TÍTULO

EQUIDAD EN APRENDIZAJE AUTOMÁTICO. APLICACIONES
=
FAIRNESS IN MACHINE LEARNING. APPLICATIONS

AUTORA

Jessica Coto Palacio

Esta edición electrónica ha sido realizada en 2025

Tutor	Dr. Antonio Javier Tallón Ballesteros
Instituciones	Universidad Internacional de Andalucía; Universidad de Huelva
Curso	*Máster Universitario en Economía, Finanzas y Computación (2023/24)*
©	Jessica Coto Palacio
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	2024

Equidad en Aprendizaje Automático. Aplicaciones

Fairness in Machine Learning. Applications

by

Jessica Coto Palacio

University of Huelva & International University of Andalusia

Julio 2024

« Equidad en Aprendizaje Automático. Aplicaciones »

« Jessica Coto Palacio »

Máster en Economía, Finanzas y Computación

« Supervisor/s » Dr. Antonio J. Tallón Ballesteros

Universidad de Huelva y Universidad Internacional de Andalucía

2024

Abstract

Nowadays, Machine Learning and Artificial Intelligence are having a significant impact on various aspects of our daily lives. This work researches how disparities in training data can affect the fairness of predictive models. Using techniques such as missing values imputation, feature selection and parameter optimization in classification algorithms, together with the incorporation of fairness metrics, several models were evaluated in terms of their ability to reduce bias. To develop these steps, a tool is implemented through which the fairness behavior is studied in three real-world classification datasets. The results indicate that the data preprocessing methods as well as the optimization of specific metrics, can improve fairness keeping the model accuracy. This research highlights the importance of considering fairness in algorithm development in order to implement fairer models.

Keywords: Supervised Learning, Classification Problems, Fairness, Mitigation.

Resumen

Actualmente, el Aprendizaje Automático y la Inteligencia Artificial están teniendo un impacto significativo en varios aspectos de nuestra vida cotidiana. En este trabajo se investiga cómo las disparidades en los datos de entrenamiento pueden afectar la justicia de los modelos predictivos. Utilizando técnicas como imputación de valores perdidos, selección de atributos, y optimización de parámetros en algoritmos de clasificación, junto a la incorporación de métricas de equidad, se evaluaron varios modelos en términos de su capacidad para reducir sesgos. Para desarrollar estos pasos se implementó una herramienta, a través de la cual se estudia el comportamiento de la equidad en tres bases de datos reales de clasificación. Los resultados indican que los métodos de preprocesamiento de datos y la optimización de métricas específicas pueden mejorar la equidad manteniendo la precisión del modelo. Esta investigación resalta la importancia de considerar la equidad en el desarrollo de algoritmos en aras de implementar modelos más justos.

Palabras clave: Aprendizaje Supervisado, Problemas de clasificación, Equidad, Mitigación.

Agradecimientos

Quiero agradecer a mi tutor por su dedicación y tiempo, los cuales han sido fundamentales para la culminación de este trabajo.

A mi madre, mi abuela, mi tía y el resto de mi familia, gracias por su inquebrantable apoyo a pesar de la distancia.

A mis amigos más cercanos, especialmente a Yailen, por su constante ayuda y apoyo incondicional.

A mis compañeros de máster, por todas las experiencias compartidas que enriquecieron esta etapa de mi vida.

Y, finalmente, al claustro de profesores de MECOFIN, quienes han sido pilares esenciales en mi formación a lo largo del máster.

Tabla de contenidos

1 Introducción 1

2 Equidad en Aprendizaje Automático 4

2.1 Tipos de Aprendizaje 4

2.1.1 Aprendizaje supervisado 5

2.1.2 Aprendizaje no supervisado 6

2.1.3 Aprendizaje semi supervisado 6

2.1.4 Aprendizaje por Refuerzo 6

2.2 Algoritmos de clasificación 6

2.2.1 Árboles de Decisión 7

2.2.2 Light Gradient Boosting Machine (LightGBM) 8

2.2.3 eXtreme Gradient Boosting (XGBoost) 8

2.2.4 Máquinas de Vector Soporte (Support Vector Machine) 9

2.3 Métricas de evaluación 10

2.3.1 Matriz de Confusión 10

2.3.2 Accuracy (Exactitud) 10

2.3.3 Precisión 11

2.3.4 Recall (Exhaustividad) 11

2.3.5 F1-Score o F-Measure 11

2.3.6 AUC (área bajo la curva (ROC)) 12

2.3.7 Coeficiente kappa 12

2.4 Equidad en Aprendizaje Automático 13

2.4.1 Métricas de Fairness 15

3 Bases de Datos 21

3.1 Reclamo de seguro de coches 21

3.2 Abandono o éxito académico 22

3.3 Admisión en la escuela de derecho 30

4 Herramienta para el análisis de equidad 32
4.1 Introducción a la herramienta desarrollada 32
4.2 Tratamiento de valores perdidos 34
4.3 Análisis gráfico de las variables 35
4.3.1 Correlograma 36
4.3.2 Gráfico de Cajas (Box Plot) 40
4.3.3 Otros gráficos 41
4.4 Separación del conjunto de datos 44
4.5 Selección de atributos (SelectKBest) 44
4.6 Algoritmos de Clasificación 45
4.6.1 LightGBM 46
4.6.2 Random Forest 46
4.6.3 SVM 47
4.6.4 XGBoost 47
4.7 Técnicas de *Fairness* y mitigación 48
5 Experimentos y Resultados 51
5.1 Reclamo de seguro de coches 52
5.1.1 Experimento 1: Valores por defecto en los parámetros de los algoritmos de clasificación y eliminación de filas con nulos 52
5.1.2 Experimento 2: Optimización de parámetros en los algoritmos de clasificación y eliminación de filas con nulo 55
5.1.3 Experimento 3: Selección de atributos, imputación de los valores perdidos con *k*NN (*k* vecinos más cercano) y valores por defecto en los algoritmos de clasificación 57
5.1.4 Conclusiones parciales 59
5.2 Abandono o éxito académico 59
5.2.1 Experimento 1: Todos los atributos y valores por defecto en los parámetros de los algoritmos de clasificación 59

5.2.2 Experimento 2: Selección de atributos y valores por defecto en los parámetros de los algoritmos de clasificación 61

5.2.3 Experimento 3: Selección de atributos y optimización de parámetros en los algoritmos de clasificación 62

5.2.4 Conclusiones parciales 64

5.3 Admisión a la escuela de derecho 64

5.3.1 Experimento 1: Todos los atributos y valores por defecto en los parámetros de los algoritmos de clasificación 64

5.3.2 Experimento 2: Selección de atributos y valores por defecto en los parámetros de los algoritmos de clasificación 66

5.3.3 Experimento 3: Todos los atributos y optimización de parámetros en los algoritmos de clasificación 67

5.3.4 Conclusiones parciales 69

6 Conclusiones y Trabajo futuro 70

7 Referencias Bibliográficas 71

8 Anexos 76

A.1 Selección de la métrica del clasificador 76

A.2 Selección de la métrica de fairness 76

A.3 Ventana de resultados de las métricas en Random Forest para el *dataset* Reclamo de seguro de coches. Experimento 1. 77

A.4 Ventana de selección de atributos para el *dataset* Reclamo de seguro de coches. Experimento 3 78

Índice de Tablas

Tabla 1. *Interpretación de valores F-Measure.* 11
Tabla 2. *Interpretación de valores ROC Area.* 12
Tabla 3. *Correlación entre Kappa y la fuerza de concordancia.* 12
Tabla 4. *Definición de las variables del dataset Reclamo de Seguros de Coches.* 21
Tabla 5. *Definición de las variables del dataset Abandono o éxito académico.* 23
Tabla 6. *Definición de las variables del dataset Admisión en la escuela de derecho.* 30
Tabla 7*. Resultados de Reclamo de seguro de coches. Experimento 1.* 53
Tabla 8. *Mitigación de los modelos LightGBM y XGBoost con Paridad Demográfica en Reclamo de seguro de coches. Experimento1.* 55
Tabla 9. *Resultados de Reclamo de seguro de coches. Experimento 2.* 56
Tabla 10. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Reclamo de seguro de coches. Experimento 2.* 57
Tabla 11. *Resultados de Reclamo de seguro de coches. Experimento 3.* 58
Tabla 12. *Mitigación de los modelos LightGBM y XGBoost con Paridad Demográfica en Reclamo de seguro de coches. Experimento 3.* 58
Tabla 13. *Resultados de Abandono o éxito académico. Experimento 1.* 60
Tabla 14*. Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Abandono o éxito académico. Experimento 1.* 60
Tabla 15. *Resultados de Abandono o éxito académico. Experimento 2.* 61
Tabla 16. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Abandono o éxito académico. Experimento 2.* 62
Tabla 17. *Resultados de Abandono o éxito académico. Experimento 3.* 63
Tabla 18. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Abandono o éxito académico. Experimento 3.* 63
Tabla 19. *Resultados de Admisión a la escuela de derecho. Experimento 1.* 65
Tabla 20. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Admisión a la escuela de derecho. Experimento 1.* 65
Tabla 21. *Resultados de Admisión a la escuela de derecho. Experimento 2.* 66
Tabla 22. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Admisión a la escuela de derecho. Experimento 2.* 67
Tabla 23*. Resultados de Admisión a la escuela de derecho. Experimento 3.* 68
Tabla 24*. Mitigación de los modelos Random Forest, LightGBM, XGBoost y SVM con Paridad Demográfica en Admisión a la escuela de derecho. Experimento 3.* 68

Índice de Figuras

Figura 1. *Pasos del Aprendizaje Automático (Raschka, 2014).* ____ *5*
Figura 2. *Matriz de confusión.* ____ *10*
Figura 3. *Ejemplo de un modelo de clasificación crediticio.* ____ *17*
Figura 4. *Paridad demográfica.* ____ *17*
Figura 5. *Uso de diferentes umbrales de ingresos para lograr la igualdad de oportunidades.* ____ *18*
Figura 6. *Uso de diferentes umbrales para lograr la equidad de probabilidades.* ____ *19*
Figura 7. *Ventana principal de la Aplicación de Fairness.* ____ *33*
Figura 8. *Segunda ventana. Datos generales del dataset (Reclamo de seguro de coches).* ____ *33*
Figura 9. *Tratamiento de valores perdidos. Reclamo de seguro de coches.* ____ *34*
Figura 10. *Ventana de selección de gráficos.* ____ *35*
Figura 11. *Distribución de las clases de Reclamo de seguro de coches.* ____ *36*
Figura 12. *Distribución de las clases de Admisión en la escuela de derecho.* ____ *36*
Figura 13. *Distribución de las clases de Abandono o éxito académico.* ____ *36*
Figura 14. *Correlograma. Abandono o éxito académico.* ____ *38*
Figura 15. *Correlograma. Admisión a la escuela de derecho.* ____ *39*
Figura 16. *Correlograma. Reclamo de seguro de coches.* ____ *39*
Figura 17. *Diagrama de Cajas. Abandono o éxito académico.* ____ *40*
Figura 18. *Diagrama de Cajas. Admisión a la escuela de derecho* ____ *41*
Figura 19. *Diagrama de Cajas. Reclamo de seguro de coches.* ____ *41*
Figura 20. *Diagrama de Barras (Género). Abandono o éxito académico.* ____ *42*
Figura 21. *Diagrama de Barras (Raza). Admisión en la escuela de derecho.* ____ *42*
Figura 22. *Diagrama de Barras (Sexo). Reclamo de seguro de coches.* ____ *43*
Figura 23. *Diagrama de Barras (Casado). Reclamo de seguro de coches.* ____ *43*
Figura 24. *Ventana de separación en conjunto de entrenamiento y prueba.* ____ *44*
Figura 25. *Ventana de selección de atributos.* ____ *45*
Figura 26. *Ventana de Algoritmos de Clasificación.* ____ *48*
Figura 27. *Ventana de Fairness.* ____ *49*
Figura 28. *Dashboard con los resultados de fairness.* ____ *50*
Figura 29. *Ventana de Mitigación de Fairness.* ____ *50*
Figura 30. *Dashboard Fairlearn en Reclamo de Seguro de coches con LightGBM y Demographic Parity. Experimento 1.* ____ *54*
Figura 31. *Mitigación del modelo Random Forest con Paridad Demográfica en Reclamo de seguro de coches. Experimento 1.* ____ *55*

1 Introducción

El Aprendizaje Automático (AA) es una rama de la Inteligencia Artificial (IA) que se enfoca en el uso de datos y algoritmos para simular la forma en la que aprenden los seres humanos, con un progreso paulatino de su precisión. En consecuencia, es capaz de inferir valores de salida a partir de muestras para las que fue entrenado. Esta capacidad de aprender es muy empleada en la industria, el ámbito sanitario, la robótica, entre muchas otras (Bundi, 2024; Nahavandi et al., 2024; Rodriguez-Fernandez & Camacho, 2024).

Uno de los tipos de aprendizaje más utilizados dentro del *Machine Learning* es el Aprendizaje Supervisado, el cual consiste en categorizar datos a partir de información previa. La clasificación se realiza en dos fases. Primero, se aplica un algoritmo de clasificación o regresión (según el problema a resolver) al conjunto de datos de entrenamiento y, a continuación, el modelo extraído se valida frente a un conjunto de datos de prueba con ejemplos etiquetados para medir el rendimiento y la precisión del modelo. Podemos encontrar diversas aplicaciones de esta rama, entre las que se encuentran la clasificación de documentos, el filtrado de *spam*, la clasificación de imágenes, la detección de fraudes, el análisis de riesgos, etc (Ribeiro et al., 2016).

Dependiendo del tipo de datos de la variable de salida podemos dividir los problemas en clasificación o regresión. Según la bibliografía consultada, algunos de los algoritmos más utilizados en la actualidad para problemas de clasificación incluyen Árboles de Decisión (Elhazmi et al., 2022; Wang & Zhang, 2023), XGBoost (Joe & Kim, 2024; Niazkar et al., 2024) y LightGBM (S. Li et al., 2024; Xi, 2024). Aunque existen muchos otros algoritmos, como las Máquinas de Vector Soporte (Han et al., 2023; Ray et al., 2024), que han demostrado ser muy efectivos para ciertos tipos de problemas, este trabajo se enfoca en los problemas de clasificación utilizando los algoritmos mencionados anteriormente.

En la era de la inteligencia artificial y el *machine learning*, la equidad se ha convertido en un tema crucial. La equidad (*fairness*) en el aprendizaje automático implica que los modelos deben tratar a todas las personas y grupos de forma justa, sin mostrar preferencia por ningún grupo en particular. Este principio abarca varios aspectos, como la equidad en términos demográficos, la igualdad de oportunidades, el logro de resultados equitativos y la imparcialidad en las consecuencias de las decisiones automatizadas (Caton & Haas, 2024).

Los problemas de equidad corresponden a aquellos problemas en los que intervienen atributos de tipo sensible, como el sexo, la nacionalidad, la raza, y muchos más, y que pueden conllevar a que los algoritmos de aprendizaje puedan ser sesgados y, por lo tanto, pueden tomar decisiones injustas o discriminatorias (Ferrara et al., 2024). Por ejemplo, un algoritmo de contratación que se entrene con datos históricos puede estar sesgado por razones de género si hay historial de selección de cierto grupo de individuos con respecto a otros. De la misma manera, un modelo predictivo de justicia penal puede discriminar a ciertos grupos raciales. Es flexible trabajar en desarrollar técnicas para evitar los sesgos, y es necesario evaluar la equidad de los modelos de manera continua para asegurar el trato justo de todas aquellas personas que puedan ser afectadas con una decisión automatizada dependiendo de los atributos de tipo sensible (Wang et al., 2024).

Por otra parte, la industria del desarrollo de software ha implementado numerosas herramientas que facilitan el trabajo de los analistas y científicos de datos. Estas herramientas no solo mejoran la eficiencia en el manejo y análisis de grandes volúmenes de datos, sino que también proporcionan capacidades avanzadas para la detección y mitigación de sesgos en los modelos de aprendizaje automático. En este contexto, el presente trabajo implementa una herramienta diseñada para asistir en el proceso de descubrimiento y análisis de sesgos en tres bases de datos específicas.

Por lo antes expuesto se establece como objetivo general de este trabajo: Realizar un análisis de equidad en tres bases de datos, cada una con al menos un atributo sensible, utilizando tres de las métricas más empleadas en la literatura.

A partir de lo anterior se plantea como objetivos específicos:

- Realizar un análisis de la literatura relacionada con la equidad en el Aprendizaje Automático.
- Determinar las métricas de equidad más utilizadas en problemas de clasificación.
- Elegir tres bases de datos que contienen al menos un atributo sensible.
- Desarrollar una herramienta que permita realizar un análisis de *fairness* en las bases de datos escogidas.
- Seleccionar el modelo que logre el mejor balance entre las métricas de clasificación y las métricas de *fairness*.

Para dar cumplimiento a los objetivos específicos el presente trabajo se estructura de la siguiente manera: un capítulo dedicado a introducir el concepto de Aprendizaje Supervisado y

cómo abordar la equidad en problemas de clasificación. En el capítulo 3 se detallan las bases de datos escogidas para realizar el análisis de equidad haciendo uso de la herramienta desarrollada, la cual se introduce en el capítulo 4. El capítulo 5 está dedicado a resumir los experimentos diseñados para cada conjunto de datos. Y finalmente en el capítulo 6 se arriba a conclusiones sobre el análisis realizado y se proponen ideas de trabajo futuro.

2 Equidad en Aprendizaje Automático

En la era digital actual, el Aprendizaje Automático y la Inteligencia Artificial están transformando numerosos aspectos de nuestra vida cotidiana. Desde sistemas de recomendación hasta decisiones de crédito, estos algoritmos están cada vez más presentes en procesos que influyen en nuestras vidas de forma significativa. Sin embargo, a medida que confiamos más en estos sistemas para automatizar decisiones importantes, surge una preocupación crítica: ¿son justos e imparciales?

El Aprendizaje Automático o *Machine Learning* es una rama de la IA que puede ser vista como una técnica que mejora el rendimiento del sistema aprendiendo a través de la experiencia, utilizando métodos computacionales. En los sistemas informáticos la experiencia viene dada en forma de datos y el principal objetivo del AA es desarrollar algoritmos de aprendizaje que construyan modelos a partir de los datos. Al alimentar este algoritmo con datos, obtenemos un modelo que puede hacer predicciones sobre nuevas observaciones (Z. H. Zhou, 2021).

El concepto de *fairness* o equidad en el aprendizaje automático ha ganado importancia en los últimos años debido a preocupaciones sobre la discriminación algorítmica y los sesgos inherentes en los datos y los modelos. A medida que las decisiones automatizadas impactan a individuos y comunidades enteras, es fundamental abordar estas preocupaciones y garantizar que nuestros sistemas sean éticos, equitativos y respeten los derechos humanos fundamentales.

En este capítulo comenzaremos introduciendo los diferentes tipos de aprendizaje y los algoritmos que se pueden encontrar dentro del AA, para posteriormente pasar a introducir las nociones básicas de *fairness* y las métricas existentes para evaluar la misma y tratar de asegurar que nuestros modelos sean justos e imparciales.

2.1 Tipos de Aprendizaje

Existen diversos algoritmos que resuelven problemas utilizando *Machine Learning,* el algoritmo empleado dependerá entre otros factores del problema a resolver, así como del número de variables involucradas. Los tipos de aprendizaje que se pueden encontrar dentro de *Machine Learning* son: Aprendizaje Supervisado, Aprendizaje no Supervisado, Aprendizaje Semi-Supervisado y Aprendizaje por Refuerzo.

2.1.1 Aprendizaje supervisado

Según (Learned-Miller, 2014), el aprendizaje supervisado es simplemente una formalización de ideas de aprendizaje a partir de ejemplos. En aprendizaje supervisado, al programa que aprende se le proveen dos conjuntos de datos, uno de entrenamiento (*train)* y otro de prueba (*test*). Como se muestra en la figura 1, la idea fundamental es aprender a partir del conjunto de entrenamiento formado por variables o etiquetas de entrada asociadas a una salida conocida y que el programa sea capaz de inferir, dado nuevos casos, la salida que más se ajusta.

Atendiendo al tipo de dato de la variable de salida (numérica o categórica), los problemas de Aprendizaje supervisado se pueden categorizar como problemas de regresión o de clasificación respectivamente. En los problemas de regresión se predice un valor de salida de naturaleza numérica o cuantitativa, los modelos más utilizados en este campo son el de regresión lineal y el de regresión polinómica.

Por otra parte, el objetivo de los problemas de clasificación es construir un modelo a partir de la distribución de etiquetas de clase (variable dependiente) en términos de características predictivas (variables independientes). El clasificador resultante es luego utilizado para asignar etiquetas de clase a las instancias donde se conocen los valores de las características predictoras pero se desconoce el valor de la etiqueta de la clase resultante (Castelli et al., 2018). Actualmente existe una amplia gama de algoritmos dedicados a resolver este tipo de problemas, podríamos mencionar algunos como la Regresión Logística, Árboles de Decisión, Máquinas de Vector Soporte (SVM), *k* vecinos más cercanos (*k*NN), entre muchos otros.

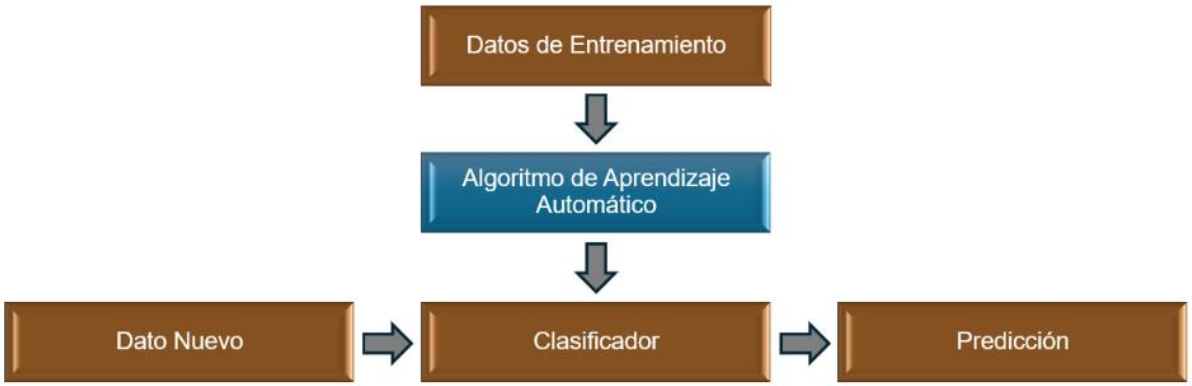


Figura 1. *Pasos del Aprendizaje Automático* (Raschka, 2014).

