/no en la interpretación de los resultados surgió como una estrategia efectiva para condensar la información cualitativa obtenida de las entrevistas en un formato más estructurado y accesible. Además, facilitó la identificación de tendencias y áreas de consenso entre los participantes, proporcionando así una base sólida para el análisis y la discusión de los hallazgos en el contexto más amplio de la investigación.

4.3 Seleccionar métricas de justicia a utilizar

En el tercer paso de la metodología, se eligió un conjunto de métricas que se usaron para medir qué tan justo era el modelo actual y las alternativas que se exploraron. Este conjunto de métricas, como se explica en la sección de antecedentes, se podían elegir a partir del tipo de sesgo que se quería tratar (según los tipos de sesgo que fueron encontrados en el paso anterior) y de las necesidades específicas del caso de estudio (identificadas a partir de las entrevistas de la sección 4.2).

4.4 Medir la justicia y el rendimiento en el modelo actual

Una vez seleccionadas las métricas de justicia que se iban a utilizar, se procedió a hacer las mediciones correspondientes sobre el modelo actual. También, se midió el rendimiento actual del

modelo usando el coeficiente de Gini, que es la métrica que se utiliza actualmente para seleccionar el modelo que se pone en producción. Para obtener robustez sobre las mediciones, se obtuvieron no solo las mediciones aplicadas directamente sobre el conjunto de pruebas o entrenamiento, sino que se utilizó la técnica *Bootstrapping* para generar intervalos de confianza al 95% para cada una de estas medidas. Se usaron 1000 muestras para cada métrica (excepto para la métrica propuesta en [20] basada en redes bayesianas, para la cual se usaron solo 100 muestras, debido a limitaciones de poder de procesamiento).

4.5 Seleccionar las técnicas de mitigación de sesgo a utilizar

Seguidamente, se escogió un conjunto de técnicas para mitigación del sesgo que fueron consideradas viables para este caso particular; a saber: Fairway, presentada por Chakraborty, Majumder, Yu y Menzies en 2020 [28]; LimeOut, presentada por Bhargava, Couceiro y Napoli en 2020 [9]; y la técnica basada en redes bayesianas presentada por Mancuhan y Clifton en 2014 [20]. Nuevamente, como se explica en los antecedentes (figura 3.3 y figura 3.4), fue posible escoger estas técnicas a partir de los tipos de sesgo que se deseaba tratar y otras necesidades específicas del caso de estudio. En este caso fue importante tomar en cuenta aspectos como el ajuste de las técnicas a los tipos de sesgo identificados, el impacto sobre la precisión del modelo, las necesidades del negocio y las restricciones legales o éticas. De igual forma que en el caso de las métricas de justicia, las necesidades de negocio y restricciones legales o éticas se obtuvieron a través de las entrevistas del paso 2. Además, se quería minimizar el impacto sobre la precisión del modelo.

4.6 Aplicar las técnicas de mitigación de sesgo seleccionadas al modelo actual

Después de seleccionar las técnicas de mitigación, diseñadas para abordar los sesgos identificados en el modelo de regresión logística actual, y para cumplir con los demás criterios identificados durante las entrevistas de la sección 4.2, se procedió a aplicarlas. En cada instancia, la aplicación de las técnicas de mitigación dio lugar a un nuevo modelo. Estos constituyeron versiones nuevas del modelo de calificación de crédito que incorporaban las técnicas de mitigación seleccionadas.

Cada modelo alternativo fue construido con el objetivo de abordar y resolver los problemas de sesgo identificados en el modelo original. A través de este proceso, se generó una serie de alternativas, cada uno de ellos con mejoras con respecto al original en términos de la mitigación de los problemas identificados.

4.7 Medir la justicia y el rendimiento de los modelos alternativos generados

Una vez generados los nuevos modelos, el siguiente paso en la metodología fue aplicar las métricas de justicia y rendimiento. Estas métricas, que fueron definidas en el marco conceptual y utilizadas en la sección 4.4 para el modelo original, se aplicaron a los nuevos modelos. Este paso fue crucial para evaluar la eficacia de las técnicas de mitigación implementadas y para entender cómo estas técnicas habían alterado el comportamiento de la calificación de crédito.

Para cada métrica aplicada, se utilizó el método *Bootstrapping* para generar intervalos de confianza al 95%, los cuales, según se explica también en el marco conceptual, consisten en intervalos en los cuales se encuentra el 95% de los resultados obtenidos para las muestras de *Bootstrapping*. Este método estadístico proporciona una medida de la variabilidad de las métricas y permite una interpretación más robusta de los resultados. Al generar estos intervalos de confianza, se pudo obtener una visión más completa de cómo los nuevos modelos se comparaban con los originales y cómo las técnicas de mitigación habían impactado en las métricas de justicia y rendimiento.

4.8 Comparar los modelos alternativos con el modelo original

Finalmente, se realizó una comparación cuantitativa para evaluar cuál de los modelos alternativos implementados proveía una mejora en la justicia respecto al modelo original y un impacto menor en la precisión. Se utilizó la prueba t de Student (sobre los resultados generados por el *Bootstrapping*) para obtener un *p-value* que funcionó como indicador de la significancia estadística en las diferencias observadas en cada métrica respecto al modelo original. En casi todos los casos, se utilizaron muestras pareadas para aplicar el método *Bootstrapping*; es decir, se calculó la respectiva métrica en las mismas muestras del conjunto de prueba. La excepción se da con la métrica de Mancuhan y Clifton [20] debido a que, por la naturaleza de los métodos de pre-procesamiento, los conjuntos de entrenamiento usados para el modelo original y los modelos alternativos basados en la técnica Fairway y la de redes bayesianas no son los mismos que el conjunto de entrenamiento del modelo original, por lo que no es posible hacer este tipo de observaciones pareadas sobre estos conjuntos. En dicho caso, se asumió la independencia de las muestras y se usó una prueba t de Student para muestras independientes.

Ahora, pasamos al capítulo de resultados, donde se explican y se discuten los resultados obtenidos al aplicar la metodología.

CAPÍTULO V. RESULTADOS

A continuación, se explican los resultados obtenidos en cada paso de la metodología. En primer lugar, se comienza por explicar cuáles fueron las posibles fuentes de sesgo encontradas, sea en el conjunto de datos de entrenamiento, en el modelo o durante el uso del modelo. Seguidamente, se resumen los resultados más importantes de las entrevistas. Luego, se explica cómo se obtuvieron las métricas que se usaron a partir de los resultados de los primeros dos pasos metodológicos. En cuarto lugar, se presentan y se discuten brevemente las mediciones realizadas sobre el modelo actual. Después, se explica cómo se obtuvieron las técnicas de mitigación de sesgo usadas y los detalles de implementación que fue necesario considerar en cada técnica. En penúltimo lugar, se presentan los resultados de aplicar las métricas seleccionadas a los modelos alternativos generados. Finalmente, se explican y discuten estos últimos resultados en relación con las mediciones obtenidas sobre el modelo original.

5.1 Posibles fuentes de sesgo del modelo actual

Como se explicó en la metodología, se usaron de manera directa las definiciones de cada tipo de sesgo para tratar de descartar su posible presencia en el modelo actual. **Así, se cumplió de manera directa con el primer objetivo específico, el cual consistía en identificar posibles fuentes de sesgo del modelo actual.** A continuación, se resumen los resultados de la aplicación de estas definiciones.

5.1.1 Sesgos provenientes de los datos de entrenamiento

Comenzando por los sesgos provenientes de los datos de entrenamiento, se usó una prueba de Chi Cuadrado para determinar si existía una correlación entre el género y la caída en impago de los clientes. El resultado obtenido para el estadístico fue de 66.63 con un *p-value* de 3.28×10^{-16}. Esto indica que es muy poco probable que las variables sean independientes; es decir, con muy alta seguridad podemos decir que existe correlación entre el género y la variable "estado" indicadora de la caída en impago, según lo establece la SUGEF. En conclusión, **no es posible descartar la existencia de tratamiento dispar en el conjunto de datos. Aun más, se obtiene un alto grado de seguridad de que sí existe**.

En segundo lugar, como se especificó anteriormente, se usó la prueba H de Kruskall-Wallis para determinar si había correlación entre los atributos predictores del modelo y la variable "género", con el fin de descartar posibles sesgos de asociación. Los resultados para este análisis se muestran en la tabla 5.1. Nótese que, con solo que uno de los atributos tuviese una correlación con

Tabla 5.1. Resultados de aplicar prueba H de Kruskal-Wallis a todas las variables predictoras en relación al género.

Atributo	Resultado del estadístico H	*p-value*	Existe correlación (p<0.05)
Atraso actual	26.51	2.62×10^{-7}	Sí
Edad en crédito	3.36×10^{-4}	0.98	No
Saldo actual	44.84	2.14×10^{-11}	Sí
Tendencia en saldo	0.42	0.52	No
Tiempo desde último atraso	17.44	2.96×10^{-5}	Sí
Atraso promedio	15.31	9.10×10^{-5}	Sí

el género, ya no hubiese sido posible descartar la presencia del sesgo de asociación. En este caso, tenemos una asociación del género con los atributos "atraso actual", "saldo actual", "tiempo desde el último atraso" y "atraso promedio", con *p-values* muy bajos, por lo que se concluye con un alto grado de certeza (al menos 95%) que la correlación existe. Por tanto, no es posible descartar la presencia de sesgo de asociación en el conjunto de datos de entrenamiento.

En tercer lugar, según lo que se explicó en la metodología, se logró obtener información suficiente de las entrevistas para determinar la presencia o no del sesgo de selección. En este caso, debido a la naturaleza de la calificación, el modelo trabaja solamente con clientes constituidos en la institución. **Esto hace muy posible la existencia de un sesgo de selección**, ya que una persona pasa a ser un cliente constituido solamente si:

1. Solicita un préstamo; y
2. La solicitud es aprobada.

Debido a esto, se puede afirmar que existe un sesgo en la selección de los participantes del conjunto de entrenamiento. En conclusión, no es posible descartar la presencia de sesgo de selección en el conjunto de datos de entrenamiento.

Finalmente, nótese que no se hace uso de ningún atributo que sea considerado típicamente sensible, debido a que no se usan ni siquiera datos demográficos de los clientes. Por tanto, **en este caso sí es posible descartar la presencia de sesgo intencional en el conjunto de datos de entrenamiento**.

5.1.2 Sesgos provenientes del entrenamiento del modelo

Como se explicó en la metodología, los sesgos introducidos durante el entrenamiento dependen del proceso seguido y del tipo de modelo utilizado. En este caso el sesgo introducido por medio del uso directo de atributos sensibles no está presente, puesto que el género no es un atributo utilizado durante el entrenamiento del modelo.

Por otra parte, de manera similar al caso del sesgo intencional en los sesgos provenientes de los datos de entrenamiento, se puede descartar este tipo de sesgo durante el entrenamiento, puesto que no se hace uso de ninguna metodología ni se toma ninguna decisión que pretenda favorecer a un grupo sobre otro injustamente [50].

Finalmente, como se menciona en la metodología, no es posible descartar los sesgos de subestimación ni el sesgo malicioso debido a la naturaleza de estos.

5.1.3 Sesgos provenientes del uso del modelo

Los sesgos provenientes del uso del modelo se pueden detectar por medio del análisis del proceso de toma de decisiones posterior al despliegue del modelo. La información obtenida en este caso a partir de las entrevistas explicadas anteriormente fue suficiente para determinar que existen casos en los que el resultado del modelo no se utiliza, debido a políticas internas de la institución. Por ejemplo, si un cliente solicita un préstamo para vivienda de más de 40 millones de colones, su calificación de cliente es calculada pero no se utiliza en la toma de decisión para otorgarle o no el crédito. Por esta razón, no es posible descartar que exista un sesgo de automatización en el uso del modelo.

A continuación, se explican los resultados de las entrevistas.

5.2 Resultados de las entrevistas

A continuación se resumen los resultados más importantes de las entrevistas en formato de respuesta sí/no a cada una de las preguntas mostradas en la sección 4.2. Cabe recalcar que, al ser estos resultados de entrevistas, las respuestas contenían mucha más información de la aquí presentada y, como toda entrevista, el resultado pasó por la interpretación del investigador antes

Tabla 5.2. Resumen de resultados de las entrevistas aplicadas a personas expertas en la institución.

Nombre corto de la pregunta	Respuesta corta
1. Conocer fuente de sesgo	Sí
2. Mitigar sesgos de origen externo a la institución	Sí
3. Hacer análisis estático del modelo	No
4. Mitigar aparición de individuos *token*	No
5. Necesidad de dar explicaciones generales del modelo	No
6. El modelo debe pasar un test situacional	Sí
7. Confidencialidad	No
8. Reducción del riesgo	Sí
9. Criterios legales	No
10. Mitigación de efecto *redlining*	Sí
11. Permite hacer recomendaciones	No
12. Interpretabilidad	Sí
13. Privacidad	Sí
14. Generalizabilidad	No
15. Monotonicidad	No
16. Bien conocida	Sí
17. Dependencia de hiperparámetros	No
18. Exactitud	Sí
19. Mitigación de sesgos en datos	Sí

de ser presentada en este documento. La tabla 5.2 muestra el resumen de los resultados de las entrevistas.

Los resultados obtenidos de las entrevistas con personas expertas en la institución financiera revelan una serie de aspectos cruciales en la evaluación y mitigación del sesgo en los

modelos de calificación crediticia. En primer lugar, se destaca la importancia atribuida al conocimiento de la fuente de sesgo, lo cual refleja la conciencia de la institución sobre la necesidad de comprender los factores subyacentes que pueden influir en los resultados del modelo. Por otro lado, la preocupación por mitigar los sesgos provenientes de fuentes externas a la institución recalca la complejidad del entorno en el que operan los modelos de calificación crediticia.

Respecto a las características específicas deseadas en las técnicas de mitigación de sesgo, se observa un énfasis en la interpretabilidad y la exactitud del modelo, lo que sugiere una preocupación por comprender y confiar en los resultados del mismo. Esta preferencia por modelos comprensibles y precisos refleja la necesidad de transparencia y fiabilidad en el proceso de toma de decisiones crediticias, lo cual es fundamental para garantizar la confianza tanto de los clientes como de los reguladores.

Además, la identificación de la mitigación del efecto *redlining* como un criterio relevante indica una sensibilidad del banco hacia la equidad y la inclusión financiera, aspectos clave en la promoción de la justicia social en el ámbito crediticio.

En resumen, los resultados de las entrevistas ofrecen una visión integral de los criterios y requisitos considerados prioritarios por parte de la institución financiera en la evaluación y mitigación del sesgo en los modelos de calificación crediticia. Estos hallazgos proporcionan una base para la implementación de estrategias de mitigación de sesgo, contribuyendo así a mejorar la equidad y la transparencia en el proceso de calificación de crédito.

5.3 Métricas de justicia que se usaron

Según se explica en la metodología, se realizó una escogencia de las métricas a utilizar con base en la posible presencia de distintos tipos de sesgo, así como otros criterios presentados por los autores de los artículos revisados.

En la figura 5.1 se muestra en más detalle la misma información de las figuras 3.1 y 3.2. Cada fila representa un tipo de sesgo según se describe en el marco conceptual y cada columna

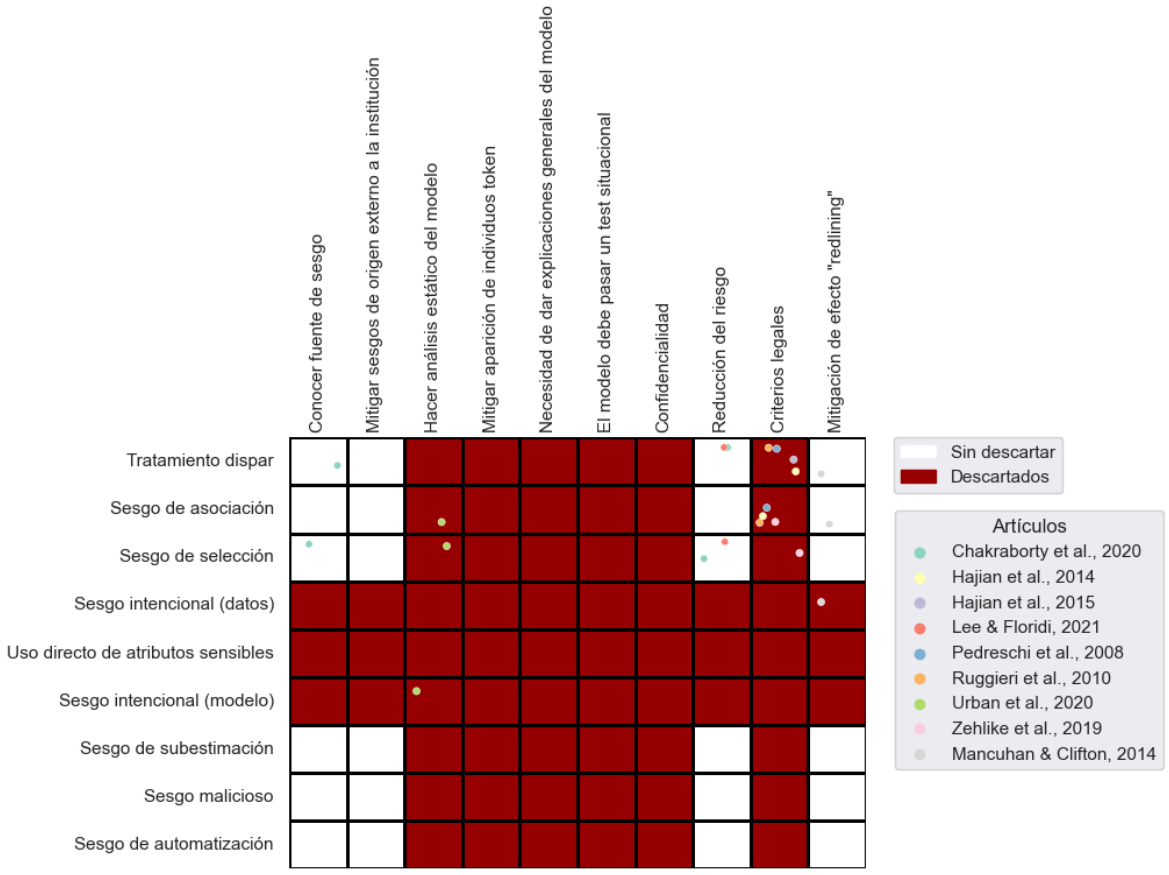

Figura 5.1. Artículos de revisión de literatura clasificados según tipo de sesgo que buscan medir y otras razones para escoger las métricas utilizadas.

representa una de las principales razones dadas para escoger las métricas por los autores de los artículos revisados durante la revisión de literatura. Las columnas, también, corresponden con las principales preguntas realizadas durante las entrevistas iniciales, mostradas en la sección anterior.

Se ha señalado en rojo las filas y columnas que se han descartado gracias a los resultados presentados en las secciones 5.1 y 5.2. Los puntos representan los artículos científicos revisados. Note que un mismo artículo puede encontrarse en más de un cuadrante, debido a que una misma métrica presentada puede funcionar para medir más de un tipo de sesgo o puesto que los autores presentan varias razones para escoger esta métrica.

Por tanto, en este trabajo, el interés radica en aquellas métricas usadas en los trabajos representados en la figura en los cuadrantes de color blanco. La tabla 5.3 presenta dichas métricas. Note que estas corresponden a las métricas de justicia explicadas en el marco conceptual.

Tabla 5.3. Métricas de justicia seleccionadas para medir el sesgo en este estudio.

Artículo	Métrica	Tipo de métrica	Sub-tipo de métrica
Chakraborty, Majumder, Yu & Menzies, 2020 [28]	Porcentaje de puntos que fallan test situacional	Justicia individual	Justicia por contraste
	EOD (Equal Opportunity Difference)	Justicia de grupo	Basada en etiquetas predichas y etiquetas reales
	AOD (Average Odds Difference)	Justicia de grupo	Basada en etiquetas predichas y etiquetas reales
Lee & Floridi, 2021 [29]	Porcentaje de personas en el grupo protegido a las que el modelo asigna un resultado negativo	Justicia de grupo	Basada en etiquetas predichas
Mancuhan & Clifton, 2014 [20]	Porcentaje de individuos discriminados según métrica BEL (Bayesian extended lift)	Justicia causal	Falta de discriminación por proxy

5.4 Mediciones sobre el modelo actual

A continuación, se presenta y se explica el resultado obtenido al aplicar las métricas al modelo de clasificación actual, así como los intervalos de confianza al 95% obtenidos para cada métrica usando el método *Bootstraping*, los cuales denotan el resultado obtenido para el 95% de las pruebas de *Bootstrapping*, como se menciona en la metodología. La tabla 5.4 resume todos estos resultados y los compara con resultados obtenidos para modelos experimentales encontrados en la literatura realizados con el conjunto de datos *German Credit Dataset* [30]. Note que se colocan todos los resultados en términos de porcentajes con el fin de facilitar la comprensión de estos.