2.1.2 Aprendizaje no supervisado

Los métodos de aprendizaje automático no supervisados son particularmente útiles en tareas de descripción porque tienen como objetivo encontrar relaciones en una estructura de datos sin tener datos previos de salida. Esta categoría de aprendizaje automático se denomina no supervisada porque carece de una variable de respuesta que pueda supervisar el análisis (James et al., 2013).

2.1.3 Aprendizaje semi supervisado

Las técnicas de aprendizaje semi supervisadas están compuestas por una combinación entre los métodos supervisados y no supervisados. En estos existen datos etiquetados y no etiquetados en el proceso de entrenamiento. Generalmente, considera una cantidad menor de datos etiquetados y una cantidad mayor de datos no etiquetados. Este tipo de técnicas pueden ajustarse por sí solas para obtener una mayor precisión. Es preferible en caso de que los datos etiquetados adquiridos necesiten recursos hábiles y apropiados para entrenar o aprender de ellos. En el resto de los casos, los datos obtenidos sin etiquetar no necesitan recursos extra (Saravanan & Sujatha, 2018).

2.1.4 Aprendizaje por Refuerzo

El Aprendizaje por Refuerzo o RL (terminología en inglés) se asocia con aprender qué hacer (cómo asociar situaciones con acciones) de forma tal que se maximice una señal numérica de recompensa. En este tipo de problemas un agente debe aprender comportamientos a través de interacciones a ‘prueba y error’ con un ambiente dinámico. Un agente es un sistema computacional situado en algún ambiente, que es capaz de actuar de manera autónoma y flexible en dicho ambiente en aras de cumplir con su objetivo (Sutton & Barto, 2018).

Teniendo en cuenta las características de los diferentes tipos de aprendizaje, este trabajo se enfocará en aprendizaje supervisado y específicamente en algoritmos de clasificación, haciendo énfasis en las técnicas que se detallan en la próxima sección.

2.2 Algoritmos de clasificación

Cuando se utiliza un algoritmo de clasificación, el resultado que se obtiene es una clase entre un número limitado de clases. Donde el término clase se refiere a categorías arbitrarias según el tipo de problema. Por ejemplo, si se desea detectar si un correo es *spam* o no, sólo hay 2

clases. Si se desea identificar en una imagen si estamos viendo un perro, un gato o una oveja entonces tendríamos tres posibles clases. O incluso se pueden tener algoritmos donde se den probabilidades como resultado de la clasificación, y entonces se puede decir que un correo tiene un 85% de ser *spam*, o que una imagen tiene 75% de probabilidad de ser un perro. Normalmente en este caso para clasificar se escoge la clase con la probabilidad más alta.

Hay varias técnicas de aprendizaje supervisado que pueden ser utilizadas para resolver problemas de clasificación, y en esta sección se detallarán las más utilizadas en la literatura.

2.2.1 Árboles de Decisión

Los árboles de decisión constituyen una de las técnicas más utilizadas para problemas de clasificación. A partir de los datos de entrenamiento, los árboles de decisión dividen recursivamente el espacio de características y asignan una etiqueta a cada partición resultante. Luego, con el árbol conformado se clasifican observaciones futuras de acuerdo con estas divisiones y etiquetas. La principal ventaja de los árboles de decisión sobre otros métodos es que son muy fáciles de interpretar y, en muchas aplicaciones, como la atención médica, esta interpretabilidad suele preferirse a otros métodos que pueden tener una mayor precisión pero que son relativamente imposibles de interpretar (Rokach & Maimon, 2010).

A lo largo de los años se han desarrollado modelos más robustos basados en árboles, que incluyen la utilización de varios árboles de decisión, ejemplo de estos es *Random Forest.* Este modelo ajusta varios árboles a un conjunto de datos para luego combinar todas las predicciones. El algoritmo primeramente selecciona numerosas muestras de arranque de los datos, aproximadamente el 63% de las originales ocurren al menos una vez. Las observaciones que no ocurren en una muestra de arranque se denominan observaciones fuera de la bolsa. Cada vez que se realiza una nueva división del árbol, se elige de forma aleatoria entre el conjunto completo de predictores una representación que pasan a formar parte del conjunto de los predictores candidatos a controlar la bifurcación. Una vez desarrollados completamente los árboles, se utilizan para predecir aquellas observaciones fuera de la bolsa. La clase pronosticada de una observación se calcula por mayoría de votos de las predicciones originales para esa observación, con empates divididos al azar (Cutler et al., 2007).

2.2.2 Light Gradient Boosting Machine (LightGBM)

LightGBM, desarrollado por *Microsoft Corporation*, se presenta como un modelo versátil que abarca tanto la clasificación como la regresión, fundamentado en árboles de decisión. Sus atributos destacados incluyen una velocidad de entrenamiento superior y eficacia ampliada, un menor consumo de memoria, una precisión elevada, y la capacidad de ofrecer soporte para el aprendizaje paralelo, distribuido y GPU. Además, se destaca por su capacidad para manejar conjuntos de datos a gran escala.

Este modelo propone la sinergia de dos técnicas innovadoras: *Gradient-based One-Side Sampling* (GOSS) y *Exclusive Feature Bundling* (EFB) (Ke et al., 2017). Con GOSS, se excluye una proporción significativa de instancias de datos que presentan gradientes pequeños, utilizando solo el resto para estimar la ganancia de información. Por otro lado, mediante EFB, se agrupan características mutuamente excluyentes, es decir, aquellas que rara vez toman valores distintos de cero simultáneamente, con el objetivo de reducir el número total de características. En el caso de EFB, se emplea un algoritmo *greedy* para encontrar una aproximación eficiente entre las características, asegurando así que no se vea comprometida la precisión del algoritmo.

En los últimos años, LightGBM ha emergido como un modelo ampliamente empleado en diversas aplicaciones. Por ejemplo, ha sido utilizado con éxito en la clasificación de imágenes del habla de caracteres chinos (Pan et al., 2023), en sistemas de detección de intrusos (Ayubkhan et al., 2023), en la predicción de la clase de masa rocosa de túneles en proyectos de construcción (L. Li et al., 2023), y en el diagnóstico inteligente de averías en agujas de ferrocarril (Lao et al., 2023). Esta versatilidad y eficacia en aplicaciones tan variadas destacan la robustez y adaptabilidad de LightGBM en distintos contextos.

2.2.3 eXtreme Gradient Boosting (XGBoost)

XGBoost representa una técnica de aprendizaje automático escalable que emplea la estrategia de *boosting* para mitigar el sobreajuste. Recientemente, ha captado la atención de la comunidad investigadora al superar a numerosos clasificadores tradicionales en el ámbito del aprendizaje automático (Liew et al., 2021).

Este método, además de estar fundamentado en árboles, optimiza el rendimiento tanto del *random forest* como del *gradient boost*. Una de sus ventajas distintivas es su capacidad para

gestionar de manera automática los datos faltantes, reduciendo al mínimo la necesidad de realizar preprocesamiento e imputación. Esto lo hace particularmente adecuado para conjuntos de datos extensos, donde la gestión eficiente de grandes cantidades de información es esencial.

También emplea procesamiento simultáneo, poda de árboles y la combinación secuencial de múltiples árboles de decisión. Este enfoque busca aprender de las salidas de los árboles anteriores y rectificar los errores generados por estos, continuando este proceso hasta que no sea posible corregir más desviaciones (T. Chen & Guestrin, 2016).

2.2.4 Máquinas de Vector Soporte (Support Vector Machine)

En 1995, Cortes y Vapnik presentaron las Máquinas de Vector Soporte más conocidas por sus siglas en inglés SVM, aunque su destacada habilidad para generalizar, su solución óptima y su eficaz capacidad de discriminación han captado el interés de las comunidades de minería de datos, reconocimiento de patrones y aprendizaje automático en los años recientes (Cervantes et al., 2020).

Este modelo genera funciones de decisión a partir de los datos de manera directa, buscando maximizar el margen de separación entre los límites de decisión en un espacio de características altamente dimensional. Una ventaja significativa de las SVM radica en su capacidad para identificar un subconjunto de vectores de soporte durante la fase de aprendizaje, el cual suele ser solo una fracción pequeña del conjunto de datos original (Cervantes et al., 2020).

La SVM lineal asume que los datos multidimensionales pueden ser separados de manera lineal en el espacio de entrada. El hiperplano óptimo, o margen máximo, se puede definir tanto matemática como geométricamente. Este hiperplano representa un límite de decisión que minimiza los errores de clasificación durante el entrenamiento (Sheykhmousa et al., 2020). Sin embargo, en la práctica, las muestras de datos de diferentes clases no siempre son linealmente separables y pueden superponerse. Por lo tanto, la SVM lineal puede no garantizar una alta precisión en la clasificación de tales datos y puede necesitar ajustes. Cortes y Vapnik (Cortes & Vapnik, 1995) introdujeron métodos como el margen suave y el *kernel* para abordar estas limitaciones de la SVM lineal.

La efectividad de la SVM está estrechamente ligada a la elección adecuada de una función de *kernel* que produzca productos de puntos en un espacio de características de dimensión superior. Este espacio, teóricamente infinito en dimensión, permite la posibilidad de

discriminación lineal. Diversos modelos de kernel se utilizan para construir distintas SVM, como Sigmoid, Radial, Polynomial y Linear (Cherkassky & Ma, 2004).

2.3 Métricas de evaluación

Las siguientes métricas brindan información sobre la capacidad de un modelo para identificar correctamente las clases en un conjunto de datos. Estas métricas tienen en común que se utilizan para evaluar el rendimiento de un modelo de clasificación y para ayudar a mejorar su desempeño.

2.3.1 Matriz de Confusión

La Matriz de Confusión es un instrumento que permite representar y valorar la exactitud de un modelo en la identificación de las clases en un conjunto de datos (Susmaga, 2004). Esta se representa en una matriz de dos dimensiones que muestra el número de veces que un modelo clasificó correctamente o incorrectamente cada clase. A través de la Matriz de Confusión se puede calcular otras métricas importantes, como la Precisión, el *Recall* y el F1-*Score* o F-*Measure*, y es esencial en el análisis de modelos de clasificación y en la mejora de su desempeño.

		PREDICCIÓN	
		0	1
REALIDAD	0	TN	FP
	1	FN	TP

Figura 2. *Matriz de confusión.*

Verdaderos Negativos [TN], Verdaderos Positivos [TP], Falsos Positivos [FP], Falsos Negativos [FN].

2.3.2 Accuracy (Exactitud)

La exactitud (*Accuracy*) mide el porcentaje de casos que el modelo ha acertado. Esta es una de las métricas más usadas. El problema con la exactitud es que nos puede llevar al engaño, es decir, puede hacer que un modelo malo parezca mucho mejor de lo que en realidad es. Esta métrica no funciona bien cuando las clases están desbalanceadas (Chicco & Jurman, 2020).

$$accuracy = \frac{TP+TN}{\mathrm{TP}+TN+FP+FN} \qquad (1)$$

2.3.3 Precisión

La Métrica de Precisión evalúa la efectividad de un modelo de clasificación en el reconocimiento de las clases en un conjunto de datos (Davis & Goadrich, 2006). La Precisión se define como el número de veces que el modelo identifica correctamente una clase positiva de entre todas las clases que el modelo ha identificado como positivas.

$$precisión = \frac{TP}{TP+FP} \qquad (2)$$

2.3.4 Recall (Exhaustividad)

Se conoce también como exhaustividad y es comúnmente utilizada para medir la capacidad del modelo para detectar correctamente todas las instancias positivas de una determinada clase (Davis & Goadrich, 2006). *Recall* se define como el número de verdaderos positivos dividido entre el número total de instancias positivas en el conjunto de datos.

$$recall = \frac{TP}{TP+FN} \qquad (3)$$

2.3.5 F1-Score o F-Measure

La métrica utilizada para combinar en un sólo valor las medidas de precisión y *recall* es comúnmente conocida como F-*Measure*. Ayuda a poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones (Chicco & Jurman, 2020).
F1 se calcula haciendo la media armónica entre la precisión y la exhaustividad:

$$f1 = 2 * \frac{precisión * recall}{precisión + recall} \qquad (4)$$

La siguiente tabla define los rangos de dicha métrica y su interpretación según (Powers, 2011).

Tabla 1. *Interpretación de valores F-Measure.*

F-Measure	Interpretación
< 0.5	Malo
0.5 – 0.8	Aceptable
0.8 – 0.9	Bueno
> 0.9	Muy bueno

2.3.6 AUC (área bajo la curva (ROC))

Esta métrica toma valores entre 0 y 1 utilizando probabilidades para decirnos que tan bien el modelo separa las clases. Comúnmente, cuanto mayor es la puntuación AUC, mejor es el rendimiento de un clasificador para una tarea de clasificación dada.

La siguiente tabla define los rangos de dicha métrica y su interpretación según (Powers, 2011).

Tabla 2. *Interpretación de valores ROC Area*.

ROC Area	Interpretación
< 0.5	sin capacidad discriminatoria diagnóstica
0.5 - 0.6	Malo
0.6 - 0.75	Regular
0.75 - 0.9	Bueno
0.9 - 0.97	Muy bueno
0.97 - 1	Excelente (valor diagnóstico perfecto)

2.3.7 Coeficiente kappa

Kappa es una medida de concordancia que se utiliza para evaluar la consistencia entre dos observadores o *raters* en la evaluación de una misma población. La concordancia se refiere a la medida en que dos observadores están de acuerdo en sus evaluaciones. Por consiguiente, existe una relación directa entre *kappa* y el nivel de concordancia.

En la siguiente tabla se definen los rangos de dicha métrica y su interpretación según (Foody, 2020).

Tabla 3. *Correlación entre Kappa y la fuerza de concordancia*.

Kappa	Fuerza de concordancia
< 0.20	Pobre
0.21 – 0.40	Débil
0.41 – 0.60	Moderada
0.61 – 0.80	Buena
0.81 – 1.00	Muy buena

2.4 Equidad en Aprendizaje Automático

Fairness es un concepto complejo que depende del contexto y la cultura, es un aspecto clave que guía nuestras interacciones y decisiones y se ve influenciado por factores sociales y personales. De ahí que sea un reto llegar a una sola definición, ya que puede variar de persona a persona o de situación en situación, en esencia, podemos decir que involucra nociones de igualdad, justicia y equidad, minimizando las desigualdades e incrementando la igualdad de oportunidades. En 2018, el Profesor Arvind Narayanan de la Universidad de Princeton, impartió un tutorial durante la conferencia ACM FAccT (*Conference on Fairness, Accountability, and Transparency*), en el cual brindó 21 definiciones de *fairness*, lo cual evidencia la complejidad del concepto, en este caso haciendo más enfocado en el punto de vista de las ciencias de la computación y de la IA.

El incremento del uso de sistemas de IA en procesos (semi-) automáticos de toma de decisiones ha generado preocupación sobre los patrones de decisión que pudieran llegar a ser dañinos o discriminatorios en contextos como la atención médica (Ledford, 2019; Obermeyer et al., 2019), la educación (Makhlouf et al., 2021) y la contratación (L. Chen et al., 2018; Raghavan et al., 2020). Estas decisiones discriminatorias diferencian a las personas basándose en características (legalmente) protegidas, por ejemplo, género, edad, origen étnico, entre otras. Dicha discriminación puede ser directa o indirecta, la primera significa que los procesos de decisión hacen referencia explícita a una característica protegida, mientras que la segunda combina dos efectos llamados discriminación por poder o impacto desigual.

Las investigaciones sobre *fairness* en algoritmos de aprendizaje automático tomaron auge en los últimos cuatro años. En 2018, IBM introdujo “AI Fairness 360”, una librería de Python con diversos algoritmos para reducir el sesgo algorítmico de un programa, aumentando así su equidad [referencia a AIF360]. Facebook anuncio también en 2018 que hacía uso de una herramienta llamada “Fairness Flow” que detectaba sesgos en su IA, aunque no se tienen muchos detalles sobre cómo funciona la misma. Google también hizo público un conjunto de herramientas para estudiar los efectos de *fairness* a largo plazo y Microsoft lanzó *fairlearn*, que en sus inicios fue un paquete de Python que se utilizó para la publicación de un artículo científico y posteriormente se convirtió en un proyecto de código abierto para ayudar a los investigadores a mejorar la equidad de los sistemas de Inteligencia Artificial (Bird et al., 2020). Este paquete, que actualmente se puede utilizar en forma de *dashboard*, será utilizado en este trabajo para medir el posible sesgo en diferentes bases de datos.

Algunos ejemplos de la vida real en los cuales puede manifestarse el ya mencionado sesgo son los siguientes:

- **Procesos de contratación:** Las herramientas basadas en IA a menudo preseleccionan y/o seleccionan a los solicitantes de empleo en los procesos de contratación. Sin embargo, si los datos utilizados para desarrollar estos modelos reflejan patrones de contratación sesgados, el sistema inadvertidamente puede estar marginando un grupo. Por ejemplo, si se utilizan datos de una empresa o industria donde históricamente han predominado los empleados varones, el modelo de IA puede aprender a favorecer a los candidatos masculinos, perpetuando los sesgos de género. Estas prácticas de contratación sesgadas obstaculizan la diversidad y la inclusión, restringiendo la igualdad de oportunidades.

- **Algoritmos de préstamos:** En el sector financiero se emplean algoritmos de IA para evaluar la solvencia y determinar las aprobaciones de préstamos. Sin embargo, si estos modelos se construyen utilizando datos históricos parciales de préstamos, pueden discriminar inadvertidamente a las comunidades desfavorecidas. Por ejemplo, si los datos indican que a ciertos grupos minoritarios se les han negado préstamos injustamente, el modelo de IA puede adoptar este comportamiento discriminatorio, perpetuando el ciclo de exclusión financiera. Este sesgo de equidad profundiza las disparidades socioeconómicas y limita el acceso a los recursos de las comunidades marginadas.

- **Disparidad en las sentencias:** Los modelos de IA se utilizan cada vez más en el sistema de justicia penal para ayudar en las decisiones de sentencia. Sin embargo, los estudios han demostrado que los algoritmos basados en IA pueden perpetuar los prejuicios raciales. Por ejemplo, un modelo de IA puede, sin saberlo, estar entrenado en datos históricos que criminalizan desproporcionadamente a ciertos grupos raciales. En consecuencia, el modelo puede recomendar sentencias más duras para personas de estos grupos, exacerbando las disparidades existentes. Este sesgo de equidad puede perpetuar la injusticia sistémica y tener consecuencias perjudiciales para las comunidades afectadas.

Además de estos ejemplos, existen muchos otros casos en los que el sesgo puede estar presente, de ahí que sea importante hacer uso de las métricas de *fairness* o equidad, las cuales nos permiten evaluar si los modelos o algoritmos de IA producen resultados que puedan considerarse justos y equitativos para diferentes grupos demográficos o atributos sensibles. Estas métricas analizan la distribución de resultados entre grupos, identifican posibles sesgos

y disparidades, y ayudan a diseñar y ajustar sistemas para promover la equidad y la imparcialidad en sus decisiones. A continuación, se explican las métricas más utilizadas en la literatura.

2.4.1 Métricas de Fairness

Como se mencionó anteriormente, en un modelo de aprendizaje automático, el sesgo se puede ver manifestado como una preferencia o prejuicio hacia una clase en específico, lo que puede entorpecer el aprendizaje y comprometer el rendimiento del modelo. Entre las métricas más utilizadas en la literatura podemos encontrar las siguientes:

Paridad Demográfica (*Demographic Parity):* La paridad demográfica es una métrica que busca garantizar que las predicciones de un modelo de Aprendizaje Automático no dependan de la pertenencia a un grupo sensible. Es decir que establece que la proporción de cada segmento de una clase protegida (por ejemplo, el género) debe recibir un resultado positivo en iguales proporciones. Por ejemplo, en el contexto de selección de currículums, significaría que la proporción de solicitantes seleccionados para una entrevista de trabajo deberían ser igual en todos los grupos (Rosenblatt & Witter, 2023).

Igualdad de oportunidades (*Equal Opportunity*): La métrica igualdad de oportunidades busca asegurar que la tasa de verdaderos positivos sea la misma para todos los grupos demográficos o subgrupos definidos por un atributo sensible. Es decir, que se centra en reducir las disparidades en la identificación correcta de casos relevantes (verdaderos positivos) entre diferentes grupos demográficos. Por ejemplo, en un caso en el que el resultado positivo representa la aprobación de un préstamo, la métrica de Igualdad de Oportunidades evaluaría si el modelo tiene tasas de verdaderos positivos similares para diferentes grupos demográficos. Si la tasa de aprobación de préstamos es mayor para un grupo demográfico en comparación con otro, sugiere una posible injusticia en las predicciones del modelo, lo que indica la necesidad de realizar ajustes para garantizar la igualdad de oportunidades para todos los grupos (Shen et al., 2022).

Equidad de probabilidades (*Equalized Odds* o Probabilidades igualadas): El objetivo de la métrica de equidad de probabilidades es garantizar que un modelo de aprendizaje automático funcione igualmente bien para diferentes grupos. Es más estricta que la paridad demográfica porque requiere que las predicciones del modelo de aprendizaje automático no sólo sean

independientes de la pertenencia a un grupo sensible, sino que los grupos tengan las mismas tasas de falsos positivos y tasas de verdaderos positivos (Tang & Zhang, 2022).

Esta distinción es importante porque un modelo podría lograr la paridad demográfica (es decir, sus predicciones podrían ser independientes de la pertenencia a un grupo sensible), pero aun así generar más predicciones falsas positivas para un grupo frente a otros. A través del uso de esta métrica se garantiza que el modelo sea equitativo tanto en la identificación de casos positivos como en la minimización de errores, independientemente de los grupos a los que pertenezcan las personas. Esto significa que el modelo debe ser igualmente preciso en detectar casos positivos y en evitar falsas alarmas para todos los grupos, sin importar sus características personales.

[1]Supongamos que estamos construyendo un modelo crediticio basado únicamente en los "ingresos". El modelo tendrá como objetivo conocer los ingresos típicos de quienes pueden pagar su préstamo en su totalidad y diferenciarlos de otros que tienden a incumplir.

Eso significa que simplemente se debe establecer un umbral de ingresos en el conjunto de entrenamiento para decidir quién recibirá un préstamo en el futuro (esta es la línea de puntos en la Figura 3). Aquellos que estén por encima del umbral obtendrán el préstamo (predicciones positivas). Los que están por debajo del umbral son los que no obtendrán un préstamo (predicciones negativas).

Los datos históricos que se utilizan son los siguientes:

- Préstamo pagado: Quienes pagaron la totalidad de sus préstamos.
- Incumplidos: Aquellos que no pudieron pagar su préstamo y entraron en incumplimiento.