En primer lugar, se tiene la métrica del porcentaje de puntos del conjunto de prueba que fallan el test situacional. En este caso particular, esta métrica se mantiene en 0.00%, debido a que el género no es utilizado como variable predictora. Si bien en Costa Rica no existen todavía leyes explícitas que prohíban el uso del género como variable predictora en un modelo de calificación de crédito, se considera generalmente que esta es una práctica que puede conducir a la discriminación y es raro utilizarla. Asimismo, en otras latitudes ya existen reglamentos o leyes explícitas que

Tabla 5.4. Resultados de mediciones de sesgo (y rendimiento) en el modelo actual.

Métrica	Resultado en modelo actual	Intervalo de confianza al 95%	Valor reportado para German Credit Dataset [30]
Porcentaje de puntos que fallan test situacional	0.00%	[0.00%, 0.00%]	8.00% [25]
Porcentaje de personas en el grupo protegido a las que el modelo asigna un resultado negativo	5.93%	[5.52%, 6.31%]	25.00% [25]
EOD	-0.47%	[-0.90%, 0.00%]	0.00% [25]
AOD	-1.70%	[-4.51%, 0.97%]	7.50% [25]
Porcentaje de individuos discriminados según métrica BEL (Bayesian extended lift)	4.12%	[1.71%, 7.77%]	1.00% [20]
Coeficiente de Gini	65.95%	[63.61%, 68.04%]	-

prohíben hacer uso de esta información para determinar la calificación de crédito de la gente [20], por lo que se espera que esta tendencia pueda llegar al país en algún momento.

A modo comparativo con el modelo implementado por [25] para el conjunto de datos *German Credit*, se observa una mejora muy significativa, puesto que este modelo obtiene un 8.00% en esta métrica.

En segundo lugar, se tiene la métrica basada en el porcentaje de personas en el grupo protegido a las que el modelo asigna un resultado negativo. En este caso, se obtuvo un valor de 5.93%, con un intervalo de confianza al 95% de [5.52%, 6.31%]. A modo comparativo, en el modelo implementado por [25] en el conjunto de datos de *German Credit*, se obtuvo un resultado del 25.00%, por lo que, considerando solamente esta métrica, se podría decir que el modelo en estudio es bastante inclusivo.

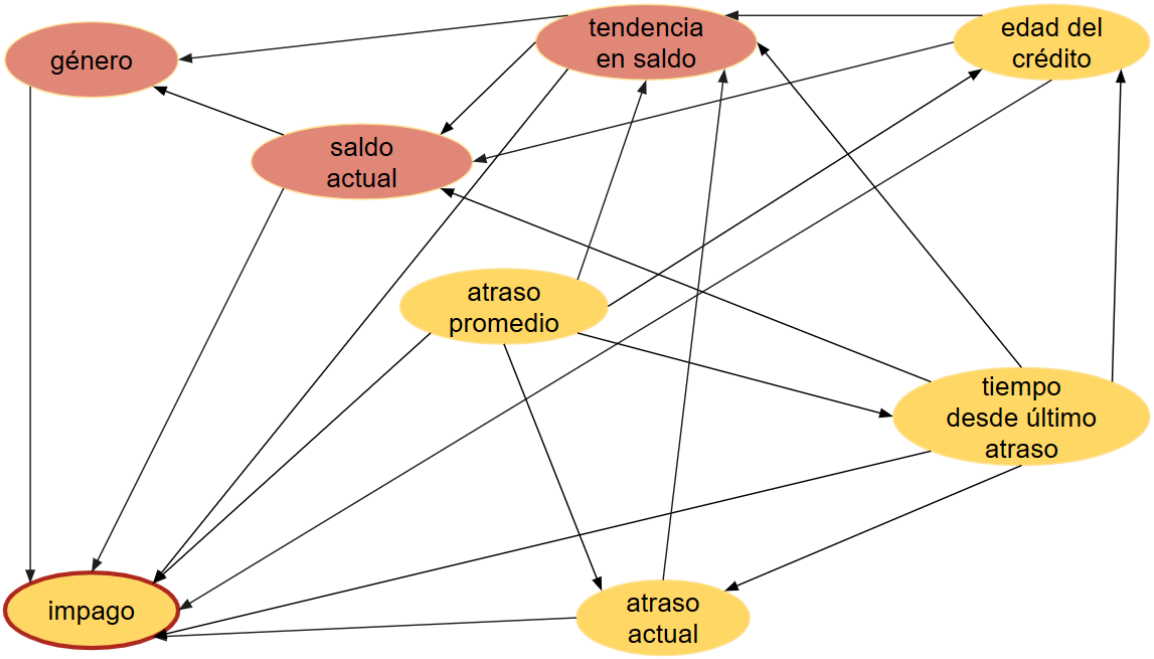

Figura 5.2. Red bayesiana que se obtiene con los datos de entrenamiento.

En tercer lugar, al hacer la medición del EOD en el modelo actual, se obtiene un resultado de -0.47% con intervalo de confianza al 95% de [-0.90%, 0.00%]. Note que un resultado negativo indica que el TPR (la proporción de clientes de bajo riesgo clasificados correctamente) para la clase no privilegiada es mayor que el TPR de la clase privilegiada. Es decir, en este caso se observa que el modelo favorece en realidad a las mujeres. Por otra parte, note que, si bien la métrica es negativa, esta se encuentra muy cercana a cero, por lo que, según el apetito de riesgo de la entidad, podría considerarse un valor aceptable.

En el caso del AOD, se obtuvo un resultado de -1.70% con un intervalo de confianza al 95% de [-4.51%, 0.97%]. Note que, si bien hay un ligero aumento (en valor absoluto) respecto al EOD, este sigue siendo un valor bastante cercano a cero (incluso el intervalo de confianza incluye a cero). Nuevamente, para dimensionar, se puede comparar con el resultado obtenido en [25] para el conjunto de datos de German Credit, el cual fue de 7.50%. Por otra parte, el hecho de que la métrica en este caso también sea negativa, sigue indicando que el modelo de clasificación actual favorece a las mujeres.

La última métrica de justicia utilizada fue la propuesta por Mancuhan y Clifton [20] basada en redes bayesianas. Las figuras 5.2 y 5.3 muestran las redes bayesianas obtenidas al aplicar la métrica. La primera es la red bayesiana que se obtiene a partir de los datos de entrenamiento y la segunda es la versión que se obtiene después de eliminar los nodos correspondientes a la variable

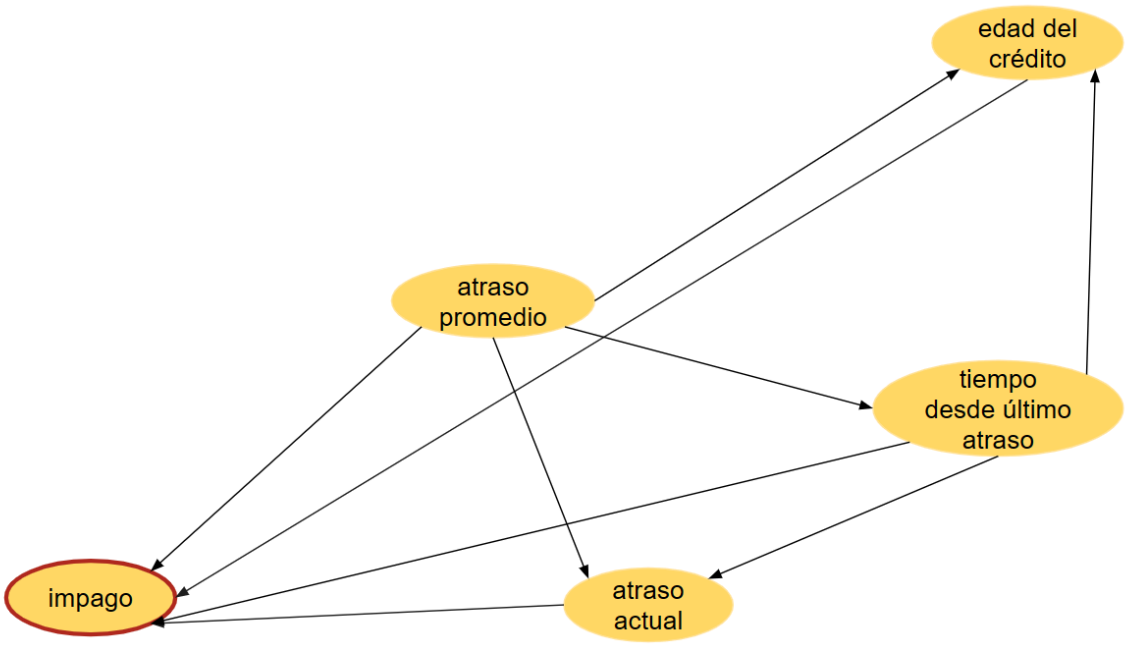

Figura 5.3. Red bayesiana que se obtiene al eliminar atributos protegidos y *redlining*.

protegida (género) y las variables correlacionadas con esta (según la definición en [20]). En este caso, se obtiene un valor de 4.12% para esta métrica de justicia con un intervalo de confianza al 95% de [1.71%, 7.77%]. Note que, según reportan los mismos autores que diseñaron la métrica [20], se obtiene un valor de 1.00% para el conjunto de datos *German Credit*. Se podría interpretar entonces que la métrica obtenida para este caso de estudio es relativamente alta, por lo que valdría la pena investigar el proceso histórico por el cual se ha generado los datos. Esto se explica en más detalle en las conclusiones del estudio.

Finalmente, para completar las métricas aplicadas al modelo actual, se mide el coeficiente de Gini de este sobre el conjunto de datos de prueba. En este caso, se obtiene un coeficiente de 65.95% con un intervalo de confianza al 95% de [63.61%, 68.04%]. Es importante tratar de mantener un coeficiente de Gini alto al aplicar los métodos de mitigación de sesgo. A modo de referencia, se ha observado que los modelos de calificación de crédito tienen normalmente un coeficiente de Gini que se encuentra entre 40% y 60% [51], [52], [53], por lo que en este caso se obtiene un coeficiente de Gini muy favorable en términos del rendimiento del modelo.

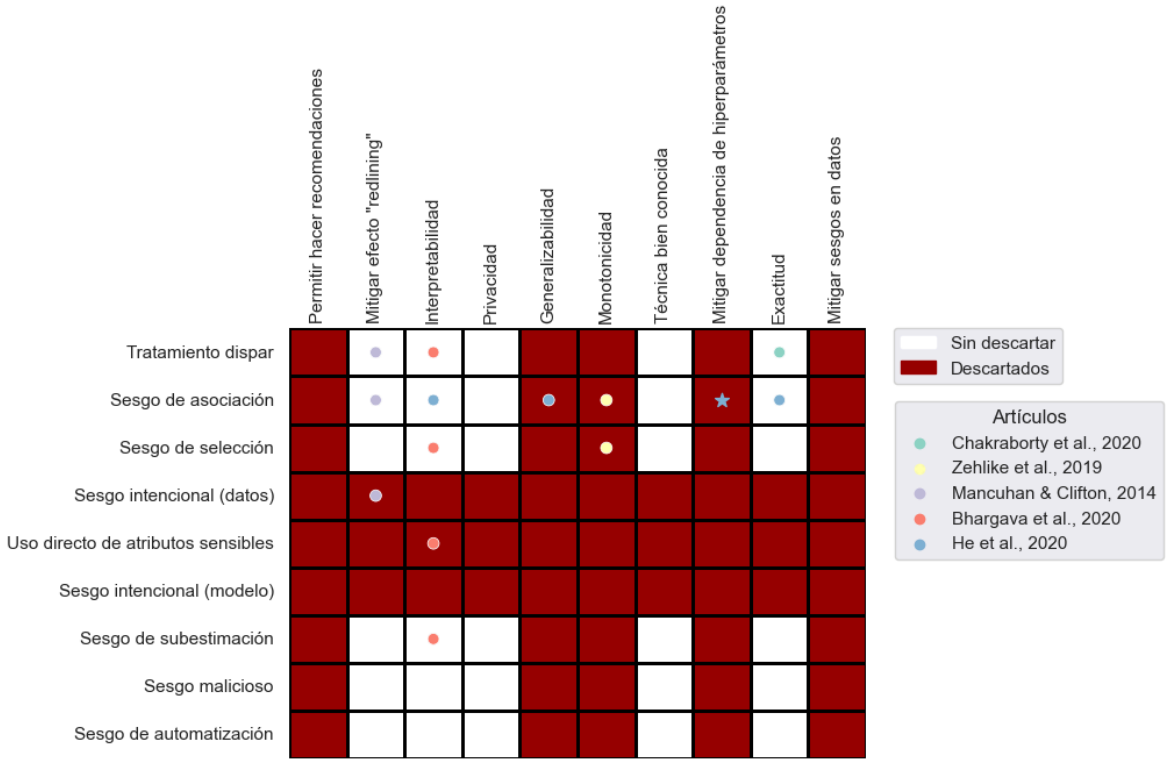

Figura 5.4. Artículos de revisión de literatura clasificados según tipo de sesgo que buscan mitigar y otras razones para escoger las técnicas de mitigación utilizadas.

Al finalizar este punto, se cumplió con el segundo objetivo específico del estudio: medir el grado de sesgo y exactitud del modelo actual.

5.5 Técnicas de mitigación de sesgo usadas

Como se menciona en la metodología, se escogieron las técnicas de mitigación de sesgo de una manera muy similar a como se realizó con las métricas de justicia. La figura 5.4 muestra los artículos revisados catalogados según el tipo de sesgo que pretenden mitigar, así como otras características de las metodologías que permiten hacer la selección de ellas. De esta manera, se descarta el uso de metodologías que pretenden mitigar tipos de sesgo cuya presencia fue descartada en la sección 5.1 y con características que no aplican en este caso de estudio según los resultados de las entrevistas de la sección 5.2. Cabe hacer la anotación de que, a pesar de que la técnica presentada por He, Burghardt y Lerman [10] presenta algunas características deseables, en última instancia se tomó la decisión de descartarla debido a la dependencia de hiperparámetros presentada por los autores.

Tabla 5.5. Técnicas de mitigación de sesgo seleccionadas para generar modelos alternativos.

Artículo	Técnica de mitigación de sesgo	Tipo de técnica	Breve descripción
Chakraborty, Majumder, Yu & Menzies, 2020 [28]	Fairway	Pre-procesamiento y procesamiento	Se eliminan instancias que fallen test situacional y se entrena modelo con optimización multi-objetivo.
Bhargava, Couceiro & Napoli, 2020 [9]	LimeOut	Procesamiento	Se entrena un conjunto de modelos que no dependen de las características protegidas o redlining.
Mancuhan & Clifton, 2014 [20]	Basada en redes bayesianas	Pre-procesamiento y procesamiento	Se identifica y corrige instancias discriminadas y se entrena una red bayesiana sobre los datos corregidos.

La tabla 5.5 resume las técnicas que fueron seleccionadas finalmente: Fairway, presentada por Chakraborty, Majumder, Yu y Menzies en 2020 [28]; LimeOut, presentada por Bhargava, Couceiro y Napoli en 2020 [9]; y la técnica presentada por Mancuhan y Clifton en 2014 [20].

A continuación, se explican detalles de implementación que fue necesario considerar en el presente caso de estudio.

5.6 Modelos alternativos

En esta sección se explican solamente los detalles de implementación que fue necesario considerar cuando se aplicaron las técnicas de mitigación para generar los modelos alternativos. **Nótese que, la aplicación de cada una de las tres técnicas de mitigación seleccionadas generó un nuevo modelo basado en la respectiva técnica que debió ser re-evaluado con las mismas métricas de rendimiento y justicia con las que se midió el modelo de clasificación de crédito original.**

En primer lugar, se generó un modelo alternativo basado en la técnica Fairway [28]. Como se explica en el marco conceptual, esta técnica se compone de dos etapas. En la etapa de pre-procesamiento, se generan dos modelos al particionar los datos en base a su atributo

protegido. En este caso de estudio, se generan entonces dos modelos: uno entrenado solo con hombres (modelo A) y otro solo con mujeres (modelo B). Un detalle importante es que Fairway no especifica el tipo de modelo a utilizar en esta etapa. Para reducir la variación en el experimento, se mantuvo el uso de regresiones logísticas, como en el modelo original.

La etapa de procesamiento de Fairway consiste en la definición de una función objetivo y en su optimización durante el entrenamiento. En este caso, se aplicó la siguiente función, la cual contiene todas las métricas que era necesario optimizar en este estudio para lograr los objetivos deseados (mitigación de sesgo y mantención de una buena exactitud en el modelo alternativo):

$$f(x) = |EOD(x)| + |AOD(x)| + \%personas\ con\ resultado\ negativo\ (x) + (1 - Gini(x))$$

Note que se usó el valor absoluto de las métricas EOD y AOD, puesto que estas métricas representan un resultado más deseable cuanto más cercanas a cero están. Asimismo, note que la métrica basada en redes bayesianas no se usa aquí directamente, puesto que esta no mide el sesgo sobre el modelo sino solamente sobre el conjunto de datos de entrenamiento[9]. Tampoco es necesario incluir la medida de los individuos que fallan el test situacional, puesto que se entrenó el modelo sin usar la variable de género de manera directa, por lo que se sabe que esta métrica siempre será 0.00%.

Al igual que en la etapa de pre-procesamiento, se entrenó el modelo alternativo basado en Fairway usando una regresión logística para reducir la variación del experimento. Sin embargo, note que esta etapa tampoco depende del tipo de modelo utilizado.

El segundo modelo alternativo generado se basa en la técnica LimeOut [9]. Como se explicó anteriormente, en la segunda etapa del método (única que fue necesario aplicar en este caso de estudio), se deben entrenar varios modelos sin cada una de las variables protegidas o *redlining*. Un detalle de implementación importante es que los autores no mencionan cómo determinar cuáles variables tienen efecto redlining, por lo que para efectos de este caso de estudio se tomó como variables redlining aquellas que se detectaron como correlacionadas con el género en la sección 5.1.1 (tabla 5.1), es decir, el atraso actual, el saldo actual, el tiempo desde el último atraso y el atraso promedio. Entonces, en este caso, se entrenó un modelo con todas las variables originales exceptuando el atraso actual, luego otro con todas las variables excepto el saldo actual, otro con todas excepto tiempo desde el último atraso y otro que excluye al atraso promedio.

[9] En Fairway, se hace interesante volver a medir esta métrica después del paso de pre-procesamiento, y así se realizó. Los resultados se observan en la sección 5.7.

Finalmente, se entrenó un modelo que excluía a estas cuatro variables. **El modelo alternativo final basado en LimeOut resultó de tomar el promedio de probabilidad de impago asignado por estos cinco modelos**. Nuevamente, para mitigar varianza innecesaria, se mantiene siempre la calibración usando modelos de regresión logística.

Por último, el tercer modelo alternativo se basa en la técnica presentada por Mancuhan y Clifton [20]. El único detalle de implementación a mencionar es que la re-calibración de la red bayesiana generada durante la medición de la métrica planteada por los mismos autores se realizó usando la versión discretizada de los datos obtenida con el algoritmo presentado en la tabla 2.3.

Al implementar estos modelos, se cumplió con el tercer objetivo específico, que consistía en generar modelos alternativos que pretendiesen mitigar el sesgo.

5.7 Justicia y rendimiento de los modelos alternativos

La tabla 5.6 presenta los resultados obtenidos a partir de la evaluación de los modelos alternativos generados con las técnicas de mitigación de sesgo escogidas. La tercera columna contiene los resultados del modelo actual, mientras que en las siguientes columnas se muestra el resultado de todas las métricas utilizadas para los modelos alternativos. Se muestra en color verde aquellas métricas que muestran una mejora estadísticamente significativa respecto al modelo original, en color rojo las que muestran deterioro estadísticamente significativo y en negro las que no muestran diferencias estadísticamente significativas. Se agrega el *p-value* obtenido al aplicar la prueba t de Student con el método *Bootstrapping* para determinar si las diferencias observadas son estadísticamente significativas, así como el intervalo de confianza al 95% obtenido con el mismo método en cada caso. Además, se coloca en negrita el mejor resultado en cada fila, i.e. en cada métrica. En la siguiente sub-sección se explican en detalle estos resultados.

Tabla 5.6. Resultados de evaluar las métricas de justicia y exactitud en los modelos alternativos y el actual.