[1] https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2

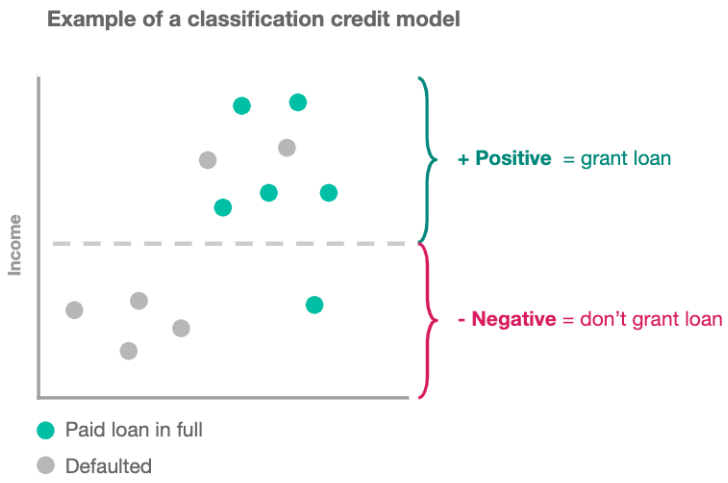


Figura 3. *Ejemplo de un modelo de clasificación crediticio*[2].

Como se explicó anteriormente, la paridad demográfica establece que la proporción de cada segmento de una clase protegida debe recibir el resultado positivo en proporciones iguales. Un resultado positivo es la decisión preferida, en este caso obtener el préstamo, por tanto, en aras de garantizar esta igualdad en las proporciones, podemos decidir utilizar diferentes niveles de requisitos para cada grupo de modo que el porcentaje de personas que obtienen un préstamo en el Grupo A sea igual al porcentaje de personas que obtienen un préstamo en el Grupo B, como se puede apreciar en la Figura 4.

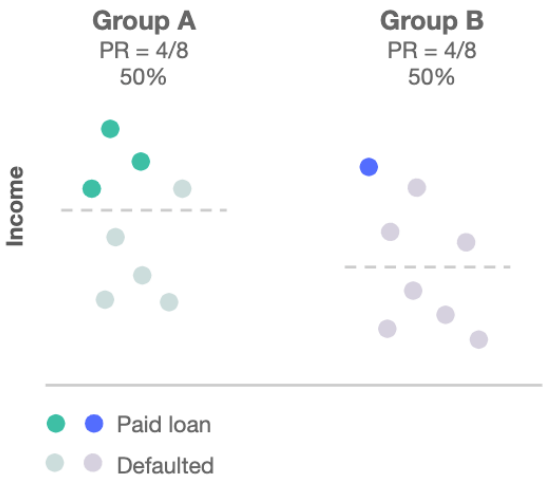


Figura 4. *Paridad demográfica.*

Después de analizar el ejemplo anterior, se puede decir que la paridad demográfica como definición de equidad es más apropiada cuando somos conscientes de que los sesgos históricos pueden haber afectado la calidad de nuestros datos y contamos con un plan para apoyar al grupo

[2] Figuras 3, 4, 5 y 6 tomadas de https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2

desfavorecido y evitar el refuerzo de dichos sesgos (por ejemplo, establecer políticas que penalicen el comportamiento no inclusivo en una junta directiva).

En el caso de la métrica igualdad de oportunidades, establece que cada grupo debe obtener el resultado positivo en proporciones iguales, asumiendo que las personas de este grupo califican para ello. Si miramos nuestro ejemplo en la Figura 5a, podemos ver que el porcentaje de positivos que se predijeron con precisión es del 50% para ambos grupos.

Si miramos la Figura 5b, podemos ver que la Tasa de Verdaderos Positivos es la misma para ambos grupos como lo requiere la Igualdad de Oportunidades. Sin embargo, si prestamos atención al Grupo B, podemos notar que se introdujeron muchos falsos positivos. Los falsos positivos en este caso son aquellos que obtienen un préstamo cuando en realidad es probable que incumplan. Eso significa que estaríamos dañando el puntaje crediticio a mayor escala en el Grupo B, lo que resulta en un impacto dispar.

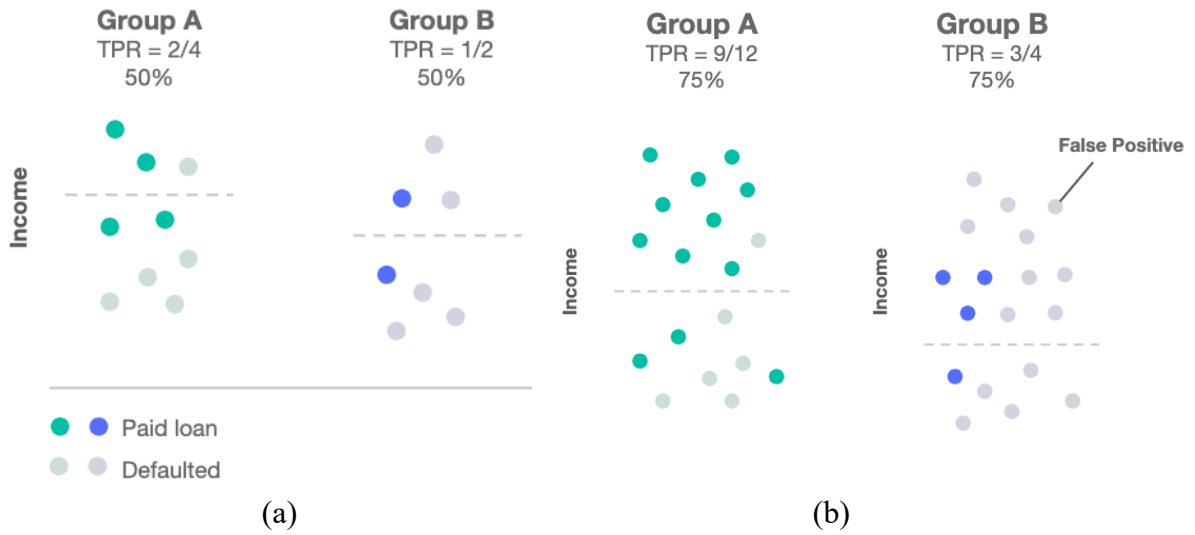


Figura 5. *Uso de diferentes umbrales de ingresos para lograr la igualdad de oportunidades.*

Después de analizar estos resultados, podemos decir que es oportuno utilizar la métrica igualdad de oportunidades cuando:

- Hay un fuerte énfasis en predecir correctamente el resultado positivo (por ejemplo: debemos ser muy buenos detectando una transacción fraudulenta).
- Introducir Falsos Positivos no es costoso para el usuario ni para la empresa (por ejemplo: notificar erróneamente a un cliente sobre una actividad fraudulenta no será necesariamente costoso para el cliente ni para el banco que envía la alerta)
- La variable objetivo no se considera subjetiva (por ejemplo, etiquetar quién es un "buen" empleado genera sesgo y, por lo tanto, es muy subjetivo).

En el caso de la métrica equidad de probabilidades, debemos recordar que es la más restrictiva ya que establece que el modelo debe identificar correctamente el resultado positivo en iguales proporciones entre los grupos (tasa de verdaderos positivos, similar a la métrica anterior), y también tener igual proporción en la tasa de falsos positivos.

En la Figura 6 podemos ver que se identifica correctamente la respuesta positiva a un 50% para ambos grupos, y se logra además una tasa de falsos positivos del 25%. Hay que tener en cuenta que tratar de lograr tasas similares en ambos grupos puede llevar a una caída en la ganancia, ya que el rendimiento del modelo puede verse comprometido al no ser capaz de optimizar la precisión en el grupo mayoritario.

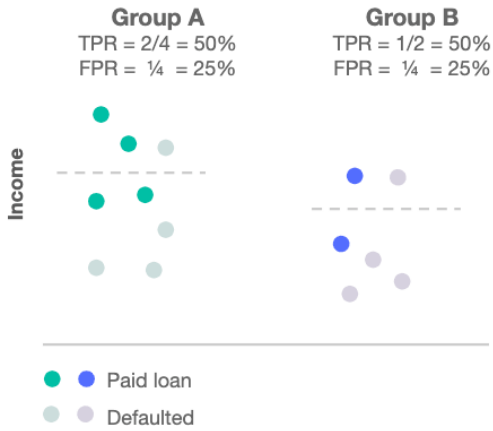


Figura 6. *Uso de diferentes umbrales para lograr la equidad de probabilidades.*

Los modelos de crédito son un excelente ejemplo del uso de equidad de probabilidades, ya que se genera valor comercial al aceptar tantos clientes que puedan pagar un préstamo en su totalidad. Sin embargo, también se tiene en cuenta que se debe minimizar los "falsos positivos", aquellos a los que se les concede un préstamo y no pueden devolverlo. Los falsos positivos pueden dañar las calificaciones crediticias de los clientes y, en consecuencia, sus oportunidades financieras en el futuro. También supondría un alto costo para el banco que genera el préstamo.

Si al evaluar un modelo utilizando las métricas anteriormente mencionadas, se detecta que el mismo está siendo injusto, se puede proceder a la mitigación, es decir, a tomar acciones que permitan reducir o eliminar sesgos y desigualdades. Existen algoritmos que permiten realizar este proceso de diferentes formas, a continuación se mencionan algunos de ellos.

Algoritmos de mitigación:

- *Grid Search*: método que busca el mejor modelo que cumpla con ciertas restricciones de *fairness* mientras optimiza el rendimiento. Es útil para encontrar un balance optimo entre equidad y precisión.
- *Exponentiated Gradient Reduction*: Un algoritmo que ajusta el modelo iterativamente para reducir las disparidades entre grupos al tiempo que maximiza la precisión del modelo.
- *Threshold Optimizer:* Ajusta los umbrales de decisión del modelo después del entrenamiento para mejorar la equidad en las predicciones.

3 Bases de Datos

En este estudio, se seleccionaron tres conjuntos de datos que comparten la particularidad de incluir al menos un atributo sensible o de naturaleza potencialmente discriminatoria, como por ejemplo, sexo, género, nacionalidad y raza.

3.1 Reclamo de seguro de coches

El conjunto de datos fue obtenido de Kaggle y presenta información sobre 10000 individuos que han adquirido pólizas de seguro de automóviles, indicando si han presentado reclamaciones asociadas a accidentes u otras circunstancias. Este conjunto de datos consta de 18 atributos (6 numéricos, 5 categóricos y 7 binarios), además de uno adicional que representa la clase (OUTCOME). En la Tabla 4 se presenta un resumen de dichos atributos.

Tabla 4. *Definición de las variables del dataset Reclamo de Seguros de Coches.*

Descripción corta	**Tipo**	**Nombre**	**Definición y codificación**
Identificador único	Numérico	ID	Valores entre 101 y 1 millón
Rango de Edades	Categórico	AGE	Rangos de edades (26-39, 40-64, 16-25, 65+)
Género	Binario	GENDER	Female (femenino) y Male (masculino)
Raza	Binario	RACE	Majority (mayoría), minority (minoría)
Años de experiencia manejando	Categórico	DRIVING_EXPERIENCE	Rangos de años de experiencia (0-9y, 10-19y, 20-29y, 30y+)
Nivel de educación	Categórico	EDUCATION	high school, university (universidad), none (ninguno)
Clase según ingresos	Categórico	INCOME	upper class (clase baja), middle class (clase media), poverty (pobreza), working class (clase trabajadora)

Puntuación de créditos	Numérico	CREDIT_SCORE	Valores entre 0 y 1
Dueño del vehículo	Binario	VEHICLE_OWNERSHIP	0 = no dueño, 1 = dueño
Vehículos del 2015	Binario	VEHICLE_YEAR	before 2015 (antes 2015), after 2015 (después 2015)
Casado(a)	Binario	MARRIED	1 = casado (a), 0 = no casado (a)
Tiene Hijos	Binario	CHILDREN	1 = tiene hijos, 0 = no tiene hijos
Código Postal	Numérico	POSTAL_CODE	Código de 5 dígitos
Kilometraje anual	Numérico	ANNUAL_MILEAGE	Valores entre 2 mil y 22 mil
Tipo de Vehículo	Binario	VEHICLE_TYPE	Sedan, sports car (coche deportivo)
Cantidad de violaciones de velocidad	Numérico	SPEEDING_VIOLATIONS	Valores entre 0 y 22
Conducir bajo la influencia de alcohol o drogas	Categórico	DUIS	Valores entre 0 y 6
Cantidad de accidentes pasados	Numérico	PAST_ACCIDENTS	Valores entre 0 y 15
Salida (Reclamo del seguro)	Binario	OUTCOME	1 = reclamo del seguro, 0 = no reclamo del seguro

3.2 Abandono o éxito académico

El conjunto de datos fue generado como parte de un proyecto destinado a combatir la deserción y el fracaso en la educación superior en el Instituto Politécnico de Portalegre. Se emplearon técnicas de aprendizaje automático con el objetivo de detectar estudiantes en situación de riesgo en las primeras etapas de su trayectoria académica, permitiendo la implementación de estrategias de apoyo tempranas (Realinho et al., 2021).

Los datos abarcan detalles disponibles en el momento de la inscripción del estudiante, como historial académico, información demográfica y factores socioeconómicos. Contiene 37 atributos (9 categóricos, 19 numéricos, 9 binarios), siendo uno de ellos la clase y consta de 3630 instancias. En la Tabla 5 se presentan las características principales de los atributos.

Tabla 5. *Definición de las variables del dataset Abandono o éxito académico.*

Descripción corta	**Tipo**	**Nombre**	**Definición y codificación**
Estado civil	Categórico	Marital status	1 = soltero(a), 2 = casado(a), 3 = viudo(a), 4 = divorciado(a), 5 = unión de hecho, 6 = legalmente separado(a)
Modo de Aplicación	Categórico	Application mode	1 = 1ª fase - contingente general, 2 = Decreto nº 612/93, 5 = 1ª fase - contingente especial (Isla de las Azores), 7 = Titulares de otros cursos superiores, 10 = Ordenanza nº 854-B/99, 15 = Estudiante internacional (licenciatura), 16 = 1ª fase - contingente especial (Isla de Madeira), 17 = 2ª fase - contingente general, 18 = 3ª fase - contingente general, 26 = Ordenanza nº 533-A/99, punto b2) (Plan diferente), 27 = Ordenanza nº 533-A/99, punto b3 (Otra institución), 39 = Mayores de 23 años, 42 = Traslado, 43 = Cambio de curso, 44 = Diplomados de especialización tecnológica, 51 = Cambio de institución/curso, 53 = Diplomados de ciclo corto 57 = Cambio de institución/curso (Internacional)
Orden de solicitud	Numérico	Application order	Orden de solicitud (entre 0 - primera opción; y 9 última opción)
Curso	Categórico	Course	33 = Tecnologías de Producción de Biocombustibles,

			171 = Animación y Diseño Multimedia, 8014 = Servicio Social (presencial nocturno), 9003 = Agronomía, 9070 = Diseño de Comunicación, 9085 = Enfermería Veterinaria, 9119 = Ingeniería Informática, 9130 = Equinocultura, 9147 = Gestión, 9238 = Servicio Social, 9254 = Turismo, 9500 = Enfermería, 9556 = Higiene Bucodental, 9670 = Gestión de Publicidad y Marketing, 9773 = Periodismo y Comunicación, 9853 = Educación Básica, 9991 = Gestión (presencial nocturno)
Asistencia diurna/noche	Binario	Daytime/evening attendance	1 = diurno 0 = nocturno
Titulación previa	Categórico	Previous qualification	1 = Educación secundaria, 2 = Educación superior – licenciatura, 3 = Educación superior – grado, 4 = Educación superior – máster, 5 = Educación superior – doctorado, 6 = Frecuencia de la educación superior, 9 = 12º año de escolarización - no completado, 10 = 11º año de escolarización - no completado, 12 = Otros - 11º año de escolarización, 14 = 10º año de escolarización, 15 = 10º año de escolarización - no completado, 19 = Educación básica 3er ciclo (9º/10º/11º año) o equiv., 38 = Enseñanza básica 2º ciclo (6º/7º/8º año) o equiv., 39 = Curso de especialización tecnológica, 40 = Enseñanza superior - licenciatura (1º ciclo),

			42 = Curso técnico superior profesional, 43 = Enseñanza superior - máster (2º ciclo)
Titulación previa (nota)	Numérico	Previous qualification (grade)	Nota de la titulación anterior (entre 0 y 200)
Nacionalidad	Categórico	Nacionality	1 = portugués, 2 = alemán, 6 = español, 11 = italiano, 13 = neerlandés, 14 = inglés, 17 = lituano, 21 = angoleño, 22 = caboverdiano, 24 = guineano, 25 = mozambiqueño, 26 = santomerano, 32 = turco, 41 = brasileño, 62 = rumano, 100 = moldavo (República de), 101 = mexicano, 103 = ucraniano, 105 = ruso, 108 = cubano, 109 = colombiano.
Título de la madre	Categórico	Mother's qualification	1 = Enseñanza Secundaria - 12º Año de Escolaridad o Equiv. 2 = Enseñanza Superior – Licenciatura, 3 = Enseñanza Superior – Grado, 4 = Enseñanza Superior – Máster, 5 = Enseñanza Superior – Doctorado, 6 = Frecuencia de Enseñanza Superior,

Título del padre	Categórico	Father's qualification	9 = 12º Año de Escolaridad - No Completado, 10 = 11º Año de Escolaridad - No Completado, 11 = 7º Año (Antiguo), 12 = Otro - 11º Año de Escolaridad, 14 = 10º Año de Escolaridad, 18 = Curso general de comercio, 19 = Enseñanza Básica 3er Ciclo (9º/10º/11º Año) o Equiv. 22 = Curso técnico-profesional, 26 = 7º año de escolarización, 27 = 2º ciclo de bachillerato general, 29 = 9º año de escolarización - No completado, 30 = 8º año de escolarización, 34 = Desconocido, 35 = No sabe leer ni escribir, 36 = Sabe leer sin tener 4º año de escolarización, 37 = Educación básica 1er ciclo (4º/5º año) o equiv. 38 = Enseñanza básica 2º ciclo (6º/7º/8º año) o equiv. 39 = Curso de especialización tecnológica, 40 = Enseñanza superior - Licenciatura (1º ciclo), 41 = Curso de estudios superiores especializados, 42 = Curso técnico superior profesional, 43 = Enseñanza superior - Máster (2º ciclo), 44 = Enseñanza superior - Doctorado (3º ciclo)

Profesión de la madre	Categórico	Mother's occupation	0 = Estudiante, 1 = Representantes del Poder Legislativo y Órganos Ejecutivos, Consejeros, Directores y Gerentes Ejecutivos, 2 = Especialistas en Actividades Intelectuales y Científicas, 3 = Técnicos y Profesiones de Grado Medio, 4 = Personal Administrativo, 5 = Trabajadores y Vendedores de Servicios Personales, Seguridad y Vigilancia, 6 = Agricultores y Trabajadores Cualificados en Agricultura, Pesca y Silvicultura, 7 = Trabajadores Cualificados en la Industria, Construcción y Artesanos, 8 = Operadores de Instalaciones y Máquinas y Montadores, 9 = Trabajadores no Cualificados, 10 = Profesiones de las Fuerzas Armadas, 90 = Otra Situación, 99 = (en blanco), 122 = Profesionales de la Salud, 123 = Profesores, 125 = Especialistas en Tecnologías de la Información y la Comunicación (TIC), 131 = Técnicos y Profesiones de Grado Medio de Ciencias e Ingeniería, 132 = Técnicos y Profesionales de nivel intermedio de sanidad, 134 = Técnicos de nivel intermedio de servicios jurídicos, sociales, deportivos, culturales y similares, 141 = Oficinistas, secretarios en general y operadores de proceso de datos, 143 = Operadores de datos, contables, estadísticos, de servicios financieros y relacionados con registros, 144 = Otro personal de apoyo administrativo, 151 = Trabajadores de servicios personales, 152 = Vendedores, 153 = Trabajadores de cuidados personales y asimilados, 171 = Trabajadores cualificados de la construcción y asimilados, excepto electricistas, 173 = Trabajadores cualificados de la imprenta, fabricación de instrumentos de precisión, joyeros, artesanos y asimilados, 175 = Trabajadores de la elaboración de alimentos, la carpintería, la confección y otras industrias y oficios, 191 = Trabajadores de la limpieza,
Profesión del padre	Categórico	Father's occupation	

			192 = Trabajadores no cualificados de la agricultura, la ganadería, la pesca y la silvicultura, 193 = Trabajadores no cualificados de la industria extractiva, la construcción, la industria manufacturera y el transporte, 194 = Ayudantes de preparación de comidas
Nota de admisión	Numérico	Admission grade	Nota de admisión (entre 0 y 200)
Desplazados	Binario	Displaced	0 = No, 1 = Si
Necesidades educativas especiales	Binario	Educational special needs	0 = No, 1 = Si
Deudor	Binario	Debtor	0 = No, 1 = Si
Tasas de matrícula al día	Binario	Tuition fees up to date	0 = No, 1 = Si
Género	Binario	Gender	1 = Masculino, 0 = Femenino
Becario	Binario	Scholarship holder	0 = No, 1 = Si
Edad en la inscripción	Numérico	Age at enrollment	Edad del estudiante en el momento de la inscripción
Internacional	Binario	International	0 = No, 1 = Si
Unidades curriculares 1er sem (acreditadas)	Numérico	Curricular units 1st sem (credited)	Número de unidades curriculares acreditadas en el 1er semestre. Valores entre 0 y 20
Unidades curriculares 1er sem (matriculados)	Numérico	Curricular units 1st sem (enrolled)	Número de unidades curriculares matriculadas en el 1er semestre. Valores entre 0 y 26
Unidades curriculares 1er sem (evaluaciones)	Numérico	Curricular units 1st sem (evaluations)	Número de evaluaciones a unidades curriculares en el 1er semestre. Valores entre 0 y 36

Unidades curriculares 1er sem (aprobado)	Numérico	Curricular units 1st sem (approved)	Número de unidades curriculares aprobadas en el 1er semestre. Valores entre 0 y 26
Unidades curriculares 1er sem (grado)	Numérico	Curricular units 1st sem (grade)	Promedio de notas en el 1er semestre (entre 0 y 20)
Unidades curriculares 1er sem (sin evaluaciones)	Numérico	Curricular units 1st sem (without evaluations)	Número de unidades curriculares sin evaluaciones en el 1er semestre. Valores entre 0 y 12
Unidades curriculares 2do sem (acreditadas)	Numérico	Curricular units 2nd sem (credited)	Número de unidades curriculares acreditadas en el 2do semestre. Valores entre 0 y 19
Unidades curriculares 2do sem (matriculados)	Numérico	Curricular units 2nd sem (enrolled)	Número de unidades curriculares matriculadas en el 2do semestre. Valores entre 0 y 23
Unidades curriculares 2do sem (evaluaciones)	Numérico	Curricular units 2nd sem (evaluations)	Número de evaluaciones a unidades curriculares en el 2do semestre. Valores entre 0 y 33
Unidades curriculares 2do sem (aprobado)	Numérico	Curricular units 2nd sem (approved)	Número de unidades curriculares aprobadas en el 2do semestre. Valores entre 0 y 20
Unidades curriculares 2do sem (grado)	Numérico	Curricular units 2nd sem (grade)	Nota media en el 2º semestre (entre 0 y 20)
Unidades curriculares 2do sem (sin evaluaciones)	Numérico	Curricular units 2nd sem (without evaluations)	Número de unidades curriculares sin evaluaciones en el 2do semestre. Valores entre 0 y 12
Tasa de desempleo	Numérico	Unemployment rate	Tasa de desempleo (%)
Tasa de inflación	Numérico	Inflation rate	Tasa de inflación (%)