Métrica	Valor	Modelo actual	Modelo Fairway	Modelo LimeOut	Red Bayesiana
Porcentaje de puntos que fallan test situacional	Métrica	**0.00%**	**0.00%**	**0.00%**	**0.00%**
	Intervalo de confianza 95%	[0.00%, 0.00%]	[0.00%, 0.00%]	[0.00%, 0.00%]	[0.00%, 0.00%]
	p-value	-	-	-	-
Porcentaje de personas en el grupo protegido a las que el modelo asigna un resultado negativo	Métrica	**5.93%**	7.44%	6.40%	9.87%
	Intervalo de confianza 95%	[5.52%, 6.31%]	[7.00%, 7.88%]	[6.00%, 6.80%]	[9.37%, 10.36%]
	p-value	-	0.00	0.00	0.00
EOD	Métrica	-0.47%	**-0.05%**	0.18%	-0.59%
	Intervalo de confianza 95%	[-0.90%, 0.00%]	[-0.58%, 0.44%]	[-0.28%, 0.65%]	[-1.18%, 0.00%]
	p-value	-	0.00	0.00	3.89×10^{-37}
AOD	Métrica	-1.70%	**-0.91%**	-1.00%	-1.36%
	Intervalo de confianza 95%	[-4.51%, 0.97%]	[-3.73%, 1.78%]	[-3.74%, 1.68%]	[-3.84%, 1.14%]
	p-value	-	1.03×10^{-139}	9.42×10^{-99}	3.68×10^{-7}
Porcentaje de individuos discriminados según métrica BEL (Bayesian extended lift)	Métrica	4.12%	4.13%	4.12%	**0.00%**
	Intervalo de confianza 95%	[1.71%, 7.77%]	[1.62%, 6.22%]	[1.71%, 7.77%]	[0.00%, 5.81%]
	p-value	-	0.02	-	2.57×10^{-23}
Coeficiente de Gini	Métrica	65.95%	**66.52%**	65.48%	62.80%
	Intervalo de confianza 95%	[63.61%, 68.04%]	[64.20%, 68.95%]	[63.16%, 67.68%]	[60.48%, 65.09%]
	p-value	-	3.98×10^{-243}	3.05×10^{-121}	0.00

5.8 Comparación de los modelos alternativos con el modelo original

A continuación, se realiza una comparación de los modelos alternativos con el modelo actual, se interpretan los resultados de la tabla 5.6 y se generan recomendaciones finales. **Cabe**

enfatizar que, con esto, se cumple con el cuarto objetivo específico: evaluar los modelos alternativos respecto a su grado de sesgo de género y exactitud. Esto, a su vez, hace que se cumpla con el objetivo general de construir y evaluar modelos de calificación de cliente que reduzcan los sesgos de género mientras mantenían resultados aceptables para el negocio.

En primer lugar, la métrica de puntos que fallan el test situacional se mantiene siempre en 0.00%. Esto se debe a que en ningún momento se utiliza el género como variable predictora, por lo que el resultado de ningún modelo cambia cuando se cambia el género de la persona. Esto es completamente deseable y puede tomarse incluso como un tipo de escenario base.

En segundo lugar, se observa que el porcentaje de personas en el grupo protegido a las cuales los modelos asignan un resultado negativo (caída en impago) tiende a aumentar en todos los casos respecto al modelo original, de un 5.93% a entre un 6 a 10%. Más adelante se provee una posible explicación de este fenómeno.

El EOD y AOD de todos los modelos alternativos tiende a mejorar, en el sentido de que estas métricas se aproximan más a cero que el modelo original (con excepción del EOD en el modelo basado en redes bayesianas, el cual pasa de un -0.47% en el modelo original a un -0.59% en el modelo de redes bayesianas). Sin embargo, al observar estos resultados en conjunto con los resultados de la métrica de personas en el grupo protegido a las que se les asigna un resultado negativo, se entiende que lo que los modelos alternativos están haciendo es aumentando más la tasa a la que predice impago para las mujeres que la tasa a la que predice impago para los hombres.

La última métrica de justicia, que es la basada en la métrica BEL, solamente mejora con el paso de pre-procesamiento de la metodología de Mancuhan y Clifton [20], pasando de un 4.12% a un 0.00%. Esto es esperable, debido a que esta es la única metodología que fue diseñada justamente con ese objetivo.

Por último, se observa que el coeficiente de Gini tiende a aumentar con el modelo Fairway, pasando de un 65.95% a un 66.52%, lo cual es deseable para esta métrica, puesto que indica una mayor exactitud en el modelo. Las demás metodologías, por el contrario, muestran un deterioro en esta métrica de menos de 5 puntos porcentuales. Si bien, en general es esperable que las métricas de precisión y exactitud de un modelo se deterioren cuando se le aplica una metodología de mitigación de sesgo, también es cierto que en experimentos se han observado mejoras [10]. Esto

podría deberse a que, al aplicar la metodología de mitigación de sesgo, se está reduciendo algún sobreajuste del modelo; sin embargo, hacen falta estudios posteriores para confirmar esto.

En conclusión, se observa que el modelo original de regresión logística presenta números considerados muy favorables en todas las métricas, excepto en la métrica basada en redes bayesianas. Esto indica que el modelo original tiende a ser bastante justo y que quizá el único problema que pueda existir está en los datos de entrenamiento. Esto último, asociado con la información de contexto obtenida de las entrevistas iniciales, sugiere que vale la pena analizar otras calificaciones de crédito utilizadas en la institución, especialmente aquellas que evalúan a clientes nuevos, ya que estas (o el proceso en el que se usan) pueden estar generando un leve sesgo de selección. Sin embargo, esto no es completamente seguro, puesto que también cabe la posibilidad de que el sesgo se produzca de manera externa a la institución. Es decir, que exista una razón externa a la institución por la cual las mujeres que aplican a créditos tienen un mejor perfil de riesgo. Estudios anteriores en este ámbito se han mostrado inconclusos [51].

En general, se observa que todos los modelos alternativos, a pesar de que tienden a mejorar las métricas EOD y AOD, lo hacen a través de un deterioro en el tratamiento del grupo protegido. Esto, puesto que el EOD y AOD inicialmente mostraban más bien un tratamiento que favorece a este grupo. Por esta razón, quizá lo más recomendable sea mantener el modelo original y centrar los esfuerzos en otras áreas. Sin embargo, esto queda a decisión de la institución financiera en estudio.

Lo más interesante del estudio quizá sea los resultados que se obtienen en los modelos alternativos para la métrica BEL. Como se mencionó anteriormente, los valores de esta métrica sugieren la posibilidad de un sesgo de selección en los datos, que sería ideal corregir antes de mitigarlo en el modelo de calificación de cliente. Sin embargo, mientras no se corrija, o si esta tarea resulta imposible, se podría considerar utilizar el paso de pre-procesamiento de la técnica de Mancuhan y Clifton [20] para mitigar dicho sesgo y luego realizar un entrenamiento tradicional del modelo usando los datos de entrenamiento corregidos (i.e. usar una regresión logística y no una red bayesiana como la que los autores proponen, ya que esta deteriora la exactitud del modelo y no mejora tanto las demás métricas de justicia). Otra posibilidad sería la combinación del pre-procesamiento de Mancuhan y Clifton y el procesamiento del método Fairway [28].

En el siguiente y último capítulo, se presentan las conclusiones de este estudio, así como limitaciones y trabajo futuro.

CAPÍTULO VI. CONCLUSIONES Y TRABAJO FUTURO

El trabajo aquí presentado constituyó un importante caso de estudio en el marco del aprendizaje de máquina ético. Se realizó un estudio integral que abarcó desde la selección de las métricas y metodologías a utilizar hasta los resultados y comparativos con el modelo actual.

Es importante hacer notar que se cumplieron los objetivos específicos de la investigación. En primer lugar, se explicó en detalle cómo descartar o no cada posible fuente de sesgo de género y, por tanto, fue posible determinar posibles fuentes de este sesgo. En segundo lugar, el grado de sesgo fue medido con las métricas explicadas previamente. Seguidamente, se construyeron modelos alternativos que buscaran reducir el sesgo de género del modelo actual, con las técnicas que se explicaron. Finalmente, se evaluó y comparó estos modelos respecto a su grado de sesgo, usando las métricas anteriormente seleccionadas, y su rendimiento, usando la ya mencionada métrica de Gini. El cumplimiento de estos objetivos a su vez llevó al cumplimiento del objetivo principal de construir y evaluar modelos de calificación de crédito que redujeran los sesgos de género y que mantuvieran resultados aceptables para el negocio según el apetito de riesgo definido por este.

Cabe destacar nuevamente que, hasta donde se ha encontrado, este es el primer caso de estudio en el cual, no solo se mide, sino que también se busca mitigar el sesgo de género en un modelo de calificación de crédito en uso en una institución financiera.

Uno de los aportes más importantes al estado del arte está dado por la metodología propuesta, la cual provee por primera vez, hasta donde se sabe, una guía práctica para practicantes en el área que deseen replicar este tipo de estudio en otra institución. Este es justamente uno de los grandes faltantes en la investigación mencionados por Balayn, Lofi y Houben [11] en 2021.

Asimismo, existen cuatro mapeos muy importantes generados durante esta investigación que también generan un aporte de gran importancia para investigaciones futuras, a saber:

- La relación entre las métricas de justicia encontradas y el tipo de sesgo que buscan medir según su fuente.
- La relación entre las métricas de justicia encontradas y criterios que se pueden usar para seleccionarlas para su uso en un caso práctico, según sus respectivos autores.
- La relación entre las técnicas de mitigación de sesgo encontradas y el tipo de sesgo que pretenden mitigar según su fuente.

- La relación entre las técnicas de mitigación de sesgo encontradas y criterios que se pueden usar para seleccionarlas para su uso en un caso práctico, según sus respectivos autores.

Estos mapeos son parte esencial de la metodología propuesta en esta investigación, y se insta a futuros practicantes a contribuir a un conjunto de datos que mapee más métricas de justicia y técnicas de mitigación de sesgo en relación con las posibles fuentes de sesgo y criterios de selección para su respectivo uso.

Cabe destacar también el aporte social que esta investigación tiene. El hecho de que este estudio se haya realizado sobre un modelo que se encuentra en uso en una institución real, implica que la institución podría a su vez tomar decisiones respecto a este modelo en base al estudio. Esto a su vez afectará a personas reales que son clientes de esta institución, por lo que en última instancia, la investigación podrá afectar positivamente la vida de muchas personas.

Se encontró que el modelo actual tiene sesgos de género menores. El mayor sesgo encontrado se podría atribuir a un sesgo de selección, ya que los datos que se usan para entrenar el modelo corresponden solamente a datos de clientes constituidos en la institución. Hace falta más estudios para comprender si este sesgo proviene de un tratamiento dispar por parte de la institución hacia los nuevos clientes o si más bien esto se debe a fenómenos sociales fuera del control de la institución.

Aun así, se realizó el ejercicio de generación de modelos alternativos con el fin de mitigar el sesgo de género encontrado. Se observó una pequeña mejora en las métricas obtenidas para los modelos alternativos respecto al modelo de regresión logística original. Sin embargo, también cabe notar que algunas de las métricas sugerían que el modelo original más bien favorece a las mujeres actualmente, por lo que, al aplicar las técnicas de mitigación de sesgo, se obtienen modelos que tienden a evaluar a las mujeres de manera más negativa que el modelo original. Asimismo, es importante notar un aumento en la complejidad de estos modelos respecto al modelo original. Por estas razones, **se recomienda mantener el modelo original y más bien repetir el estudio con otros modelos de calificación de crédito de la institución**, específicamente con modelos de aplicación (según los define Anderson [6]), es decir, modelos utilizados durante la etapa de originación de una nueva operación de crédito para seleccionar a nuevos aplicantes.

6.1 Limitaciones del estudio y trabajo futuro

Existen algunas limitaciones del estudio. En primer lugar, cabe mencionar la exclusión de personas no binarias o que tienen un género que no cabe dentro del binario tradicional. Si bien, considerar a estas personas haría que el estudio fuese más completo, lamentablemente no se contó

con información de personas que tuvieran otros géneros que no fueran masculino y femenino, debido a que tradicionalmente esta información no ha sido recolectada. Un futuro estudio podría considerar la inclusión de más géneros.

En la misma línea, otras formas de ampliar un estudio de este tipo pueden ser: realizar el estudio con base en otras características protegidas, como la etnia o la edad de las personas, o incluso considerar casos interseccionales, en los cuales puede haber discriminación con base en dos o más características protegidas de forma simultánea.

En tercer lugar, cabe recalcar también la exclusión de personas jurídicas. El modelo con el cual se trabajó fue en realidad calibrado usando datos tanto de personas físicas como personas jurídicas. Un trabajo futuro podría considerar incluir al menos algunas de las personas jurídicas. Por ejemplo, MIPYMES unipersonales cuyas dueñas sean mujeres. Ya algunos estudios han sugerido que existe algún tipo de discriminación crediticia hacia este sector en el país [52].

Como se mencionó anteriormente, valdría la pena hacer un estudio más completo dentro de la institución, en el cual se estudie también el posible sesgo proveniente de otros modelos de calificación crediticia en uso. Cabe recalcar que la interacción entre estos modelos se vuelve compleja en la práctica, por lo que un estudio de este tipo sería muy provechoso.

También es importante mencionar que, si bien las probabilidades son bajas, muchas de las técnicas estadísticas aquí utilizadas no nos dan una verdad absoluta, y por tanto existe una posibilidad mínima de haber encontrado correlaciones espurias u otro tipo de conclusiones falaces.

Como se menciona también en el capítulo de resultados, se observó mejoras en los pasos de pre-procesamiento de la metodología basada en redes bayesianas de Mancuhan y Clifton [20], pero no así en el paso de procesamiento. Valdría la pena hacer un estudio posterior en el que se combine la técnica de pre-procesamiento de estos autores con otras técnicas de procesamiento.

Finalmente, al observar las figuras 5.1 y 5.4, se puede notar que hay cuadrantes vacíos. Los estudios que fueron ubicados en estas matrices son aquellos provenientes de una revisión de literatura realizada específicamente con el enfoque del contexto de sesgos de género y calificaciones crediticias. Sin embargo, una revisión sistemática de literatura más completa podría tratar de ubicar un mayor número de estudios en estos cuadrantes. Esto sería una contribución importante para personas en la industria que desean hacer un estudio similar a este. Como se vio aquí, estas matrices fueron unas de las herramientas más importantes al inicio del trabajo para realizar la selección de las métricas y técnicas de mitigación y son uno de los aportes más importantes de este estudio, puesto que anteriormente, y según se menciona en [11], no se había

propuesto una metodología sistemática para seleccionar estas métricas y técnicas para su uso en un caso práctico.

Bibliografía

[1] N. Hermes y R. Lensink, "Impact of microfinance: A critical survey", *Econ. Polit. Wkly.*, vol. 42, núm. 6, pp. 462–465, 2007.

[2] NobelPrize.org, "The Nobel Peace Prize 2006 - Presentation Speech". Consultado: el 31 de mayo de 2022. [En línea]. Disponible en: https://www.nobelprize.org/prizes/peace/2006/ceremony-speech/

[3] M. Hudon, "Should access to credit be a right?", *J. Bus. Ethics*, vol. 84, núm. 1, pp. 17–28, 2009, doi: 10.1007/s10551-008-9670-y.

[4] B. Wittlinger, L. Carranza, y T. Mori, "Best Practices in Collections Strategies", ACCION International, 2008.

[5] K. Brown y P. Moles, *Credit risk management*. Edinburgh: Edinburgh Business School, 2014.

[6] R. Anderson, "The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation", pp. 1–790, 2007.

[7] V. Belavadi, Y. Zhou, M. Kantarcioglu, y B. Thuriasingham, "Attacking Machine Learning Models for Social Good", en *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, pp. 457–471. doi: 10.1007/978-3-030-64793-3_25.

[8] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, y F. Giannotti, *Discrimination- and privacy-aware patterns*, vol. 29, núm. 6. Springer US, 2015. doi: 10.1007/s10618-014-0393-7.

[9] V. Bhargava, M. Couceiro, y A. Napoli, "LimeOut: An Ensemble Approach to Improve Process Fairness", en *ECML PKDD 2020 Workshops*, 2020, pp. 475–491.

[10] Y. He, K. Burghardt, y K. Lerman, "A geometric solution to fair representations", en *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 279–285. doi: 10.1145/3375627.3375864.

[11] A. Balayn, C. Lofi, y G. Houben, "Managing bias and unfairness in data for decision support : a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems", *VLDB J.*, 2021, doi: 10.1007/s00778-021-00671-8.

[12] N. Bailey, B. Carr, y A. Delle-Case, "Machine Learning Thematic Series Part II: Bias and Ethical Implications in Machine Learning", 2019.

[13] Dastin J, "Amazon scraps secret AI recruiting tool that showed bias against women - Reuters", *Reuters*, pp. 1–6, 2018.

[14] J. Angwin, J. Larson, S. Mattu, y L. Kirchner, "Machine Bias — ProPublica", Propublica. Consultado: el 15 de octubre de 2021. [En línea]. Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[15] S. Lohr, "Facial recognition is accurate, if you're a white Guy", *N. Y. Times*, pp. 2018–2021, 2018.

[16] K. Peachey, "Sexist and biased? How credit firms make decisions - BBC News", BBC News. Consultado: el 15 de octubre de 2021. [En línea]. Disponible en: https://www.bbc.com/news/business-50432634

[17] A. Stupnytska, K. Koch, A. MacBeath, S. Lawson, y K. Matsui, "Giving credit where it is due: How closing the credit gap for women owned SMEs can drive global growth", *Goldman Sachs*, vol. 81, núm. 3, 2014.

[18] J. Powers y B. Magnoni, "Dueña de tu propia empresa : Identificación , análisis y superación de las limitaciones a las pequeñas empresas de las mujeres en América Latina y el Caribe Jennifer Powers y Barbara Magnoni EA Consultants", *Fondo Multilater. Inversiones Miemb. Grupo BID*, núm. January 2010, pp. 1–98, 2010.

[19] M. Favaretto, E. De Clercq, y B. S. Elger, "Big Data and discrimination: perils, promises and solutions. A systematic review", *J. Big Data*, vol. 6, núm. 1, 2019, doi: 10.1186/s40537-019-0177-4.

[20] K. Mancuhan y C. Clifton, "Combating discrimination using Bayesian networks", *Artif. Intell. Law*, vol. 22, núm. 2, pp. 211–238, 2014, doi: 10.1007/s10506-014-9156-4.
[21] G. James, D. Witten, T. Hastie, R. Tibshirani, y J. E. Taylor, *An introduction to statistical learning: with applications in Python*. en Springer texts in statistics. Cham, Switzerland: Springer, 2023.
[22] D. Koller y N. Friedman, *Probabilistic graphical models: principles and techniques*, Nachdr. en Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2010.
[23] G. F. Cooper y E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data", *Mach. Learn.*, vol. 9, núm. 4, pp. 309–347, oct. 1992, doi: 10.1007/BF00994110.
[24] I. Schatz, "Using the Gini coeficient to evaluate the performance of credit score models". Consultado: el 27 de febrero de 2022. [En línea]. Disponible en: https://towardsdatascience.com/using-the-gini-coefficient-to-evaluate-the-performance-of-credit-score-models-59fe13ef420
[25] S. Verma y J. Rubin, "Fairness definitions explained", en *Proceedings - International Conference on Software Engineering*, IEEE Computer Society, may 2018, pp. 1–7. doi: 10.1145/3194770.3194776.
[26] P. Cunningham y S. J. Delany, "Underestimation Bias and Underfitting in Machine Learning", vol. 12641, 2021, pp. 20–31. doi: 10.1007/978-3-030-73959-1_2.
[27] O. Trejo, "What Is Bias in Machine Learning?", Scalable Path. Consultado: el 4 de enero de 2024. [En línea]. Disponible en: https://www.scalablepath.com/machine-learning/bias-machine-learning
[28] J. Chakraborty, S. Majumder, Z. Yu, y T. Menzies, "Fairway: A way to build fair ML software", *ESECFSE 2020 - Proc. 28th ACM Jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, pp. 654–665, 2020, doi: 10.1145/3368089.3409697.
[29] M. S. A. Lee y L. Floridi, "Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs", *Minds Mach.*, vol. 31, núm. 1, pp. 165–191, 2021, doi: 10.1007/s11023-020-09529-4.
[30] H. Hofmann, "UCI Machine Learning Repository: Statlog (German Credit Data) Data Set", University of California, Irvine Machine Learning Repository. Consultado: el 6 de junio de 2021. [En línea]. Disponible en: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[31] "BNLearn's Documentation — bnlearn bnlearn documentation". Consultado: el 3 de octubre de 2023. [En línea]. Disponible en: https://erdogant.github.io/bnlearn/pages/html/index.html
[32] M. T. Ribeiro, S. Singh, y C. Guestrin, ""Why Should I Trust You?': Explaining the Predictions of Any Classifier", en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, ago. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
[33] D. Garreau y U. von Luxburg, "Explaining the Explainer: A First Theoretical Analysis of LIME". arXiv, el 13 de enero de 2020. Consultado: el 29 de septiembre de 2023. [En línea]. Disponible en: http://arxiv.org/abs/2001.03447
[34] D. Garreau y U. von Luxburg, "Looking Deeper into Tabular LIME". arXiv, el 18 de julio de 2022. Consultado: el 2 de octubre de 2023. [En línea]. Disponible en: http://arxiv.org/abs/2008.11092
[35] R. S. Witte y J. S. Witte, *Statistics*, 9th ed. Hoboken, NJ: J. Wiley & Sons, 2010.
[36] R. Corrales-Barquero, G. Marín-raventós, y E. G. Barrantes, "A Review of Gender Bias Mitigation in Credit Scoring Models", *2021 Ethics Explain. Responsible Data Sci. EE-RDS*, pp. 1–10, 2021, doi: 10.1109/EE-RDS53766.2021.9708589.
[37] C. Kuhlman, M. A. VanValkenburg, y E. Rundensteiner, "Fare: Diagnostics for fair ranking using pairwise error metrics", en *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2019, pp. 2936–2942. doi: 10.1145/3308558.3313443.
[38] P. Delobelle, P. Temple, G. Perrouin, B. Frénay, P. Heymans, y B. Berendt, "Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning", vol. 23, núm. 1, pp. 32–41,