PIB	Numérico	GDP	Producto Interno Bruto (PIB)
Éxito o Fracaso escolar	Binario	Target	Graduate (Graduado), Dropout (Abandono)

3.3 Admisión en la escuela de derecho

Este *dataset* representa un estudio realizado en los Estados Unidos por el *Law School Admission Council* (LSAC) sobre la aprobación del examen de acceso a la abogacía. Los datos fueron recopilados de la promoción que inició la carrera de Derecho en el otoño de 1991 e incluye información proporcionada por los estudiantes, sus facultades de derecho, y las juntas estatales de examinadores del colegio de abogados a lo largo de un período de cinco años (Wightman, 1998). Contiene un total de 12 atributos, incluido el de la clase (6 numéricos, 4 binarios, 2 categóricos) y un total de 20796 instancias (Ver Tabla 6)

Tabla 6. *Definición de las variables del dataset Admisión en la escuela de derecho.*

Descripción corta	**Tipo**	**Nombre**	**Definición y codificación**
Decil primer curso	Numérico	decile1b	Decil del alumno en la escuela dadas sus notas en el primer curso (Valores entre 1.0 y 10.0)
Decil tercer curso	Numérico	decile3	Decil del alumno en la escuela dadas sus notas en el tercer curso (Valores entre 1.0 y 10.0)
Puntuación LSAT	Numérico	lsat	Puntuación del estudiante en el LSAT (*Law School Admissions Test*). Valores entre 11.0 y 48.
Nota media	Numérico	ugpa	Nota media del estudiante. Valores entre 1.5 y 4.0
Nota media del primer año	Numérico	zfygpa	La nota media del primer año de Derecho. Valores entre -3.35 y 3.48
GPA acumulativo	Numérico	zgpa	El GPA (Promedio de notas - *Grade Point Average*) acumulativo de la facultad de Derecho. Valores entre -6.44 y 4.01

Trabajo tiempo completo	Binario	fulltime	Si el estudiante trabajará a tiempo completo o parcial (1 = tiempo completo, 2 = tiempo parcial)
Nivel de renta familiar	Categórico	fam_inc	Nivel de renta familiar del estudiante. Valores entre 1 y 5
Hombre	Binario	male	Si el estudiante es hombre o mujer (0 = mujer, 1 = hombre)
Nivel del estudiante	Categórico	tier	Nivel del estudiante. Valores entre 1 y 6
Raza	Binario	race	White (blanca), Non-White (no blanca)
Admisión a la escuela	Binario	pass_bar	1 = admitido(a), 0 = no admitido(a)

4 Herramienta para el análisis de equidad

En este capítulo se establece la metodología para desarrollar la herramienta que facilita el análisis exploratorio de datos, la construcción de los modelos de clasificación propuestos, el examen del comportamiento de los modelos en términos de equidad, y la mitigación de posibles injusticias que puedan resultar en la elección de una opción negativa en lugar de una positiva. Los pasos a seguir se resumen de la siguiente manera:

- Realizar un estudio de los datos para obtener las características de las variables, observar la correlación que existe entre ellas e identificar valores perdidos.
- Preprocesar el conjunto de datos para evitar anomalías en los mismos.
- Confeccionar los conjuntos de entrenamiento y de prueba de los clasificadores.
- Evaluar los clasificadores mediante diversas métricas.
- Elegir el clasificador óptimo teniendo en cuenta las métricas definidas para este estudio, que incluyen tanto las métricas tradicionales de los clasificadores como las métricas de *fairness*.
- Mitigar el modelo seleccionado considerando tanto las métricas tradicionales de los clasificadores como las métricas de *fairness.*

4.1 Introducción a la herramienta desarrollada

Para llevar a cabo el análisis de los conjuntos de datos, se desarrolló una herramienta (https://github.com/jcotopalacio/AplicacionFairness) utilizando Python 3.10, haciendo uso de diversas bibliotecas de Aprendizaje Automático, tales como pandas, seaborn, matplotlib, numpy, scikit-learn y xgboost. En la sección dedicada a la equidad, se utilizó *fairlearn* y la herramienta *FairnessDashboard* proveniente de la librería *raiwidgets*, y para la visualización se empleó PyQt5. La aplicación consta de ventanas que facilitan la navegación a lo largo de las distintas etapas del proceso de Descubrimiento de Conocimientos Útiles a partir de Datos (KDD por sus siglas en inglés) que serán explicados detalladamente más adelante.

En un principio, el usuario se ubica en una ventana principal donde tiene la opción de elegir uno de los conjuntos de datos previamente mencionados (Figura 7). A continuación, se muestran los atributos y el tipo de cada uno, con la posibilidad de modificarlos en caso de detectarse algún error. También se visualiza el número de instancias y se presenta un gráfico

que representa la distribución de las clases, permitiendo al usuario verificar la existencia de desequilibrios (Figura 8).

Figura 7*. Ventana principal de la Aplicación de Fairness.*

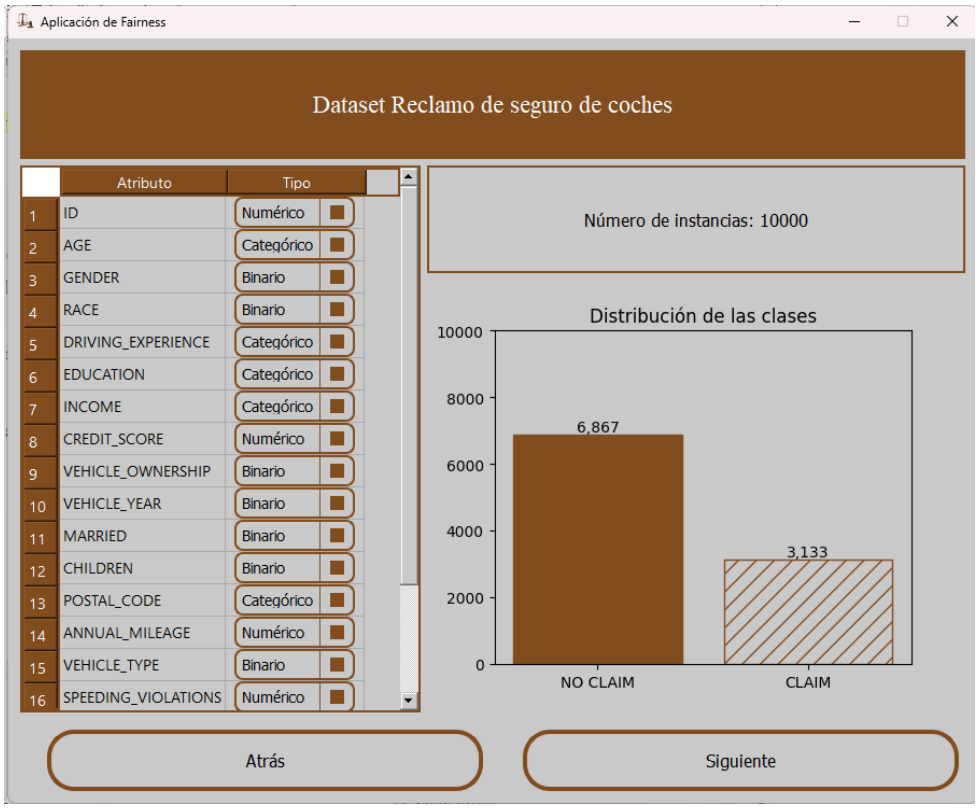


Figura 8. *Segunda ventana. Datos generales del dataset (Reclamo de seguro de coches).*

4.2 Tratamiento de valores perdidos

En las etapas iniciales del análisis de un conjunto de datos, es esencial investigar la presencia de valores perdidos, ya que numerosas técnicas en el ámbito del Aprendizaje Automático carecen de la capacidad para gestionar adecuadamente datos faltantes. Identificar y abordar esta ausencia de valores se convierte en un paso crucial, dado que la calidad y la fiabilidad de los resultados del análisis dependen en gran medida de la integridad y la completitud de los datos. La falta de información puede afectar negativamente la capacidad del modelo para generalizar patrones y realizar predicciones precisas.

En el contexto de este estudio, de los tres conjuntos de datos analizados, únicamente el relacionado con los Reclamos de Seguro de Coches exhibe la presencia de valores ausentes en dos de sus atributos específicos: CREDIT_SCORE y ANNUAL MILEAGE. Para abordar esta situación, se brinda al usuario la capacidad de elegir entre tres técnicas disponibles, las cuales se ajustan de manera diferenciada al conjunto de datos en cuestión (consultar Figura 9). Es importante destacar que la aplicación no permite avanzar al siguiente paso hasta que se hayan resuelto de manera efectiva los valores nulos en los datos, garantizando así la integridad y la validez del análisis siguiente.

Figura 9. *Tratamiento de valores perdidos. Reclamo de seguro de coches.*

4.3 Análisis gráfico de las variables

En esta sección, exploraremos el comportamiento de las variables a través de una variedad de representaciones gráficas, tales como gráficos de violín, correlogramas, gráficos de barras, diagramas de caja y tabulaciones cruzadas de múltiples variables (Figura 10). Esta aproximación proporciona una perspectiva más nítida para comprender las interrelaciones entre las variables, al mismo tiempo que facilita la identificación de valores atípicos o extremos.

Inicialmente, se examinará el desequilibrio de las clases de salida en cada uno de los conjuntos de datos. Se puede observar en las figuras 11, 12 y 13 que hay desequilibrios en todos los conjuntos de datos, siendo más pronunciado en el conjunto de datos de Admisión en la escuela de derecho.

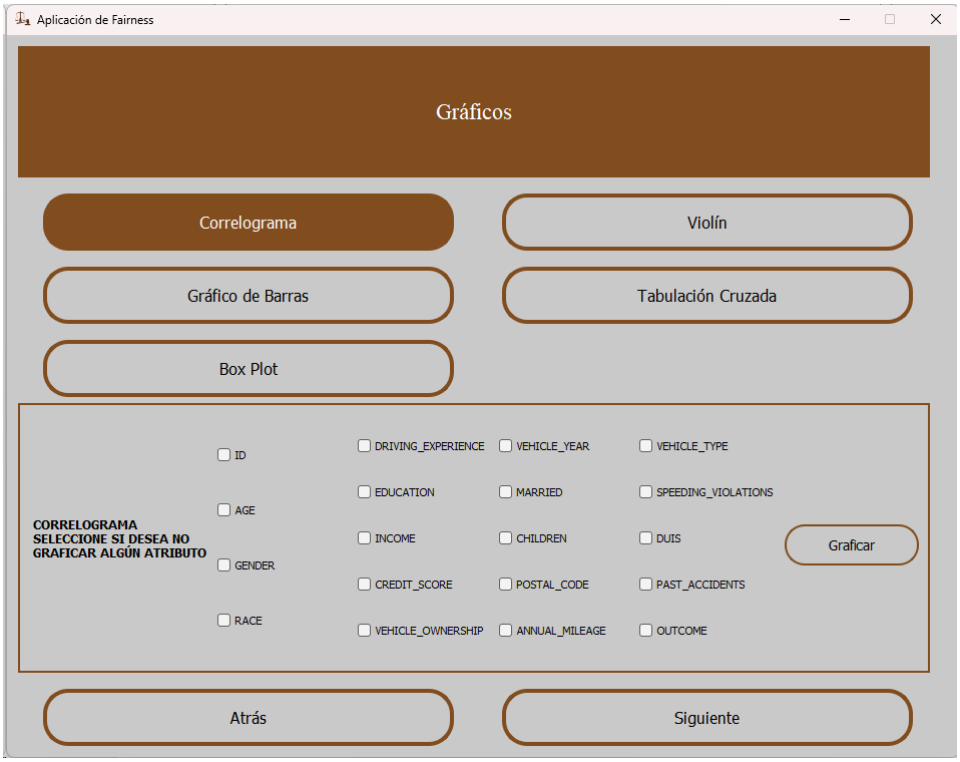


Figura 10. *Ventana de selección de gráficos.*

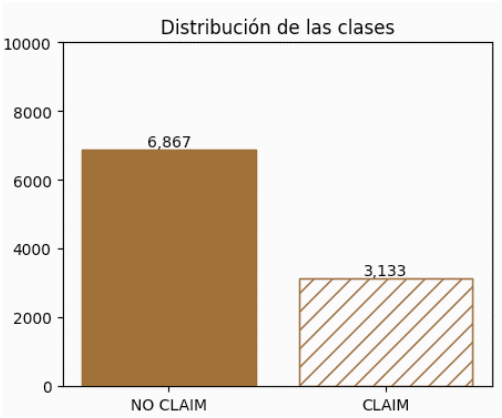


Figura 11. *Distribución de las clases de Reclamo de seguro de coches.*

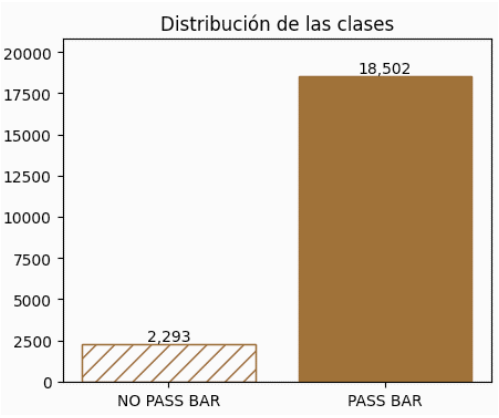


Figura 12. *Distribución de las clases de Admisión en la escuela de derecho.*

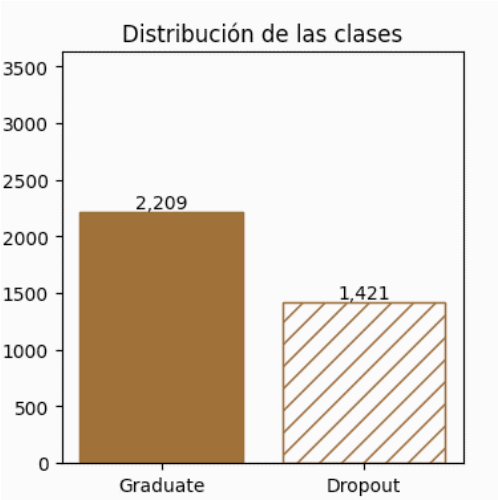


Figura 13. *Distribución de las clases de Abandono o éxito académico.*

4.3.1 Correlograma

La matriz de correlación de variables es una herramienta fundamental en el análisis de datos. Este gráfico utiliza el coeficiente de correlación, que varía entre -1 y 1. Un valor de -1 indica una correlación negativa perfecta, 1 representa una correlación positiva perfecta, y 0 señala la ausencia de correlación entre las variables. Este método nos permite evaluar la fuerza y dirección de las relaciones entre las diferentes variables, proporcionando una valiosa visión de cómo éstas interactúan en conjunto.

En la figura 14 que representa el conjunto de datos sobre el abandono o éxito académico, se observa que las variables "International" y "Nationality" están fuertemente vinculadas con un

coeficiente de correlación de 0.8, lo que sugiere que proporcionan información similar en este contexto específico.

Asimismo, se destaca una fuerte correlación positiva entre las unidades curriculares del mismo tipo en ambos semestres, como "Curricular units 1st sem (credited)" con "Curricular units 2nd sem (credited)", y "Curricular units 1st sem (enrolled)" con "Curricular units 2nd sem (enrolled)". También se observan relaciones positivas de intensidad moderada entre todas las unidades curriculares de ambos semestres, a excepción de aquellas sin evaluación, que solo muestran correlación positiva entre sí.

Adicionalmente, se evidencia una relación positiva significativa, con coeficientes por encima de 0.5, entre la cantidad de unidades curriculares aprobadas en ambos semestres y la calificación promedio correspondiente a cada semestre en relación con la variable de clasificación "Target". La correlación de 0.58 entre la nota de la titulación anterior y la nota de admisión indica un patrón consistente en el desempeño académico del estudiante, sugiriendo que aquellos con buen rendimiento en su titulación anterior tienden a obtener puntuaciones más altas en los exámenes de admisión.

Al analizar el diagrama de correlación del conjunto de datos de Admisión a la escuela de derecho (Figura 15), se destaca una correlación positiva considerable de 0.87 entre los deciles del primer curso y los del tercer curso de los estudiantes. Esta relación sugiere un patrón coherente en el rendimiento del estudiante a lo largo de ambos cursos. Además, se observa una fuerte y positiva asociación entre los deciles y la nota media del estudiante en el primer año (zfygpa), así como entre la nota media del primer año y el promedio general acumulativo de la facultad (zgpa). Estos hallazgos indican un rendimiento académico consistente y estable por parte del estudiante.

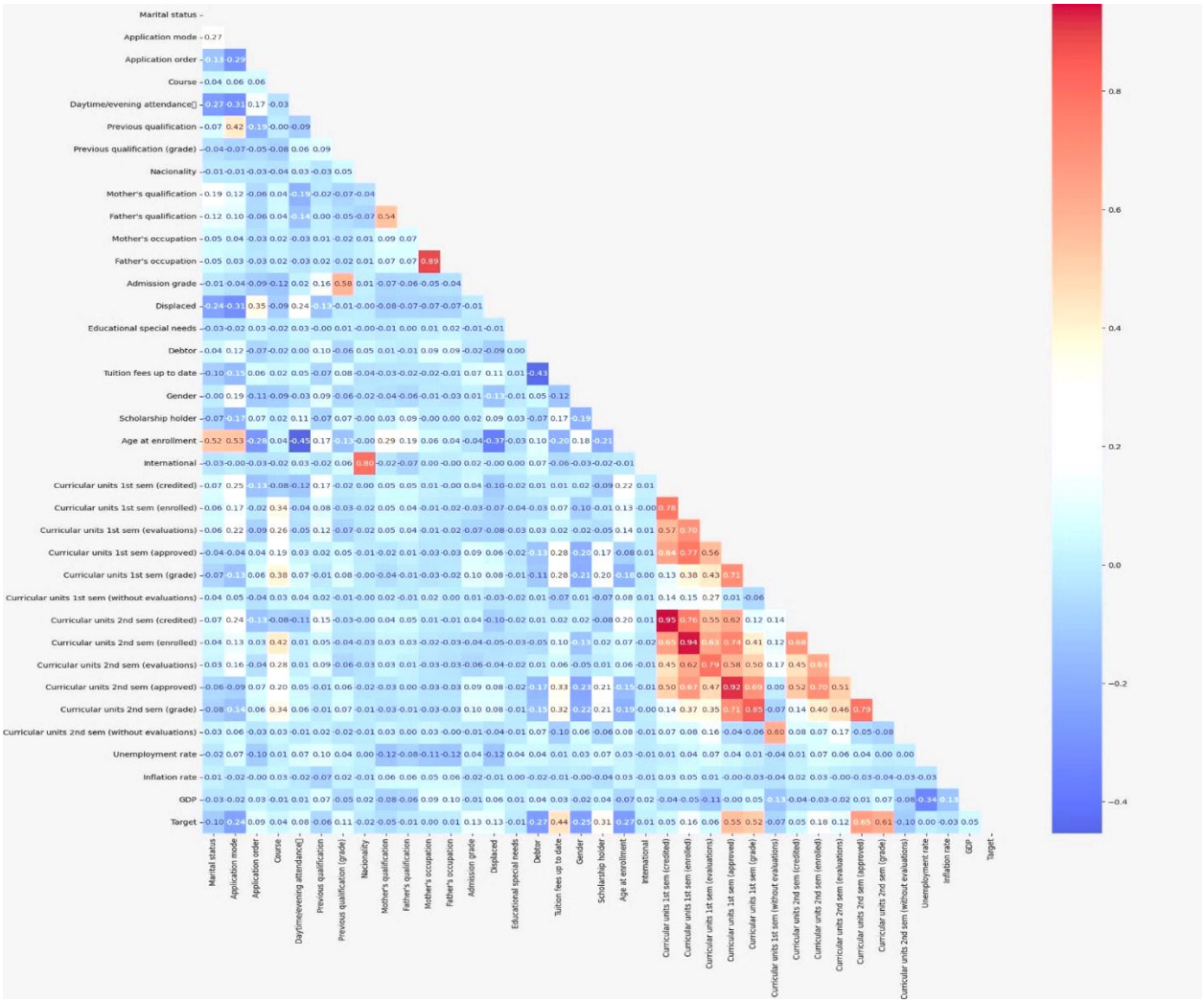


***Figura 14**. Correlograma. Abandono o éxito académico.*

En el conjunto de datos sobre Reclamo de seguro de coches, se resalta una correlación positiva notable de 0.71 entre la edad de los conductores y su experiencia al volante, como se muestra en la Figura 16. Este descubrimiento sugiere que, en términos generales, el aumento en la edad de los conductores está asociado con un incremento en su experiencia de conducción. Este vínculo podría tener implicaciones significativas al analizar reclamos de seguros para automóviles, ya que la pericia al volante podría afectar la probabilidad de eventos y la forma en que los conductores enfrentan situaciones específicas de manejo. Este fenómeno se refleja más específicamente en la relación positiva entre los Años de experiencia manejando (DRIVING_EXPERIENCE) y tanto la Cantidad de violaciones de velocidad (SPEEDING_VIOLATIONS) como la Cantidad de accidentes pasados (PAST_ACCIDENTS).