2020.
[39] M. Zehlike, P. Hacker, y E. Wiedemann, "Matching code and law: achieving algorithmic fairness with optimal transport", *Data Min. Knowl. Discov.*, vol. 34, núm. 1, pp. 163–200, 2020, doi: 10.1007/s10618-019-00658-8.
[40] B. Kulynych, R. Overdorf, C. Troncoso, y S. Gürses, "POTs: Protective Optimization Technologies", en *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 177–188. doi: 10.1145/3351095.3372853.
[41] C. Urban, M. Christakis, V. Wüstholz, y F. Zhang, "Perfectly parallel fairness certification of neural networks", *Proc. ACM Program. Lang.*, vol. 4, núm. OOPSLA, 2020, doi: 10.1145/3428253.
[42] B. Green y Y. Chen, "The principles and limits of algorithm-in-the-loop decision making", *Proc. ACM Hum.-Comput. Interact.*, vol. 3, núm. CSCW, 2019, doi: 10.1145/3359152.
[43] J. Cesaro y F. G. Cozman, "Measuring Unfairness Through Game-Theoretic Interpretability", en *Machine Learning and Knowledge Discovery in Databases*, vol. 8190, 2013, pp. 253–264. doi: 10.1007/978-3-030-43823-4.
[44] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer, y F. Giannotti, "Fair Pattern Discovery", *SAC 14*, pp. 113–120, 2014, doi: 10.1145/2554850.2555043.
[45] Y. Li, Y. Ning, R. Liu, Y. Wu, y W. Hui Wang, "Fairness of Classification Using Users? Social Relationships in Online Peer-To-Peer Lending", en *The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*, 2020, pp. 733–742. doi: 10.1145/3366424.3383557.
[46] S. Ruggieri, D. Pedreschi, y F. Turini, "Data mining for discrimination discovery", *ACM Trans. Knowl. Discov. Data*, vol. 4, núm. 2, pp. 1–40, 2010, doi: 10.1145/1754428.1754432.
[47] Q. Tao, Y. Dong, y Z. Lin, "Who can get money? Evidence from the Chinese peer-to-peer lending platform", *Inf. Syst. Front.*, vol. 19, núm. 3, pp. 425–441, 2017, doi: 10.1007/s10796-017-9751-5.
[48] D. Pedreschi, S. Ruggieri, y F. Turini, "Discrimination-aware Data Mining", *KDD 08*, pp. 560–568, 2008.
[49] SUGEF, "2022 BOTTOM UP STRESS TEST Guía Metodológica", SUGEF, mar. 2022. [En línea]. Disponible en: https://www.sugef.fi.cr/informacion_relevante/manuales/manual_sicveca/Prueba_de_%20Tension-BUST/2022%20BUST%20Guia%20Metodologica.pdf
[50] G. C. Arrieta Portuguez, "Sin título", Banco Nacional de Costa Rica, San José, Costa Rica, 2022.
[51] G. Li, "Gender-Related Differences in Credit Use and Credit Scores", jun. 2018, Consultado: el 30 de septiembre de 2023. [En línea]. Disponible en: https://www.federalreserve.gov/econres/notes/feds-notes/gender-related-differences-in-credit-use-and-credit-scores-20180622.html
[52] "III Informe brechas entre hombres y mujeres, en el acceso y uso del sistema financiero en Costa Rica, 2023", INAMU, MEIC, SUGEF, Banca para el Desarrollo, SUGESE, SUPEN, SUGEVAL, Costa Rica, 3, 2023. Consultado: el 30 de septiembre de 2023. [En línea]. Disponible en: https://www.sugef.fi.cr/informacion_relevante/informe%20brechas%20de%20genero/Informe%20Brechas%20de%20Genero%202023.pdf

ANEXO A. $LIME_{Global}$

En este anexo se explica en más detalle la metodología $LIME_{Global}$ mencionada en el marco conceptual. Como se mencionó en dicho capítulo, en este caso particular no fue necesario aplicar este paso para mitigar el sesgo de acuerdo a lo planteado en la metodología LimeOut. Aun así, se hizo el ejercicio de aplicar esta metodología por completitud y se presentan en este anexo los resultados.

Se recalca, como se mencionó en el marco conceptual, que el objetivo de esta etapa de la metodología LimeOut es determinar los atributos más importantes para el modelo de clasificación actual, y que se basa en el uso de la metodología LIME. A continuación se explica en más detalle esta metodología.

A.1 LIME

La metodología LIME (*Local Interpretable Model-agnostic Explanations*) está diseñada para generar explicaciones que aproximen localmente a un modelo en estudio. El término "explicaciones" se refiere a un conjunto de modelos sencillos de entender, como modelos lineales, y el término "aproximación local" se refiere a modelos que tengan resultados aproximadamente iguales a los del modelo original en un vecindario de la instancia que cada uno busca explicar, donde el vecindario está dado por alguna métrica de distancia entre las instancias de datos. Además, como su nombre lo señala, la metodología es agnóstica al modelo, lo cual quiere decir que, en teoría, funciona de igual manera, sin importar el tipo de modelo en estudio [32].

Para explicar el algoritmo, es necesario dar algunas definiciones. En primer lugar, se denota al modelo en estudio como $f: \mathbb{R}^d \rightarrow \mathbb{R}$, donde $f(x)$ es la probabilidad de que x pertenezca a una clase dada. Se denota por G la clase de modelos interpretables (i.e. las posibles explicaciones). Para cada $g \in G$, su dominio es $\{0,1\}^{d'}$ (note que d' puede depender de la instancia x que se pretende explicar). En esta metodología, se hace una distinción entre las características que usa f y las características *explicables* que usan las $g \in G$. Los autores [32] mencionan que las explicaciones actúan sobre la presencia o ausencia de componentes interpretables, y mencionan como ejemplos el uso de superpixeles en el caso de modelos que trabajen con imágenes o vectores de *bag-of-words* en el caso de modelos de lenguaje natural.

Se define también una medida de complejidad de las explicaciones, denotada por $\Omega(g)$. Esta dependerá del tipo de explicaciones que se estén usando; por ejemplo, en el caso de árboles

Tabla A.1. Técnica LIME. Tomada de [48].

```
LIME(f, N, x, pi_x, Omega)
% Obtiene una explicación g localmente parecida a f en un vecindario de x.
%
% Entradas:
%     - f: el modelo en estudio
%     - N: cantidad de puntos para entrenar g
%     - x: elemento que se desea interpretar
%     - pi_x: medida de cercanía a x
%     - Omega: medida de complejidad de g

Z <- {}

para cada i en 1..N:
    z_i <- muestrear_cerca_de(x)
    z_i' <- transformar(z_i, x)
    Z <- Z ∪ {⟨z_i', f(z_i), pi_x(z_i)⟩}

g <- ajustar_modelo(Z, Omega) % se minimiza Omega y se ponderan las muestras según pi_x

retornar g
```

de decisión puede ser la profundidad del árbol, o en el caso de un modelo lineal puede ser la cantidad de coeficientes no-cero [32].

Para cada $x \in \mathbb{R}^d$, se define una función $\pi_x: \mathbb{R} \to \mathbb{R}$ para definir la cercanía a x. Finalmente, se define $L(f, g, \pi_x)$ como una métrica de qué tan diferente es g de f ponderando según π_x para dar mayor importancia al vecindario cercano a x. Entonces, el algoritmo LIME busca obtener:

$$\xi(x) = argmin_{g \in G}[L(f, g, \pi_x) + \Omega(g)]$$

El algoritmo en sí se observa en la tabla A.1. En resumen, lo que se realiza para una instancia x cuyo resultado bajo f se desea explicar es entrenar un modelo g en base a puntos cercanos a x usando sus versiones interpretables (las cuales pueden depender de x) y el resultado asignado por f para cada uno de estos puntos y ponderando según su cercanía con x, mientras que se minimiza la complejidad de g medida por Ω [32].

Los detalles de las funciones `muestrear_cerca_de`, la cual obtiene un punto cerca de x, `transformar`, que transforma un punto de su versión utilizable por f a su versión localmente interpretable, `pi_x`, que mide la distancia de un punto $z \in \mathbb{R}^d$ a x, `Omega`, que mide la complejidad de g, y `ajustar_modelo`, que entrena un modelo interpretable g usando el conjunto de datos Z y ponderando por los pesos respectivos W, dependen de la aplicación específica. Ribeiro, Singh y Guestrin [32] plantearon estas definiciones para los casos de modelos f que trabajan con lenguaje

natural e imágenes. Estos casos no son de interés para este caso de estudio, por lo que a continuación se salta a la propuesta de Garreau y von Luxburg [33], [34] para datos tabulares.