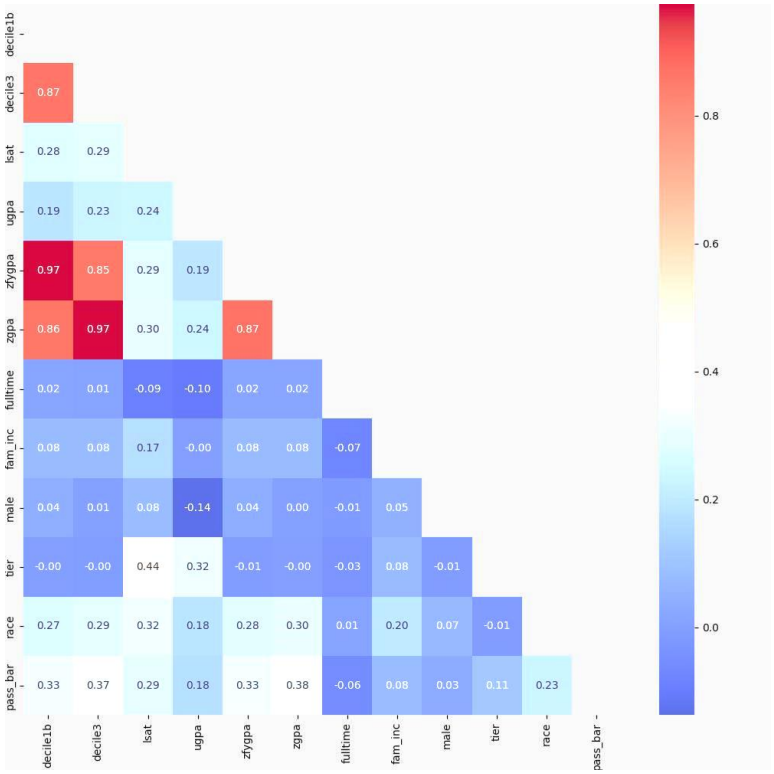


***Figura 15**. Correlograma. Admisión a la escuela de derecho.*

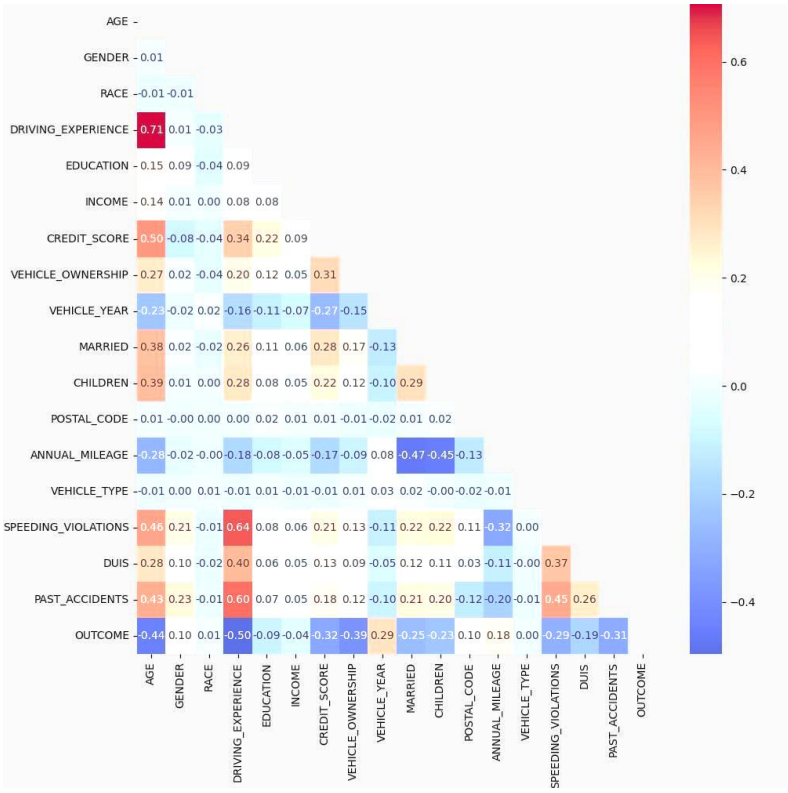


***Figura 16**. Correlograma. Reclamo de seguro de coches.*

4.3.2 Gráfico de Cajas (Box Plot)

Un diagrama de caja es una representación estandarizada que resume la distribución de datos utilizando cinco puntos clave: el valor mínimo, el primer cuartil (Q1), la mediana, el tercer cuartil (Q3) y el valor máximo. Estos gráficos ofrecen información sobre la presencia y magnitud de valores atípicos, la simetría de los datos, la concentración de valores y la presencia de sesgos, proporcionando así una visión rápida y visual de la variabilidad en un conjunto de datos.

En la Figura 17, se observa que, en el contexto del Abandono o Éxito académico, hay valores atípicos con respecto a la media para las variables de Orden de solicitud, Titulación previa (nota), Nota de admisión y Edad en la inscripción. Sin embargo, es importante destacar que todos estos valores atípicos están dentro de los límites aceptables según los rangos permitidos para dichas variables.

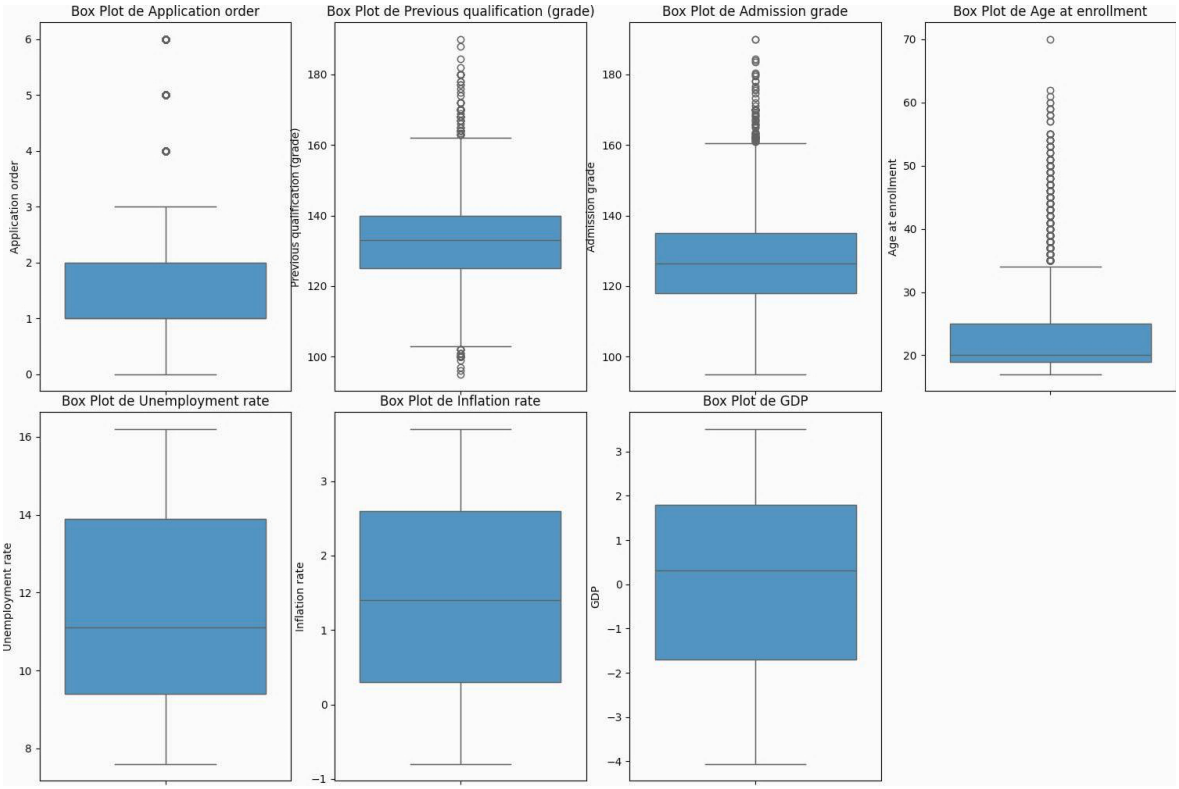


Figura 17. *Diagrama de Cajas. Abandono o éxito académico.*

Para los conjuntos de datos de Admisión a la escuela de derecho (Figura 18) y Reclamo de seguro de coches (Figura 19), no se observan valores que estén fuera de los rangos permitidos. Sin embargo, en algunas variables se pueden notar valores extremos con una incidencia reducida.

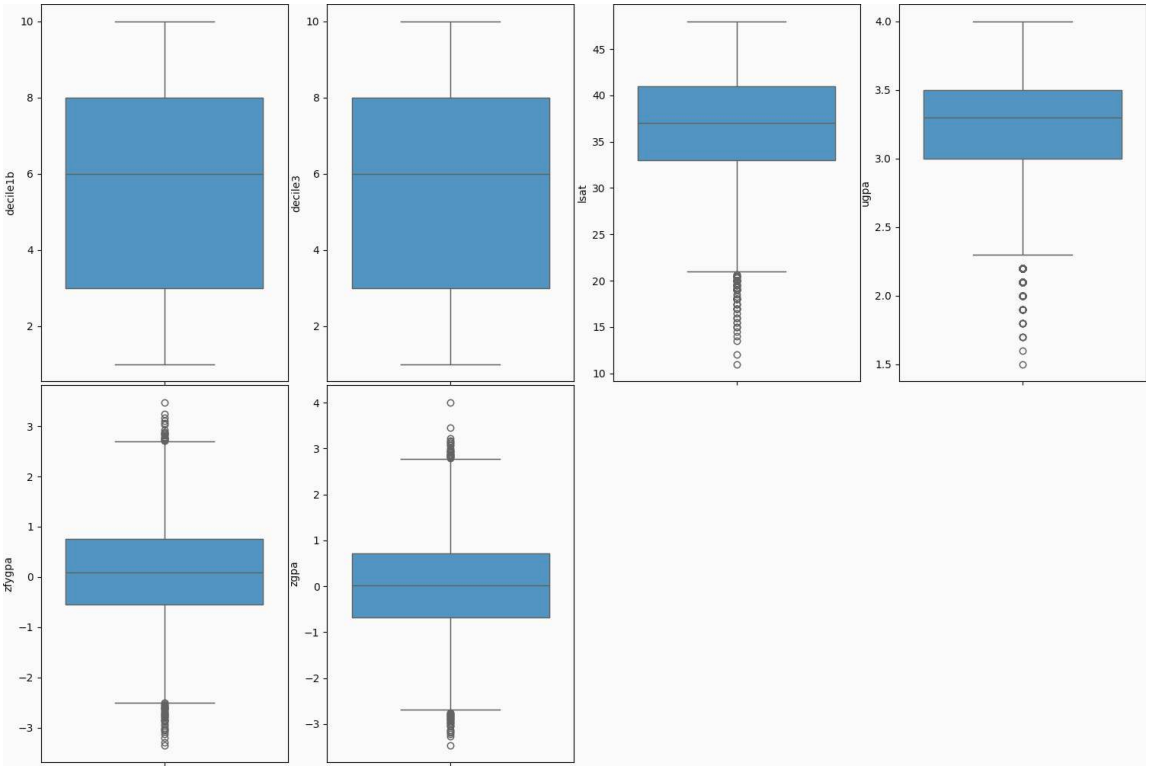


Figura 18. *Diagrama de Cajas. Admisión a la escuela de derecho*

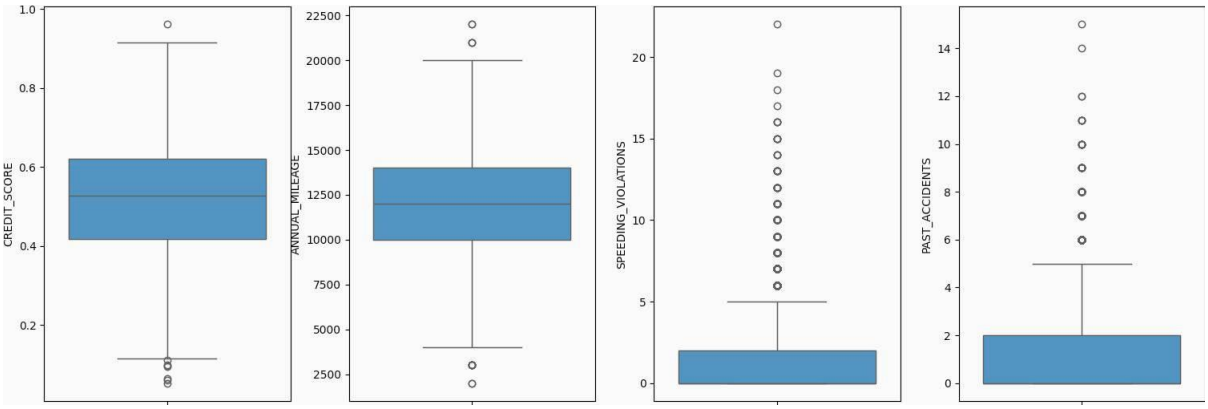


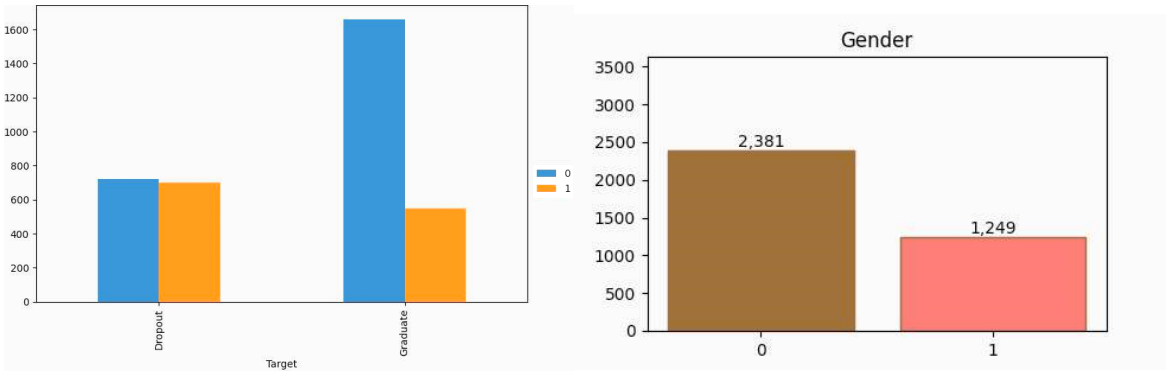
Figura 19. *Diagrama de Cajas. Reclamo de seguro de coches.*

4.3.3 Otros gráficos

Siendo consecuentes con los objetivos planteados en este trabajo, es necesario analizar cómo se comportan los atributos sensibles en comparación con otros, especialmente en relación con el atributo que contiene las clases. Este análisis permitirá evaluar la relación y la influencia de los atributos sensibles en la determinación de las clases y proporcionará información crucial para entender posibles sesgos o patrones específicos en los datos.

En el conjunto de datos sobre Abandono o Éxito académico, se observa una clara prevalencia del género femenino (0) en términos de la cantidad de muestras (Figura 20b). Además, se

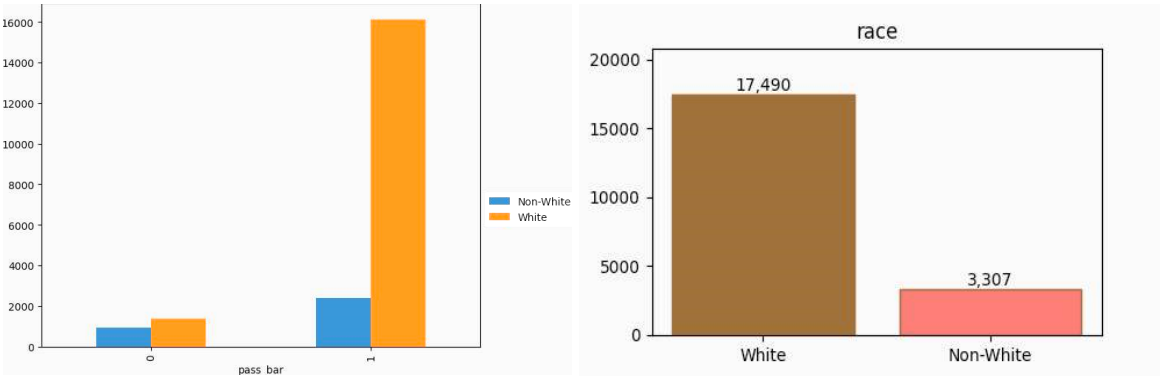
destaca que hay un mayor número de mujeres (0) que se gradúan (GRADUATE) en comparación con la cantidad de hombres (1), según se ilustra en la Figura 20a. Esta disparidad de género podría tener implicaciones importantes en el análisis de posibles graduaciones, y es crucial considerar estos desequilibrios al evaluar la equidad y diversidad en el proceso de predecir. La representación desigual de género también puede influir en la interpretación de resultados y en la toma de decisiones relacionadas con políticas educativas y de igualdad de oportunidades.



(a) Diagrama Tabulación Cruzada. *(b) Diagrama de Barras.*

Figura 20. *Diagrama de Barras (Género). Abandono o éxito académico.*

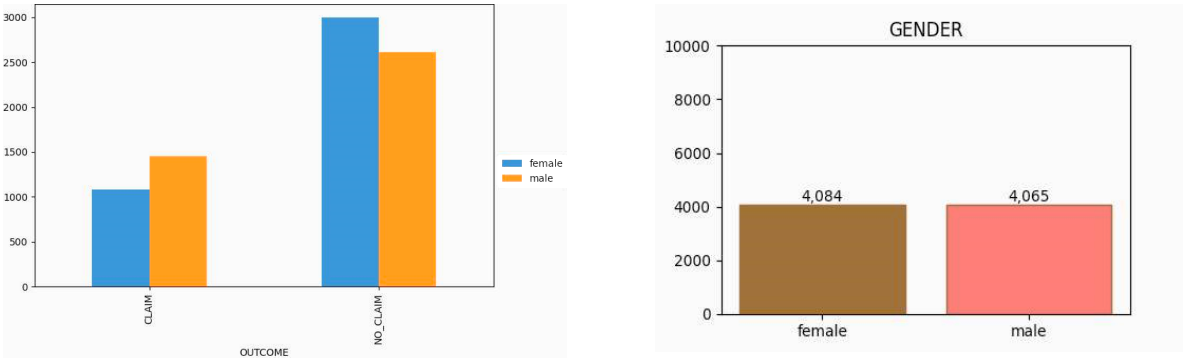
En el caso de la Admisión en la escuela de derecho, se evidencia una marcada disparidad en la cantidad de admitidos en la escuela (1) entre individuos blancos (White) y no blancos (Non-White), representando los blancos casi el 85% del total (Figura 21b). Esta discrepancia podría introducir un sesgo en el análisis, ya que la desigual cantidad de observaciones entre las categorías étnicas podría influir en las conclusiones. Sin embargo, como se puede apreciar en la Figura 21a, existe una diferencia aún más notable entre las razas cuando son admitidos que cuando no, lo que subraya aún más la importancia de considerar la variable étnica en el análisis.



(a) Diagrama Tabulación Cruzada. *(b) Diagrama de Barras.*

Figura 21. *Diagrama de Barras (Raza). Admisión en la escuela de derecho.*

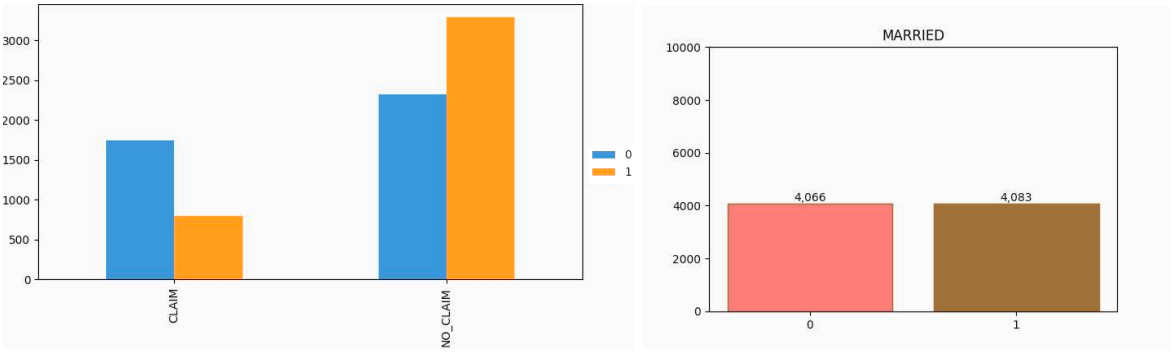
Por otra parte, en el *dataset* de Reclamo de seguro de coches, se observa que hay una cantidad similar de muestras para hombres y mujeres, como se muestra en la Figura 22b. Además, se nota una tendencia hacia una mayor proporción de mujeres que no realizan reclamos de seguro de coche debido a accidentes u otras circunstancias, en comparación con los hombres. Asimismo, al observar los datos, no se percibe una diferencia significativa entre hombres y mujeres en cada una de las categorías (Figura 22ª). Ante esta situación, es necesario llevar a cabo un análisis más exhaustivo, utilizando técnicas de vanguardia que se abordarán en próximas secciones.



(a) Diagrama Tabulación Cruzada. *(b) Diagrama de Barras.*

Figura 22. *Diagrama de Barras* (Sexo). *Reclamo de seguro de coches.*

Si consideramos otro factor sensible como si está casado, podemos observar que no ocurre lo mismo que con el género (Figura 23a). A pesar de que existe un equilibrio entre la cantidad de personas casadas y no (Figura 23b), aquellos que no están casados tienden a presentar reclamaciones al seguro con mayor frecuencia, por consecuencia, los casados tienden a ser los que menos reclaman al seguro de coche.



(a) Diagrama Tabulación Cruzada. *(b) Diagrama de Barras.*

Figura 23. *Diagrama de Barras* (Casado). *Reclamo de seguro de coches.*

4.4 Separación del conjunto de datos

En la Figura 24, se presenta la interfaz que facilita la partición del *dataset* en entrenamiento y prueba. Sin embargo, es necesario seleccionar previamente el atributo sensible que se empleará en el análisis de equidad en etapas posteriores. La separación de datos se realiza mediante un muestreo estratificado, donde se debe especificar en los parámetros de la función el porcentaje destinado a las pruebas (*test_size*), y en consecuencia, el resto se asignará al entrenamiento del algoritmo. Además, existe la opción de especificar uno o varios atributos para excluir(los) de los nuevos datos. Una vez configurados todos los parámetros, se puede proceder a realizar la división mediante el botón "Aplicar".

Figura 24. *Ventana de separación en conjunto de entrenamiento y prueba.*

Después de hacer clic en el botón "Aplicar", se habilitará la opción "Selección de Atributos" para reducir la cantidad de columnas en los conjuntos de datos, si así se desea. También estará disponible la opción "Siguiente", la cual permite avanzar a la siguiente ventana y continuar con el análisis.

4.5 Selección de atributos (SelectKBest)

Al elegir la opción de "Selección de Atributos", se desplegará una segunda ventana, tal como se ilustra en la Figura 25, que le permitirá seleccionar características utilizando las k puntuaciones más altas (SelectKBest). Además, podrá elegir entre dos funciones de puntuación

bien reconocidas en la literatura y provenientes de la biblioteca scikit-learn: el valor F de ANOVA (f_classif) y las estadísticas de Chi-cuadrado para características no negativas (chi2).

Para efectuar las modificaciones, se requiere pulsar el botón "Aceptar" para confirmarlas o, de lo contrario, seleccionar "Cancelar". Ambas alternativas resultan en el cierre de la ventana actual, dejando en primer plano la ventana previa (referenciada como Figura 24), desde donde se procederá a la ejecución de los algoritmos de clasificación al presionar el botón "Siguiente".

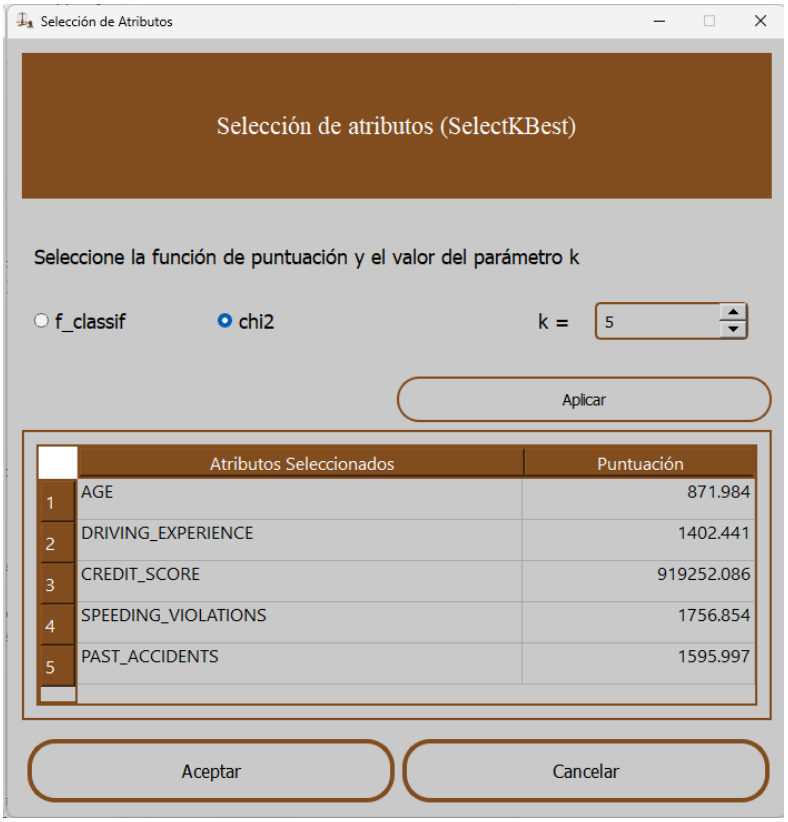


Figura 25. *Ventana de selección de atributos.*

4.6 Algoritmos de Clasificación

En este proyecto, como se ha mencionado previamente, se emplearán cuatro algoritmos de clasificación: LightGBM, Random Forest, SVM y XGBoost. Para cada uno de ellos, se podrán realizar ajustes de parámetros manualmente, variando los valores de aquellos atributos que son frecuentemente modificados según la literatura, o activando la opción de "Optimizar Parámetros" (ver Figura 26). Esta última explorará la mejor combinación de parámetros en busca de un rendimiento óptimo, siguiendo una métrica especificada por el usuario. A continuación, se enumeran los atributos que pueden ser optimizados para cada modelo.

4.6.1 LightGBM

En este modelo se podrán optimizar los siguientes parámetros.

- **max_leaf_nodes**: Número máximo de nodos hoja permitidos en un árbol. Limitar este número puede ayudar a prevenir el sobreajuste al restringir la complejidad del árbol.
- **min_samples_leaf**: Número mínimo de muestras requeridas para que un nodo sea considerado como un nodo hoja. Establecer este valor más alto puede ayudar a regularizar el modelo, evitando árboles con muy pocas muestras en los nodos hoja.
- **learning_rate**: También conocida como tasa de aprendizaje, controla la magnitud de los ajustes realizados durante el proceso de optimización del modelo. Una tasa de aprendizaje más baja generalmente conduce a una convergencia más lenta pero puede mejorar la generalización del modelo.
- **l2_regularization**: Fuerza de la regularización L2, también conocida como regularización de Ridge. La regularización L2 penaliza los coeficientes grandes en el modelo, lo que ayuda a prevenir el sobreajuste al reducir la complejidad del modelo.
- **max_iter**: Indica el número máximo de iteraciones permitidas durante el entrenamiento del modelo. Si el modelo no converge después de alcanzar este número de iteraciones, el entrenamiento se detendrá.
- **max_bins**: Controla el número máximo de contenedores (*bins*) para discretizar características continuas. Reducir este número puede ayudar a mejorar la eficiencia computacional, especialmente en conjuntos de datos con muchas características continuas.

4.6.2 Random Forest

Los posibles parámetros a optimizar son:

- **min_samples_split**: Número mínimo de muestras requeridas para dividir un nodo interno en dos nodos hijos durante la construcción de un árbol. Si el número de muestras en un nodo es menor que este valor, la división no se realizará.
- **max_depth**: Profundidad máxima de los árboles en el bosque aleatorio. Limitar la profundidad de los árboles puede ayudar a prevenir el sobreajuste y a mejorar la generalización del modelo.

- **n_estimators**: Indica el número de árboles en el bosque aleatorio. Cuanto mayor sea el número de estimadores, más robusto será el modelo, pero también aumentará el tiempo de entrenamiento.
- **Criterion**: Especifica la función para medir la calidad de una división en un árbol. Los criterios comunes incluyen "gini" para la impureza de Gini y "entropy" para la ganancia de información. La elección del criterio depende del problema y puede afectar el rendimiento del modelo.

4.6.3 SVM

En SVM se podrán optimizar los siguientes parámetros.

- **C**: Controla el parámetro de regularización, que ajusta la compensación entre maximizar el margen y minimizar la clasificación errónea. Valores más altos de C permiten un margen más estrecho y pueden conducir a un modelo más propenso al sobreajuste, mientras que valores más bajos de C promueven un margen más amplio y pueden conducir a un modelo más generalizado.
- **kernel**: Especifica el tipo de función de *kernel* a utilizar en el algoritmo SVM. Los *kernels* comunes incluyen el lineal, el polinómico y el RBF. La elección del *kernel* afecta la capacidad del modelo para separar datos no lineales en un espacio de alta dimensión.
- **gamma**: Influye en la influencia de un solo ejemplo de entrenamiento, con valores más altos que indican una influencia más limitada y valores más bajos que indican una influencia más amplia. Una baja gamma genera un modelo más suave (con influencia más amplia), mientras que una alta gamma genera un modelo más ajustado (con influencia más limitada).

4.6.4 XGBoost

En este modelo se podrán ajustar los siguientes parámetros para optimizar su rendimiento.

- **max_depth, n_estimators**: Son similares a los parámetros homónimos en Random Forest.
- **subsample**: Proporción de muestras (filas) utilizadas para entrenar cada árbol.
- **colsample_bytree**: Determina la fracción de características (columnas) a considerar al construir cada árbol. Un valor menor reduce el sobreajuste al introducir más

variabilidad, mientras que uno mayor considera más características para modelos más robustos.

- **learning_rate**: También conocida como tasa de aprendizaje, es similar al parámetro homónimo en LightGBM. Controla la magnitud de los ajustes realizados durante el proceso de optimización del modelo.

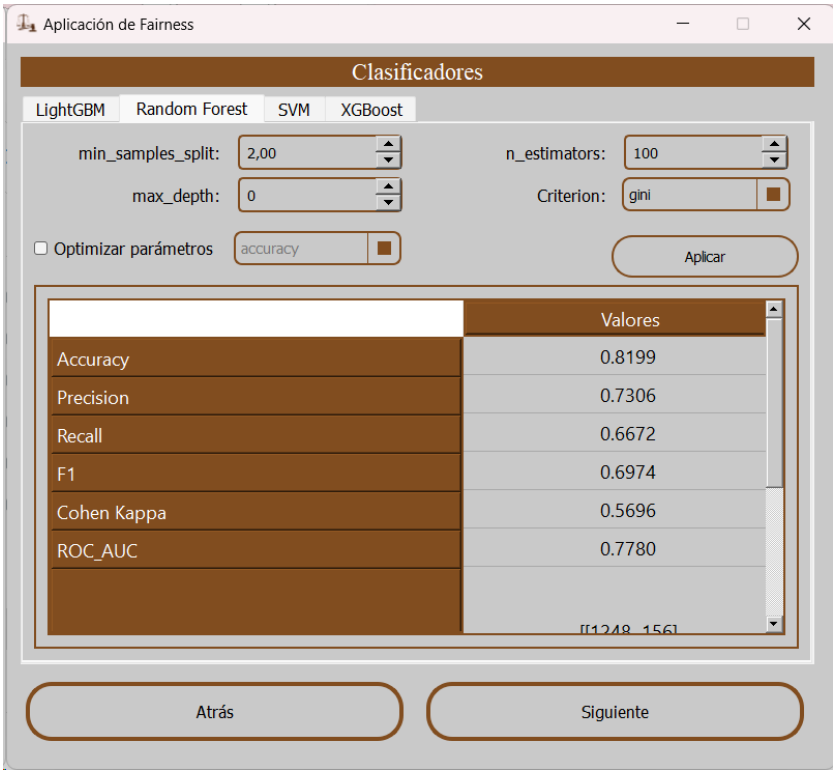


Figura 26. *Ventana de Algoritmos de Clasificación.*

4.7 Técnicas de *Fairness* y mitigación

El siguiente paso en este estudio implica el análisis de *fairness*, al cual se puede acceder a través de la aplicación una vez que se ha ejecutado uno o varios clasificadores. En la Figura 27 se muestra el panel de control perteneciente a la librería *raiwidgets* integrada en la aplicación. Sin embargo, por razones de conveniencia, también se proporciona la URL (http://localhost:5000) que se puede abrir en un navegador para acceder a todas las opciones de manera más práctica.

El panel de control permite realizar análisis en múltiples atributos sensibles. Sin embargo, dado que el objetivo de este trabajo es enfocarse en un atributo sensible y no en múltiples, se procederá a explicar un paso posterior donde se debe seleccionar primero la métrica de clasificación más adecuada al problema (Anexo A.1) y luego elegir la relacionada con la

equidad (Anexo A.2). Finalmente, el panel de control muestra los resultados en una tabla y un gráfico de barras, los cuales se analizarán más adelante (Figura 28).

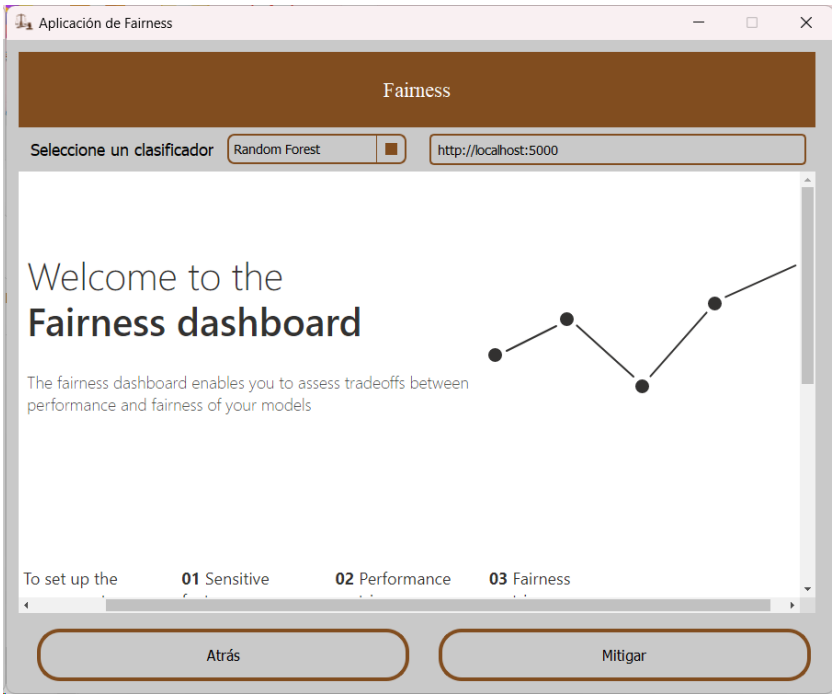


Figura 27. *Ventana de Fairness.*

Finalmente, se lleva a cabo la mitigación de la equidad (en caso de existir), siguiendo pasos iniciales similares al cálculo de la equidad, pero mostrando varios modelos mitigados. En este caso, el eje X representa la métrica del aprendizaje automático y el eje Y representa la métrica de equidad (Figura 29). El mejor modelo es aquel que logre combinar una menor equidad y una mayor métrica de aprendizaje.

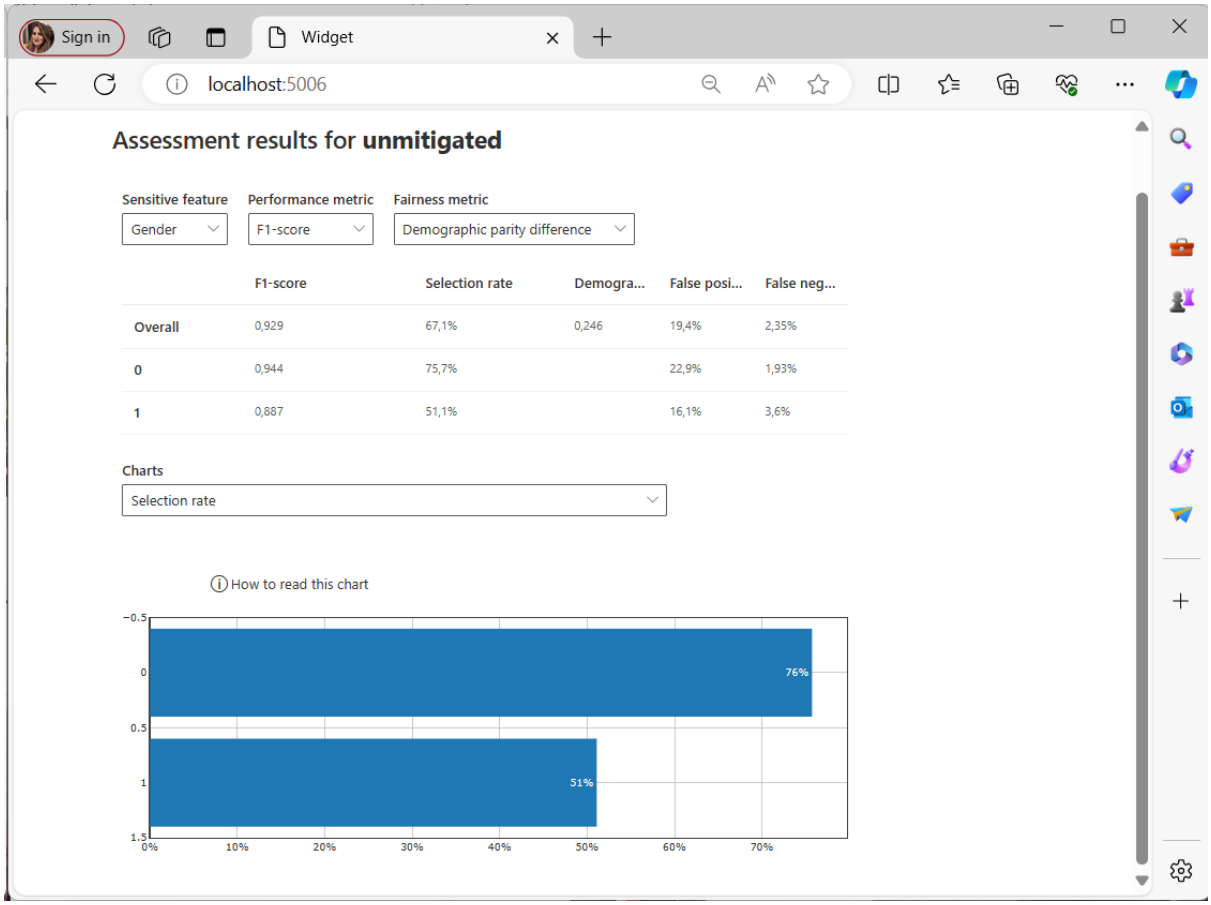


Figura 28. *Dashboard con los resultados de fairness.*

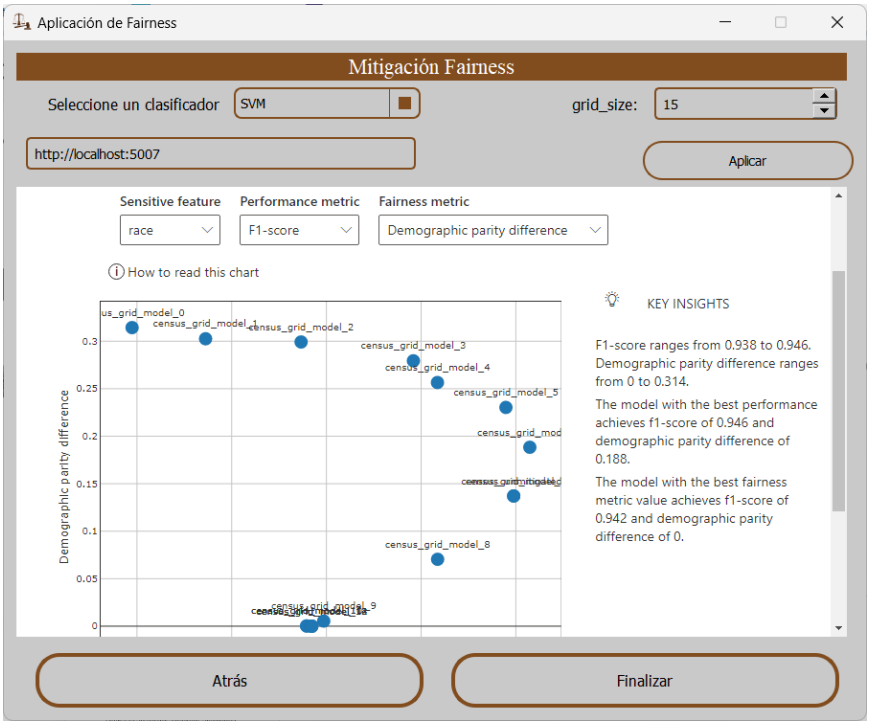


Figura 29. *Ventana de Mitigación de Fairness.*

5 Experimentos y Resultados

En esta sección se llevarán a cabo tres experimentos para cada conjunto de datos, y se realizará una comparación entre ellos. El objetivo de estos experimentos es determinar si la equidad permanece constante o si varía con las modificaciones de los parámetros de los algoritmos de clasificación, la selección de atributos y/o el tipo de procesamiento de los valores perdidos.

En todos los escenarios, dividiremos el conjunto de datos en conjuntos de entrenamiento (75%) y prueba (25%) utilizando un enfoque de muestreo estratificado, donde se asegura que las proporciones de las clases en el conjunto de datos original se mantengan en los conjuntos de entrenamiento y prueba.

Los algoritmos de clasificación mencionados en este estudio tienen ciertos valores predeterminados en sus parámetros más significativos, los cuales se pueden visualizar en la herramienta y son los siguientes:

- **Random Forest**: ‘min_samples_split’: 2, ‘max_depth’: None, ‘n_estimators’: 100, ‘criterion’ :gini.
- **LightGBM**: ‘max_leaf_nodes’: 31, ‘min_samples_leaf’: 20, ‘learning_rate’: 0.1, ‘l2_regularization’: 0, ‘max_iter’: 100, ‘max_bins’: 255.
- **XGBoost**: ‘max_depth’: 6, ‘n_estimators’: 100, ‘subsample’: 1, ‘colsample_bytree’: 1, ‘learning_rate’: 0.3.
- **SVM**: ‘C’: 1.0, ‘kernel’: rbf, ‘gamma’: scale.

En el proceso de optimización de parámetros, se emplea RandomizedSearchCV de la biblioteca sklearn. Los valores potenciales para cada clasificador son los siguientes:

- **Random Forest**: 'n_estimators': [10, 50, 100, 200, 300, 400, 500], 'criterion': ['gini', 'entropy'], 'max_depth': [1 al 32], 'min_samples_split': [1 al 30].
- **LightGBM**: 'learning_rate': [0.01, 0.05, 0.1], 'max_iter': [10 al 1000], 'max_leaf_nodes': [2 al 500], 'min_samples_leaf': [2 al 300], 'l2_regularization': [0.0 al 100.0], 'max_bins': [32 al 255]
- **XGBoost**: 'learning_rate': [0.01, 0.05, 0.1], 'max_depth': [1 al 32], 'n_estimators': [10, 50, 100, 200, 300, 400, 500], 'colsample_bytree': [0.5, 0.6, 0.7, 0.8, 0.9, 1.0], 'subsample': [0.5, 0.6, 0.7, 0.8, 0.9, 1.0].

- **SVM**: 'C': [0.1, 0.25, 0.5, 0.75, 1, 10, 100], 'kernel': ['linear', 'rbf'], 'gamma': ['scale', 'auto', 1, 0.1, 0.01, 0.001, 0.0001].

Para abordar la mitigación de la equidad, se consideró la paridad demográfica, que es ampliamente utilizada en la literatura y se considera la medida más efectiva de equidad entre distintos grupos. En dicho proceso de mitigación se generan 15 modelos.

5.1 Reclamo de seguro de coches

Este conjunto de datos presenta características distintivas en comparación con los otros dos, principalmente la presencia de valores perdidos. Teniendo en cuenta esta particularidad y que se escoge como atributo sensible 'MARRIED' (si está casado o no), se diseñaron los siguientes experimentos.

5.1.1 Experimento 1: Valores por defecto en los parámetros de los algoritmos de clasificación y eliminación de filas con nulos

El primer experimento consiste en utilizar todos los atributos y en el caso de CREDIT_SCORE y ANNUAL MILEAGE que contienen nulos, simplemente se eliminan esas filas, estas serían un total de 1851 que representan un 18.51% del total de muestras.

En la Tabla 7 se presentan los resultados de los experimentos para cada uno de los clasificadores según diversas métricas, incluyendo aquellas relacionadas con la equidad mencionadas anteriormente (ver Anexo A.3). Se observa que el SVM muestra el peor rendimiento en todos los aspectos. Sin embargo, no hay un clasificador que sobresalga por encima de los demás en todas las métricas. Por lo tanto, debido al desequilibrio de las clases, la métrica más apropiada para evaluar el rendimiento del clasificador es F1. Siguiendo esta lógica, los clasificadores LightGBM y XGBoost exhiben comportamientos similares, pero en términos de equidad, LightGBM muestra una ligera ventaja sobre XGBoost.

***Tabla 7**. Resultados de Reclamo de seguro de coches. Experimento 1.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
Exactitud	0.8283	0.8376	**0.8381**	0.6968
Precisión	**0.7407**	0.7378	0.7397	0.5270
Exhaustividad	0.6893	**0.7413**	0.7397	0.2461
F1-Score	0.7141	**0.7396**	**0.7397**	0.3355
ROC_AUC	**0.7902**	0.6216	0.6222	0.1714
Cohen Kappa	0.5916	**0.8112**	**0.8111**	0.5732
Matriz de Confusión	1251 153 197 437	1237 167 164 470	**1239 165** **165 469**	1264 140 478 156
Fairness (F1)	40% no casados, 18% casados	43% no casado, 20% casado	43% no casado, 19% casado	22% no casado, 7% casado
Demographic parity difference	0.2260	0.2350	0.2460	**0.1510**
Equalized odds difference	0.1150	**0.1040**	0.1210	0.1330
True positive rate difference	**0.0501**	0.0750	0.0727	0.1330

La Figura 30 presenta la ejecución del algoritmo LightGBM en el *dashboard* de *fairness*, donde se señala que, en general, el 31.3% de las muestras fueron clasificadas como reclamos de seguro de coche (CLAIM). Dentro de este grupo, el 43.1% está compuesto por personas casadas (1) y el 19.6% por personas no casadas (0), evidenciando una disparidad del 23.5% (valor de la métrica *demographic parity*).