A.2 LIME tabular

La versión tabular de LIME propuesta por Garreau y von Luxburg [33], [34] necesita de un conjunto de entrenamiento $X \subseteq \mathbb{R}^d$ de donde se obtendrá una cantidad p de cuantiles empíricos para cada uno de los d atributos de los datos. Continuando con la notación de la sección anterior, se definen las funciones anteriormente mencionadas:

- `muestrear_cerca_de`: Esta es la parte más compleja del algoritmo. Teniendo $x = (x_1, x_2, ..., x_d) \in \mathbb{R}^d$, se construye, con los cuantiles anteriormente mencionados, una versión discreta de x de la forma $(\chi_1, \chi_2, ..., \chi_d)$, donde cada $\chi_i \in \{1, ..., p\}$ indica en cuál de los respectivos p cuantiles para la dimensión i se encuentra el atributo x_i de x. Luego, se obtiene, al azar con distribución uniforme, un vector $\theta = (\theta_1, \theta_2, ..., \theta_d) \in \{1, ..., p\}^d$. Finalmente, para obtener la salida $z = (z_1, z_2, ..., z_d)$ de la función, se des-discretiza el vector θ de tal forma que cada z_i se escoja al azar con una distribución normal (con media μ_{i,θ_i} y desviación estándar σ_{i,θ_i} que se obtienen de los datos, para cada dimensión y cada cuantil en la dimensión) truncada al cuantil en la dimensión i correspondiente al valor de θ_i. Por ejemplo, si $\theta = (4, 3, ..., 8)$, entonces z_1 se escoge al azar con una distribución normal truncada al cuantil 4 de la dimensión 1, z_2 se escoge al azar con una distribución normal truncada al cuantil 3 de la dimensión 2, y así sucesivamente [33], [34].

- `transformar`: La versión explicable $z' = (z'_1, z'_2, ..., z'_d) \in \{0, 1\}^d$ de z (la salida de la función anterior) está definida de tal forma que z'_i vale 1 si z'_i está en el mismo cuantil que x_i y 0 si no; o dicho de otra forma, vale 1 si $\theta_i = \chi_i$ o 0 si no [33], [34].

- pi_x: Esta función está dada por:

$$\pi_x(z) = exp\left(\frac{-||z-x||^2}{2\nu^2}\right)$$

donde $||\cdot||$ denota la norma euclidiana en $\mathbb{R}^d$ y $\nu > 0$ es un parámetro que los autores llaman "ancho de banda". El valor de este parámetro aumenta o disminuye la distancia a la

Tabla A.2. Escogencia sub-modular. Tomado de [48].

```
escogencia_sub_modular(X, presupuesto)
% Obtiene un conjunto de instancias de X que.
%
% Entradas:
%     - X: conjunto de datos a explicar
%     - presupuesto: límite de instancias a explicar

para cada x en X:
    g[i] <- LIME(x)
    W[i] <- obtener_parametros(g[i])

para cada j=1..d':
    suma <- 0
    para cada i=1..|X|:
        suma <- suma + abs(W[i][j])
    I[j] <- sqrt(suma)

V <- {}

mientras que |V| < presupuesto:
    V <- V ∪ argmax(lambda i: c(V ∪ {i}, W, I))

retornar V
```

cual se considera que un valor de z es "suficientemente" cercano a x para darle un peso significativo, y como los mismos autores señalan, es un parámetro difícil de afinar [33], [34].

- `Omega` y `ajustar_modelo`: En este punto, estas funciones todavía dependen del tipo de explicaciones que estemos buscando, por lo que aún se dejan a la libre. Sin embargo, los autores de LIME tabular centran sus experimentos en modelos lineales (ponderados), por lo que toman $\Omega = 0$ y definen la función `ajustar_modelo` de tal forma que utilice siempre regularización Ridge (L^2) o Lasso (L^1) para minimizar la complejidad de los modelos explicativos [33], [34].

Bhargava, Couceiro y Napoli [9] usan esta versión tabular de LIME para definir $\text{LIME}_{\text{Global}}$. Solo hace falta explicar cómo combinar las explicaciones locales generadas por LIME tabular para obtener una explicación global y de ahí seleccionar las características más importantes para el modelo en estudio.

A.3 Escogencia sub-modular

Ribeiro, Singh y Guestrin [32] presentan, además de la metodología LIME, un algoritmo al que llaman "escogencia sub-modular", el cual está diseñado para escoger un subconjunto razonable de instancias de un conjunto de datos explicadas por LIME para mostrar a un usuario y que este pueda entender el funcionamiento del modelo en estudio. Bhargava, Couceiro y Napoli [9] se basan

en este algoritmo para combinar las explicaciones locales generadas por LIME en una explicación global del modelo.

El algoritmo de escogencia sub-modular se resume en la tabla A.2. Este es un algoritmo ávido (*greedy*) que busca maximizar una medida de cobertura dada por:

$$c(V, W, I) = \sum_{j=1}^{d'} [\![\exists i \in V : W_{ij} > 0]\!] I_j$$

donde V es el subconjunto de instancias que deseamos evaluar, W es una matriz en la que cada fila corresponde a una explicación dada por LIME para cada instancia del conjunto de datos en estudio (se asume que la explicación puede ser expresada de manera vectorial, e.g. los parámetros de una regresión lineal) e I es un vector que expresa un tipo de "media geométrica" (norma L1) de la importancia de cada variable en cada explicación local [32].

A.4 LIME$_{\text{Global}}$

Bhargava, Couceiro y Napoli [9] notan que el cálculo del vector I en el algoritmo de escogencia sub-modular acaba siendo una forma de combinar las explicaciones locales generadas por LIME tabular en una explicación global. Nuevamente, suponiendo que W es una matriz tal que la entrada W_{ij} en la fila i y la columna j corresponde al parámetro j del modelo g explicable obtenido a partir de la aplicación de LIME tabular a la muestra i en un conjunto de datos de tamaño n, entonces se calcula, para cada j:

$$I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|} \, .$$

En particular, cuando se trabaja con modelos lineales, cada fila i en W contiene un valor que puede ser interpretado como un grado de importancia de los atributos (o sus componentes explicables) para el resultado del modelo original en la instancia i. Al combinar estos atributos en el vector I, se obtiene entonces un vector de importancia de cada atributo en el modelo original a nivel global [9].

Los autores utilizan el algoritmo de escogencia sub-modular para obtener un subconjunto de datos que maximice el cubrimiento explicado anteriormente y luego calculan el vector I usando solamente las instancias obtenidas de la escogencia sub-modular [9].

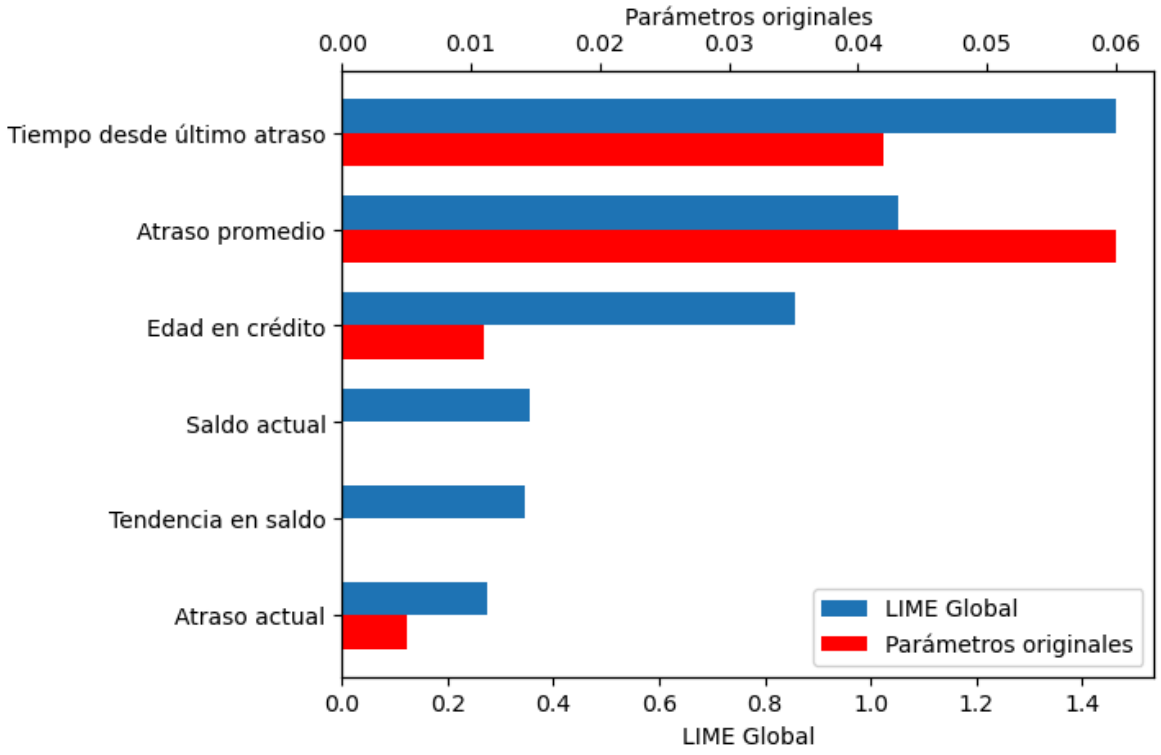

Figura A.1. Resultados de aplicación de LIME Global al caso de estudio.

En la siguiente subsección, se explican los detalles de implementación que se realizaron para aplicar esta técnica al presente caso de estudio, y se detallan los resultados obtenidos.

A.5 Resultados de aplicación al caso de estudio

En la aplicación de este método al presente caso de estudio, se debió tomar en cuenta lo siguiente:

- Debido a limitaciones de poder de procesamiento, no fue posible aplicar LIME al 100% de los datos para obtener W tal como se muestra en el algoritmo de escogencia sub-modular, por lo que se aplicó solamente a una muestra aleatoria de un 1% de ellos.
- Se usaron los siguientes valores para los parámetros:
    - N (cantidad de puntos a muestrear alrededor de cada instancia de datos): 1000
    - ν ("ancho de banda" de métrica de cercanía en $\mathbb{R}^d$, con $d = 6$ en este caso): 50
    - p (cantidad de cuantiles calculados para cada dimensión): 4

Tomando en cuenta dichos detalles, se obtuvieron los resultados que se muestran en la figura A.1. A modo comparativo, se muestra también el valor del parámetro asociado a cada atributo en el modelo de regresión logística del modelo original.