Aunque puede haber algoritmos que superen a otros en términos de equidad, aún persiste un sesgo en los datos que se puede intentar mitigar para promover una mayor equidad social. Dado el comportamiento del algoritmo SVM en este conjunto de datos, no se llevará a cabo la mitigación de su equidad.

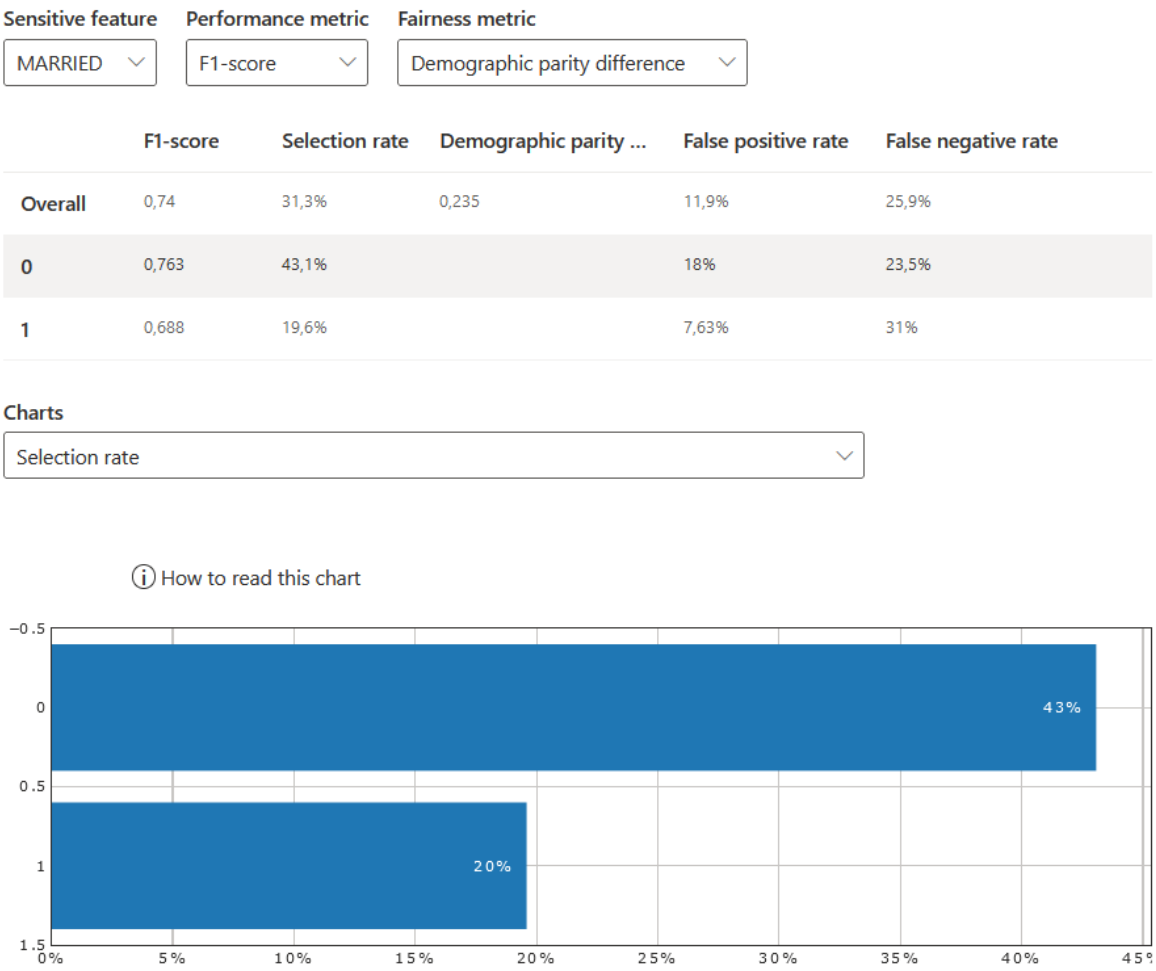


Figura 30. *Dashboard Fairlearn en Reclamo de Seguro de coches con LightGBM y Demographic Parity. Experimento 1.*

En la Figura 31 se ilustra la mitigación para Random Forest, donde se aprecia que la discrepancia en la paridad demográfica oscila entre 0.206 y 0.496 (eje Y), mientras que en el F1-score varía entre 0.206 y 0.722. Esto indica que no se logra una mejora significativa en cuanto a equidad. Sin embargo, la situación es diferente para LightGBM y XGBoost (ver Tabla 8), ya que se consigue mejorar la equidad a pesar de obtener valores ligeramente más bajos en la métrica F1. La elección del mejor modelo depende siempre de los objetivos del análisis y, en este caso, el Modelo 4 en XGBoost parece ser una buena opción para este conjunto de datos.

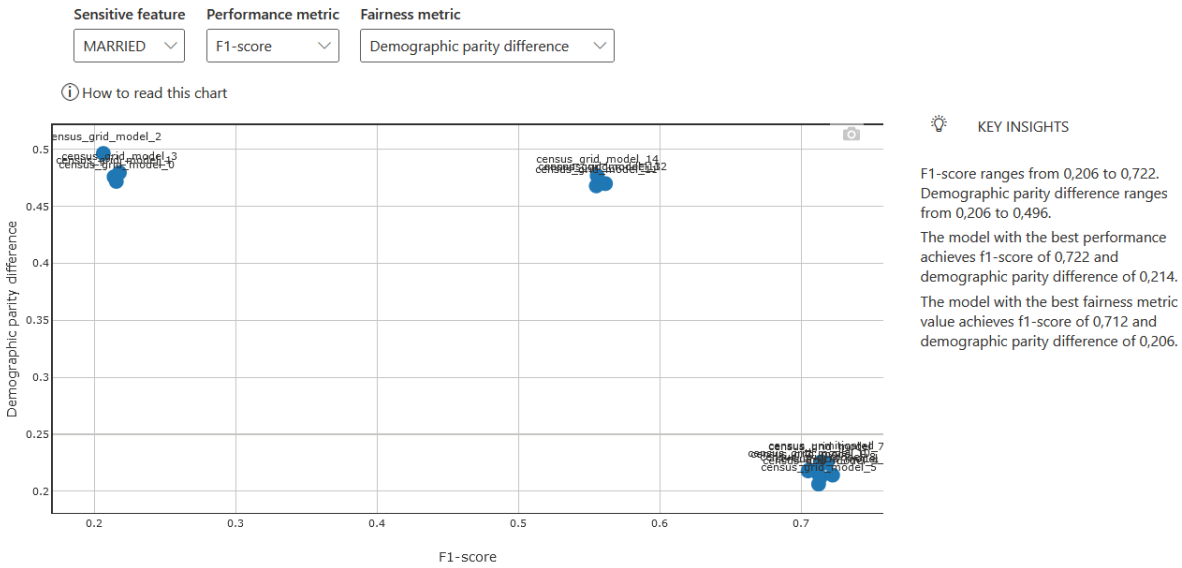


Figura 31. *Mitigación del modelo Random Forest con Paridad Demográfica en Reclamo de seguro de coches. Experimento 1.*

Tabla 8. *Mitigación de los modelos LightGBM y XGBoost con Paridad Demográfica en Reclamo de seguro de coches. Experimento1.*

	LightGBM		XGBoost	
	F1-Score	Paridad Demográfica	F1-Score	Paridad Demográfica
Modelo 4	0.6418	0.0504	**0.6718**	**0.0811**
Modelo 5	0.6940	0.1323	0.7090	0.1679
Modelo 6	0.7318	0.2034	0.7149	0.1993
Modelo 7	0.7396	0.2349	0.7397	0.2457

5.1.2 Experimento 2: Optimización de parámetros en los algoritmos de clasificación y eliminación de filas con nulo

Para el segundo experimento, conforme al título de la sección, se procederá a eliminar las filas con valores perdidos, similar al experimento 1. En esta ocasión, se evaluarán los resultados de los clasificadores únicamente utilizando la métrica F1, dada la desigualdad en las clases (ver Tabla 9). Cabe destacar que el tiempo de procesamiento computacional es considerablemente mayor para el clasificador SVM.

Se puede notar que el rendimiento de Random Forest, LightGBM y XGBoost es bastante similar, aunque XGBoost destaca ligeramente sobre los otros. En la tabla también se muestran los mejores parámetros encontrados mediante RandomizedSearchCV.

En términos del análisis de equidad, se observa que LightGBM, Random Forest, XGBoost y SVM identificaron respectivamente el 31.3%, 30.1%, 30.1% y 23.4% de las muestras como 'CLAIM'. Sin embargo, se reitera la falta de equidad en este conjunto de datos en relación con la paridad demográfica, ya que en todos los clasificadores el valor supera 0.2.

Luego se procede con la mitigación, la cual resulta en una mejora en todos los clasificadores excepto en el SVM. En la Tabla 10 se evidencia cómo, al sacrificar un poco de rendimiento en los algoritmos de clasificación, se puede alcanzar una mayor equidad. Una vez más, como en el experimento 1, el clasificador XGBoost parece ser la mejor opción, con un F1 de 0.6637 y un índice de equidad de 0.0572, aunque la diferencia entre los tres clasificadores no es significativa.

Tabla 9. *Resultados de Reclamo de seguro de coches. Experimento 2.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
F1-Score	0.7374	0.7488	**0.7573**	0.6911
Parámetros optimizados	min_samples_split: 17	max_leaf_nodes: 127	max_depth: 2	C: 1.0
	max_depth: 23	min_samples_leaf: 3	colsample_bytree: 0.7	
		max_iter: 162	subsample: 1.0	kernel: linear
	n_estimators: 10	learning_rate: 0.1	n_estimators: 300	
	criterion: entropy	l2_regularization: 57.2956	learning_rate: 0.1	gamma: auto
		max_bins: 156		
Fairness (F1)	42% no casados, 19% casados	44% no casado, 19% casado	42% no casado, 18% casado	36% no casado, 11% casado
Demographic parity difference	**0.2330**	0.2560	0.2390	0.2540
Equalized odds difference	**0.1080**	0.1220	**0.1060**	0.2580
True positive rate difference	**0.0565**	0.0988	0.0754	0.2580

Tabla 10. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Reclamo de seguro de coches. Experimento 2.*

	Random Forest		LightGBM		XGBoost	
	F1-Score	Paridad Demográfica	F1-Score	Paridad Demográfica	F1-Score	Paridad Demográfica
Modelo 4	0.6561	0.0858	0.6532	0.0564	**0.6637**	**0.0572**
Modelo 5	0.7151	0.1461	0.7239	0.1512	0.7142	0.1282
Modelo 6	0.7235	0.1894	0.7434	0.2073	0.7432	0.1973

5.1.3 Experimento 3: Selección de atributos, imputación de los valores perdidos con *k*NN (*k* vecinos más cercano) y valores por defecto en los algoritmos de clasificación

El experimento 3 difiere de los dos anteriores en su enfoque para imputar los valores perdidos. En este caso, se emplea el algoritmo *k*NN con 5 vecinos más cercanos y una distribución uniforme. Además, se realiza una selección de 11 atributos utilizando chi2. La elección exacta de esta cantidad se basa en las puntuaciones obtenidas, ya que al seleccionar un atributo adicional, su puntuación fue significativamente inferior a las demás (consulte Anexo A.4).

En la Tabla 11, se observa que el clasificador LightGBM muestra mejores resultados en todas las métricas, aunque la diferencia con XGBoost y Random Forest no es muy significativa. En contraste, SVM presenta una diferencia más notable, por lo que no se tomará en cuenta para futuros análisis de equidad.

El comportamiento de la equidad es bastante similar en los tres clasificadores y en las tres métricas. LightGBM logra identificar el 32.4% de muestras positivas con una diferencia de paridad demográfica entre casados y no casados de 0.224. Por otro lado, Random Forest, XGBoost y SVM identifican el 31%, 32.9% y 13.5% respectivamente, con métricas similares, excepto para SVM, como se muestra en la Tabla 11.

Tabla 11. *Resultados de Reclamo de seguro de coches. Experimento 3.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
Exactitud	0.8224	**0.8432**	0.8316	0.7020
Precisión	0.7187	**0.7417**	0.7202	0.5562
Exhaustividad	0.7114	**0.7663**	0.7561	0.2401
F1-Score	0.7150	**0.7538**	0.7377	0.3354
ROC_AUC	0.7922	**0.8223**	0.8111	0.5764
Cohen Kappa	0.5860	**0.6388**	0.6138	0.1807
Matriz de Confusión	1499 218 226 557	**1508 209** **183 600**	1487 230 191 592	1567 150 595 188
Fairness (F1)	42% no casados, 20% casados	45% no casado, 20% casado	45% no casado, 21% casado	22% no casado, 7% casado
Demographic parity difference	0.2240	0.2480	0.2400	**0.1400**
Equalized odds difference	**0.1030**	0.1090	0.1060	0.1460
True positive rate difference	**0.0479**	0.0695	0.0663	0.1460

Una vez confirmada la falta de equidad, se procede a la mitigación. Sin embargo, los valores para Random Forest en cuanto a equidad oscilan de 0.224 a 0.435 para la paridad demográfica, lo que indica que no mejora el comportamiento del algoritmo. Lo mismo ocurre con las otras dos métricas de equidad. Por lo tanto, en la Tabla 12 se muestran los mejores 4 modelos encontrados para LightGBM y XGBoost en cuanto a la paridad demográfica.

Nuevamente, la elección del modelo depende de los objetivos del análisis. Si se busca mejorar la equidad sacrificando un poco el F1, el Modelo 4 de XGBoost sería la opción. En cambio, si se busca reducir la falta de equidad sin sacrificar tanto el F1, se puede elegir el Modelo 5 de LightGBM o alguno con un comportamiento similar.

Tabla 12. *Mitigación de los modelos LightGBM y XGBoost con Paridad Demográfica en Reclamo de seguro de coches. Experimento 3.*

	LightGBM		XGBoost	
	F1-Score	Paridad Demográfica	F1-Score	Paridad Demográfica
Modelo 4	0.6382	0.0777	**0.6446**	**0.0845**
Modelo 5	0.7122	0.1541	0.7000	0.1685
Modelo 6	0.7370	0.1962	0.7213	0.2006
Modelo 7	0.7538	0.2485	0.7377	0.2399

5.1.4 Conclusiones parciales

Tras realizar los tres experimentos con este conjunto de datos, se puede concluir que los algoritmos con mejor rendimiento fueron LightGBM y XGBoost, ambos con valores de F1 superiores a 0.73, aunque presentaron paridades demográficas superiores a 0.2. Esta paridad demográfica pudo ser mitigada, pero a costa de una disminución en el rendimiento de los clasificadores. La optimización de parámetros junto con la eliminación de filas con valores perdidos o nulos (Experimento 2) fue el enfoque que mejores resultados arrojó para el clasificador XGBoost, obteniendo una métrica F1 de 0.7573 y una paridad demográfica de 0.2390. La paridad demográfica fue mitigada hasta 0.0572, lo que resultó en una métrica F1 de 0.6637.

5.2 Abandono o éxito académico

A diferencia del conjunto de datos sobre reclamos de seguro de coches, este conjunto no presenta valores perdidos, pero cuenta con un mayor número de atributos. El atributo sensible elegido para todos los experimentos es el género (GENDER).

5.2.1 Experimento 1: Todos los atributos y valores por defecto en los parámetros de los algoritmos de clasificación

En el primer experimento, se emplean los 36 atributos que conforman el conjunto de datos y no se realizan modificaciones en los valores de los parámetros de los clasificadores. Como se puede apreciar en la Tabla 13, Random Forest muestra un mejor rendimiento en la mayoría de las métricas, pero la diferencia con LightGBM y XGBoost no es muy significativa. Por otro lado, con SVM se observa una mayor disparidad en las métricas en comparación con los otros clasificadores.

En lo que respecta a la equidad, todos los clasificadores muestran una discrepancia en la paridad demográfica por encima de 0.2 y lograron identificar más del 66% de muestras positivas (Graduate), excepto el SVM, que identificó un 64.3%.

Tabla 13. *Resultados de Abandono o éxito académico. Experimento 1.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
Exactitud	**0.9108**	0.9053	0.9075	0.8073
Precisión	0.8869	0.8872	**0.8915**	0.8236
Exhaustividad	**0.9783**	0.9675	0.9656	0.8698
F1-Score	**0.9304**	0.9256	0.9271	0.8461
ROC_AUC	**0.8920**	0.8880	0.8913	0.7898
Cohen Kappa	**0.8071**	0.7960	0.8011	0.5889
Matriz de Confusión	**286 69** **12 541**	287 68 18 535	290 65 19 534	252 103 72 481
Fairness (F1)	76% mujeres, 51% hombres	75% mujeres, 51% hombres	74% mujeres, 50% hombres	72% mujeres, 50% hombres
Demographic parity difference	0.2530	0.2360	0.2390	**0.2130**
Equalized odds difference	0.0787	**0.0617**	0.0671	0.0702
True positive rate difference	0.0191	0.00457	**0.00215**	0.0567

Debido a los buenos resultados obtenidos por Random Forest, LightGBM y XGBoost en términos de las métricas de clasificación, se seleccionaron estos tres clasificadores para abordar la falta de equidad presente. Como se muestra en la Tabla 14, todos ellos lograron encontrar al menos cuatro modelos con un rendimiento mejorado en cuanto a la equidad. Como se ha explicado en experimentos anteriores, la elección del modelo se basa en los objetivos específicos del análisis. En este conjunto de datos, es importante destacar que el género no debería influir significativamente en el éxito académico. Por lo tanto, aunque Random Forest mostró un mejor rendimiento en sus métricas antes de la mitigación de la equidad, LightGBM demuestra un mejor equilibrio entre el F1-score y la paridad demográfica en el modelo 12 (consulte la Tabla 14).

Tabla 14*. Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Abandono o éxito académico. Experimento 1.*

Random Forest			LightGBM			XGBoost		
M[3]	F1-Score	Paridad Demográfica	M[3]	F1-Score	Paridad Demográfica	M[3]	F1-Score	Paridad Demográfica
0	0.7960	0.2456	10	0.9211	0.2240	9	0.9302	0.2252
11	0.7991	0.0437	11	0.9054	0.1670	11	0.8619	0.0840
12	0.7991	0.0244	12	**0.8592**	**0.0767**	12	0.8413	0.0534
13	0.7974	0.0145	13	0.8454	0.0154	13	0.8289	0.0024

[3] Número del modelo

5.2.2 Experimento 2: Selección de atributos y valores por defecto en los parámetros de los algoritmos de clasificación

Otra estrategia interesante para este conjunto de datos sería llevar a cabo una selección de atributos antes de aplicar los clasificadores y después de dividir el conjunto en entrenamiento y prueba. Se han seleccionado 20 atributos utilizando el método chi2, teniendo en cuenta las puntuaciones más altas. La Tabla 15 muestra los resultados para este nuevo escenario, donde se observa un claro dominio de Random Forest en cuanto a las métricas de clasificación. Sin embargo, al igual que en el experimento anterior, no se aprecia una gran diferencia con LightGBM y XGBoost.

Tabla 15. *Resultados de Abandono o éxito académico. Experimento 2.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
Exactitud	**0.9108**	0.9075	0.9009	0.8095
Precisión	**0.8907**	0.8889	0.8814	0.8265
Exhaustividad	**0.9729**	0.9693	0.9675	0.8698
F1-Score	**0.9300**	0.9273	0.9224	0.8476
ROC_AUC	**0.8935**	0.8903	0.8823	0.7926
Cohen Kappa	**0.8077**	0.8007	0.7860	0.5940
Matriz de Confusión	**289 66** **15 538**	288 67 17 536	283 72 18 535	254 101 72 481
Fairness (F1)	76% mujeres, 50% hombres	75% mujeres, 50% hombres	74% mujeres, 50% hombres	72% mujeres, 50% hombres
Demographic parity difference	0.2570	0.2460	0.2620	**0.2150**
Equalized odds difference	0.0841	0.0786	0.0846	**0.0700**
True positive rate difference	0.0214	**0.00699**	0.0430	0.0567

El patrón de equidad es consistente con el observado en el experimento 1, lo que nos lleva a implementar medidas de mitigación en los resultados obtenidos para Random Forest, LightGBM y XGBoost. En este escenario, el Random Forest exhibe un equilibrio favorable entre el F1-Score y la paridad demográfica en su modelo 11 (consulte la Tabla 16). Además, el XGBoost también demuestra un buen desempeño en su modelo 13, con un F1 de 0.8349 y un índice de equidad de 0.0523.

Tabla 16. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Abandono o éxito académico. Experimento 2.*

Random Forest			LightGBM			XGBoost		
M[4]	F1-Score	Paridad Demográfica	M[4]	F1-Score	Paridad Demográfica	M[4]	F1-Score	Paridad Demográfica
0	0.8053	0.1972	11	0.9068	0.2133	11	0.8876	0.1457
11	**0.8091**	**0.0049**	12	0.8833	0.1357	12	0.8574	0.1090
12	0.8053	0.0175	13	0.8569	0.1219	13	**0.8349**	**0.0523**
13	0.8055	0.0189	14	0.7618	0.0130	14	0.7477	0.0143

5.2.3 Experimento 3: Selección de atributos y optimización de parámetros en los algoritmos de clasificación

En el último experimento, se optó por seleccionar los 20 mejores atributos según la puntuación de chi2 y por buscar los mejores parámetros para los clasificadores utilizando RandomizedSearchCV. Como se muestra en la Tabla 17, se logró encontrar una combinación de parámetros para SVM que supera el rendimiento de los experimentos anteriores para este clasificador, mientras que en el resto de los clasificadores los resultados de las métricas fueron similares.

Una vez ejecutados los algoritmos y comprobado su buen rendimiento, se procedió a analizar la equidad. Se demostró la falta de equidad en cuanto a la Paridad Demográfica, ya que todos los valores superan 0.2. El LightGBM identificó un 65.5% de las muestras como graduados, el Random Forest un 67.4%, el XGBoost un 66.6% y el SVM un 67.4%. La distribución de cuántas mujeres y cuántos hombres fueron identificados en estos porcentajes se muestra en la Tabla 17 (Fairness F1).

[4] Número del modelo

Tabla 17. *Resultados de Abandono o éxito académico. Experimento 3.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
F1-Score	0.9258	0.9253	**0.9278**	0.9261
Parámetros optimizados	min_samples_split: 2	max_leaf_nodes: 230	max_depth: 9	C: 0.1
	max_depth: 10	min_samples_leaf: 75	colsample_bytree: 0.5	
		max_iter: 422	subsample: 0.5	kernel: linear
	n_estimators: 500	learning_rate: 0.1	n_estimators: 200	
	criterion: entropy	l2_regularization: 46.5310	learning_rate: 0.05	gamma: 0.01
		max_bins: 171		
Fairness (F1)	76.1% mujeres, 51.4% hombres	74.9% mujeres, 48.3% hombres	75.6% mujeres, 50.2% hombres	75.7% mujeres, 52% hombres
Demographic parity difference	0.2470	0.2660	0.2540	**0.2370**
Equalized odds difference	0.0789	0.0727	**0.0562**	0.0619
True positive rate difference	0.0094	0.0526	0.0454	**0.0022**

Finalmente, se intentó mitigar este problema en el modelo. Para SVM no se logró una mejora significativa en los resultados anteriores. Sin embargo, en los demás casos (Tabla 18), los modelos 12 y 13 en LightGBM y el modelo 13 en XGBoost mostraron un buen equilibrio entre la métrica F1 y la Paridad Demográfica.

Tabla 18. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Abandono o éxito académico. Experimento 3.*

Random Forest			LightGBM			XGBoost		
M[5]	F1-Score	Paridad Demográfica	M[5]	F1-Score	Paridad Demográfica	M[5]	F1-Score	Paridad Demográfica
10	0.9265	0.2357	10	0.9256	0.2409	9	0.9221	0.2361
11	0.9015	0.1887	11	0.9180	0.2239	11	0.9135	0.1940
12	0.8891	0.1382	12	0.8792	0.1445	12	**0.8667**	**0.0985**
13	0.8634	0.1029	13	**0.8515**	**0.0894**	13	**0.8500**	**0.0745**
14	0.8398	0.0626	14	0.8155	0.0384	14	0.8083	0.0463

[5] Número del modelo

5.2.4 Conclusiones parciales

Los experimentos 1 y 2 realizados con este conjunto de datos demostraron que, aunque Random Forest obtuvo las mejores métricas de F1, no fue tan efectivo en la mitigación, ya que no logró un equilibrio tan adecuado entre F1 y Paridad Demográfica como LightGBM y XGBoost. Por otro lado, la optimización de parámetros no produjo resultados significativamente mejores que los parámetros por defecto de los algoritmos de clasificación. Dado el alto coste computacional de optimizar los parámetros, utilizar los valores predeterminados de los clasificadores sería la opción más recomendable.

5.3 Admisión a la escuela de derecho

Este conjunto de datos incluye dos atributos sensibles: género (*male*) y raza (*race*). Para los experimentos, se decidió seleccionar la raza, dado que en el *dataset* de Abandono o éxito académico se había abordado previamente el género.

5.3.1 Experimento 1: Todos los atributos y valores por defecto en los parámetros de los algoritmos de clasificación

Este primer ejemplo se centra en analizar los datos en su forma más básica, sin ajustar los parámetros de los clasificadores ni excluir ningún atributo, excepto la raza. Como se puede observar en la Tabla 19, ningún clasificador se destaca por ser superior en todas las métricas. Además, debido al desbalance del *dataset*, la métrica más confiable es el F1, y en todos los casos el resultado es similar, superando 0.94.

Por otro lado, según la métrica de Paridad Demográfica, se observa falta de equidad en Random Forest, LightGBM y XGBoost. Sin embargo, esto no se aplica a SVM, donde la paridad demográfica es solo de 0.0311. Los clasificadores lograron identificar como muestras positivas (admitidos) el 94.6%, 95.4%, 94.7% y 99.5%, respectivamente.

Tabla 19. *Resultados de Admisión a la escuela de derecho. Experimento 1.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
Exactitud	0.8973	**0.9012**	0.8969	0.8927
Precisión	**0.9162**	0.9147	0.9153	0.8933
Exhaustividad	0.9736	0.9803	0.9743	**0.9987**
F1-Score	0.9440	**0.9464**	0.9439	0.9431
ROC_AUC	**0.6273**	0.6211	0.6233	0.0607
Cohen Kappa	0.3271	0.3245	0.3193	**0.5177**
Matriz de Confusión	161 412 122 4505	**150 423** **91 4536**	156 417 119 4508	21 552 6 4621
Fairness (F1)	97.3% blancos, 79.2% no blancos	98% blancos, 80.8% no blancos	97.4% blancos, 79.8% no blancos	100% blanco, 96.8% no blanco
Demographic parity difference	0.1810	0.1710	0.1760	**0.0311**
Equalized odds difference	0.3150	0.3110	0.3010	**0.0779**
True positive rate difference	0.0867	0.0782	0.0854	**0.0106**

Como se observa en la Tabla 20, hay un modelo en LightGBM (modelo 9) que supera al SVM, logrando un F1 superior a 0.9431 con una baja Paridad Demográfica. Aunque no supera al SVM en términos de paridad demográfica, sí obtiene un valor muy bajo. Por otro lado, el modelo 11 de LightGBM logra reducir aún más la paridad demográfica en comparación con el SVM, manteniendo una excelente métrica de F1.

Tabla 20. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Admisión a la escuela de derecho. Experimento 1.*

Random Forest			LightGBM			XGBoost		
M[6]	F1-Score	Paridad Demográfica	M[6]	F1-Score	Paridad Demográfica	M[6]	F1-Score	Paridad Demográfica
11	0.9414	0.0050	9	0.9441	0.0778	10	0.9417	0.0486
12	0.9406	0.0020	10	0.9418	0.0177	11	0.9396	0.0078
13	0.9408	0.0024	11	**0.9407**	**0.0025**	12	0.9396	0.0046
14	0.9412	0.0005	12	0.9411	0.0039	13	0.9386	0.0120
			13	0.9393	0.0060	14	0.9384	0.0036
			14	0.9396	0.0071			

[6] Número del modelo

5.3.2 Experimento 2: Selección de atributos y valores por defecto en los parámetros de los algoritmos de clasificación

En el segundo experimento, se seleccionaron un total de 8 atributos utilizando las puntuaciones de chi2. En la tabla 21 se exhiben los resultados del experimento, los cuales muestran comportamientos muy similares al experimento anterior.

Al igual que en el caso anterior, el SVM no muestra señales de falta de equidad en su modelo y presenta un valor de F1 superior a 0.94, lo que indica un rendimiento sobresaliente. Los clasificadores Random Forest, LightGBM, XGBoost y SVM lograron identificar como muestras positivas (admitidos) un 94.3%, 95.5%, 94.8% y 99.5% respectivamente.

El siguiente paso consiste en abordar la falta de equidad en los modelos, con el objetivo de identificar, si es posible, un modelo superior al SVM. Como se observa en la Tabla 22, los modelos 9 y 10 de LightGBM exhiben métricas de F1 mayores que el SVM y una baja Paridad Demográfica. Por lo tanto, estos dos modelos podrían ser excelentes candidatos para la predicción de nuevos datos en el futuro.

Tabla 21. *Resultados de Admisión a la escuela de derecho. Experimento 2.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
Exactitud	0.8950	**0.9004**	0.8985	0.8929
Precisión	**0.9162**	0.9136	0.9160	0.8935
Exhaustividad	0.9708	0.9808	0.9754	**0.9987**
F1-Score	0.9427	**0.9460**	0.9447	0.9432
ROC_AUC	**0.6268**	0.6160	0.6264	0.5185
Cohen Kappa	0.3214	0.3136	**0.3281**	0.0636
Matriz de Confusión	162 411 135 4492	**144 429** **89 4538**	159 414 114 4513	22 551 6 4621
Fairness (F1)	97.2% blancos, 78.1% no blancos	98% blancos, 81.6% no blancos	97.5% blancos, 79.7% no blancos	99.9% blancos, 96.8% blancos
Demographic parity difference	0.1920	0.1640	0.1780	**0.0308**
Equalized odds difference	0.3120	0.2990	0.2850	**0.0750**
True positive rate difference	0.1020	0.0747	0.0907	**0.0106**

Tabla 22. *Mitigación de los modelos Random Forest, LightGBM y XGBoost con Paridad Demográfica en Admisión a la escuela de derecho. Experimento 2.*

Random Forest			LightGBM			XGBoost		
M[7]	F1-Score	Paridad Demográfica	M[7]	F1-Score	Paridad Demográfica	M[7]	F1-Score	Paridad Demográfica
11	0.9419	0.0016	**9**	**0.9448**	**0.0626**	10	0.9422	0.0500
12	0.9420	0.0005	**10**	**0.9439**	**0.0388**	11	0.9410	0.0073
13	0.9412	0.0036	11	0.9411	0.0009	12	0.9412	0.0000
14	0.9414	0.0029	12	0.9405	0.0027	13	0.9381	0.0099
			13	0.9411	0.0026	14	0.9386	0.0010
			14	0.9414	0.0079			

5.3.3 Experimento 3: Todos los atributos y optimización de parámetros en los algoritmos de clasificación

En el último experimento, se incorporó la optimización de parámetros a los algoritmos de clasificación (ver Tabla 23). Todos los algoritmos muestran valores similares de F1, así como métricas de equidad. Sin embargo, a diferencia de los experimentos anteriores, en este caso se observa falta de equidad para SVM.

Los clasificadores Random Forest, LightGBM y XGBoost lograron identificar un mayor número de muestras positivas (admitidos) en comparación con los experimentos anteriores, alcanzando un 96.5%, 95.6% y 96.5% respectivamente. Sin embargo, esto no ocurre con SVM, que identifica un 97.1% de las muestras positivas.

[7] Número del modelo

***Tabla 23**. Resultados de Admisión a la escuela de derecho. Experimento 3.*

Clasificador/ Resultado	Random Forest	LightGBM	XGBoost	SVM
F1-Score	0.9459	0.9459	**0.9461**	0.9449
Parámetros optimizados	min_samples_split: 17	max_leaf_nodes: 3	max_depth: 1	C: 0.5
	max_depth: 8	min_samples_leaf: 42	colsample_bytree: 0.6	
		max_iter: 279	subsample: 1.0	kernel: rbf
	n_estimators: 200	learning_rate: 0.1	n_estimators: 300	
	criterion: gini	l2_regularization: 25.4874	learning_rate: 0.05	gamma: 0.001
		max_bins: 219		
Fairness (F1)	99% blancos, 82% no blancos	98.5% blancos, 79.1% no blancos	99% blancos, 83% no blancos	99% blancos, 85% no blancos
Demographic parity difference	0.1710	0.1950	0.1610	**0.1370**
Equalized odds difference	0.3620	0.4130	0.3480	**0.2800**
True positive rate difference	0.0725	0.0810	0.0630	**0.0622**

A diferencia de los experimentos anteriores, en este caso es crucial abordar la falta de equidad en el modelo SVM. Como se puede apreciar en la Tabla 24, LightGBM destaca una vez más, presentando un modelo con un excelente puntaje F1 y una baja disparidad demográfica. Es importante destacar que todos los modelos presentados en la tabla obtuvieron resultados muy prometedores en este experimento.

***Tabla 24**. Mitigación de los modelos Random Forest, LightGBM, XGBoost y SVM con Paridad Demográfica en Admisión a la escuela de derecho. Experimento 3.*

	Random Forest		LightGBM		XGBoost		SVM		
M[8]	F1-Score	Paridad Demográfica	F1-Score	Paridad Demográfica	F1-Score	Paridad Demográfica	M[8]	F1-Score	Paridad Demográfica
9	0.9449	0.0510	**0.9453**	**0.0528**	0.9441	0.0427	8	0.9443	0.0704
10	0.9429	0.0116	0.9425	0.0080	0.9417	0.0035	9	0.9419	0.0054
11	0.9416	0.0004	0.9421	0.0018	0.9418	0.0008	11	0.9416	0.0002
12	0.9418	0.0007	0.9416	0.0032	0.9419	0.0010	14	0.9417	0.0000
13	0.9420	0.0007	0.9425	0.0015					

[8] Número del modelo

5.3.4 Conclusiones parciales

Después de realizados los experimentos en este conjunto de datos se puede concluir que el clasificador SVM muestra un buen comportamiento, presentando una baja paridad demográfica en los experimentos 1 y 2, lo cual no se observó en las dos bases de datos anteriores. Además, aunque las estrategias de mitigación aplicadas en los experimentos 1 y 2 lograron mejorar los resultados, estas mejoras solo fueron significativas para el algoritmo LightGBM en el experimento 1, donde consiguió reducir su paridad demográfica. En el último experimento, el modelo XGBoost obtuvo los mejores resultados generales; sin embargo, al aplicar la mitigación, el modelo LightGBM demostró un mejor equilibrio entre la métrica F1 y la paridad demográfica.

6 Conclusiones y Trabajo futuro

En el contexto del aprendizaje automático, la equidad se ha convertido en un aspecto crucial que debe ser tenido en cuenta pues los modelos, inadvertidamente, pueden aprender y replicar prejuicios presentes en los datos de entrenamiento, resultando en decisiones discriminatorias.

Este TFM se enfocó en problemas de clasificación, donde el objetivo principal es asignar una etiqueta a cada instancia desconocida teniendo en cuenta sus características. Los problemas de clasificación son particularmente relevantes en contextos de equidad, ya que las decisiones tomadas por los modelos pueden tener un impacto significativo, por ejemplo, en el sector bancario, más específicamente en la concesión de préstamos, en la selección de candidatos para empleos, o en el sistema judicial.

En el desarrollo de la investigación se analizaron varios algoritmos de clasificación, los cuales fueron evaluados en cuanto a desempeño y utilizando métricas de equidad para determinar además cómo se comportan en términos de imparcialidad.

Para realizar este análisis, se implementó una herramienta de equidad basada en tres conjuntos de datos, cada uno con al menos un atributo sensible. Esta herramienta permite, de manera intuitiva, llevar a cabo el preprocesamiento de los datos (incluyendo la imputación de valores perdidos, la selección de atributos y la generación de gráficos), optimizar los parámetros de los algoritmos de clasificación, ejecutar el *dashboard* de equidad y mitigar la falta de equidad.

Se realizaron tres experimentos para cada uno de los conjuntos de datos, mostrando que los clasificadores XGBoost y LightGBM, obtuvieron con más frecuencia un mejor rendimiento en términos de tiempo de procesamiento y capacidad para encontrar un buen equilibrio entre la métrica F1 y la paridad demográfica. Los resultados indican que no existe un algoritmo superior en todos los contextos y que la elección del modelo debe estar guiada tanto por el rendimiento como por la equidad.

Como trabajo futuro se propone:

- Extensión de la herramienta para que se puedan analizar más bases de datos.
- Diseñar otros experimentos con el objetivo de analizar el comportamiento de los algoritmos bajo otras condiciones.
- Utilizar otros algoritmos de clasificación.
- Brindar la opción de utilizar otro *dashboard* de *fairness*.

7 Referencias Bibliográficas

Ayubkhan, S. A. H., Yap, W.-S., Morris, E., & Rawthar, M. B. K. (2023). A practical intrusion detection system based on denoising autoencoder and LightGBM classifier with improved detection performance. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7427–7452.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Bundi, D. N. (2024). Adoption of machine learning systems within the health sector: a systematic review, synthesis and research agenda. *Digital Transformation and Society*, *3*(1). https://doi.org/10.1108/DTS-06-2023-0041

Castelli, M., Vanneschi, L., & Largo, Á. R. (2018). Supervised learning: classification. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1*, 342–349.

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, *56*(7). https://doi.org/10.1145/3616865

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174225

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, *17*(1), 113–126.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240.

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2024). Fairness-aware machine learning engineering: how far are we? *Empirical Software Engineering*, *29*(1). https://doi.org/10.1007/s10664-023-10402-y

Foody, G. M. (2020). Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment*, *239*. https://doi.org/10.1016/j.rse.2019.111630

Han, Y., Yu, J., Zhang, N., Meng, C., Ma, P., Zhong, W., & Zou, C. (2023). Leverage Classifier: Another Look at Support Vector Machine. *Statistica Sinica*. https://doi.org/10.5705/ss.202023.0124

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Joe, H., & Kim, H. G. (2024). Multi-label classification with XGBoost for metabolic pathway prediction. *BMC Bioinformatics*, *25*(1). https://doi.org/10.1186/s12859-024-05666-0

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Lao, Z., He, D., Wei, Z., Shang, H., Jin, Z., Miao, J., & Ren, C. (2023). Intelligent fault diagnosis for rail transit switch machine based on adaptive feature selection and improved LightGBM. *Engineering Failure Analysis*, *148*, 107219.

Learned-Miller, E. G. (2014). Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts*, *3*.

Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. In *Nature* (Vol. 574, Issue 7780). https://doi.org/10.1038/d41586-019-03228-6

Li, L., Liu, Z., Shen, J., Wang, F., Qi, W., & Jeon, S. (2023). A LightGBM-based strategy to predict tunnel rockmass class from TBM construction data for building control. *Advanced Engineering Informatics*, *58*, 102130.

Li, S., Jin, N., Dogani, A., Yang, Y., Zhang, M., & Gu, X. (2024). Enhancing LightGBM for Industrial Fault Warning: An Innovative Hybrid Algorithm. *Processes*, *12*(1). https://doi.org/10.3390/pr12010221

Liew, X. Y., Hameed, N., & Clos, J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. *Machine Learning with Applications*, *6*, 100154. https://www.sciencedirect.com/science/article/pii/S2666827021000773

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). On the Applicability of Machine Learning Fairness Notions. *ACM SIGKDD Explorations Newsletter*, *23*(1), 14–23. https://doi.org/10.1145/3468507.3468511

Nahavandi, S., Alizadehsani, R., Nahavandi, D., Lim, C. P., Kelly, K., & Bello, F. (2024). Machine learning meets advanced robotic manipulation. *Information Fusion* (Vol. 105). https://doi.org/10.1016/j.inffus.2023.102221

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environmental Modelling and Software*, 105971. https://doi.org/10.1016/j.envsoft.2024.105971

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464). https://doi.org/10.1126/science.aax2342

Pan, H., Li, Z., Tian, C., Wang, L., Fu, Y., Qin, X., & Liu, F. (2023). The LightGBM-based classification algorithm for Chinese characters speech imagery BCI system. *Cognitive Neurodynamics*, *17*(2), 373–384.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Raschka, S. (2014). Naive Bayes and Text Classification I - Introduction and Theory. *ArXiv Preprint ArXiv:1410.5329*.

Ray, S., Haque, M., Rahman, M. M., Sakib, M. N., & Al Rakib, K. (2024). Experimental investigation and SVM-based prediction of compressive and splitting tensile strength of ceramic waste aggregate concrete. *Journal of King Saud University - Engineering Sciences*, *36*(2). https://doi.org/10.1016/j.jksues.2021.08.010

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict students' dropout and academic success. *UCI Machine Learning Repository*, *10*.

Rodriguez-Fernandez, V., & Camacho, D. (2024). Recent trends and advances in machine learning challenges and applications for industry 4.0. *Expert Systems* (Vol. 41, Issue 2). https://doi.org/10.1111/exsy.13506

Rokach, L., & Maimon, O. (2010). Classification trees. *Data Mining and Knowledge Discovery Handbook*, 149–174.

Rosenblatt, L., & Witter, R. T. (2023). Counterfactual Fairness Is Basically Demographic Parity. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, *37(12), 14461-14469*. https://doi.org/10.1609/aaai.v37i12.26691

Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 945–949.

Shen, A., Han, X., Cohn, T., Baldwin, T., & Frermann, L. (2022). Optimising Equal Opportunity Fairness in Model Training. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 4073–4084. https://doi.org/10.18653/v1/2022.naacl-main.299

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325.

Susmaga, R. (2004). Confusion matrix visualization. *Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM '04 Conference Held in Zakopane, Poland, May 17–20, 2004*, 107–116.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tang, Z., & Zhang, K. (2022). Attainability and Optimality: The Equalized Odds Fairness Revisited. *Proceedings of Machine Learning Research*, *177*.

Wang, Z., Huang, C., & Yao, X. (2024). *Procedural Fairness in Machine Learning*. arXiv preprint arXiv:2404.0187

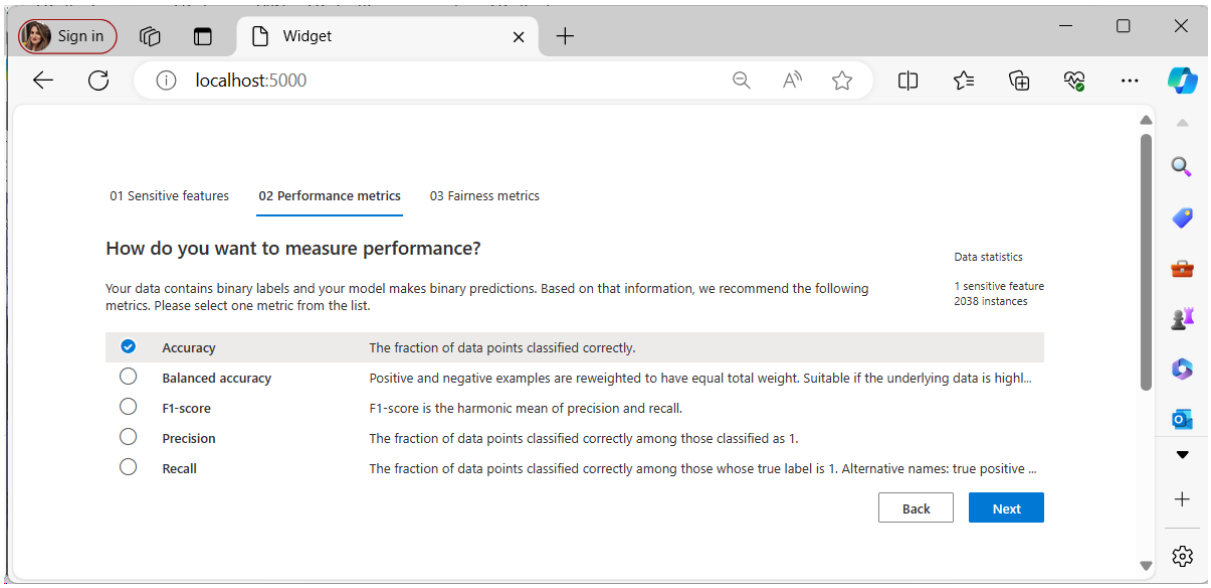
Wightman, L. F. (1998). LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. *ERIC*.

Xi, X. (2024). The role of LightGBM model in management efficiency enhancement of listed agricultural companies. *Applied Mathematics and Nonlinear Sciences*, *9*(1). https://doi.org/10.2478/amns.2023.2.00386

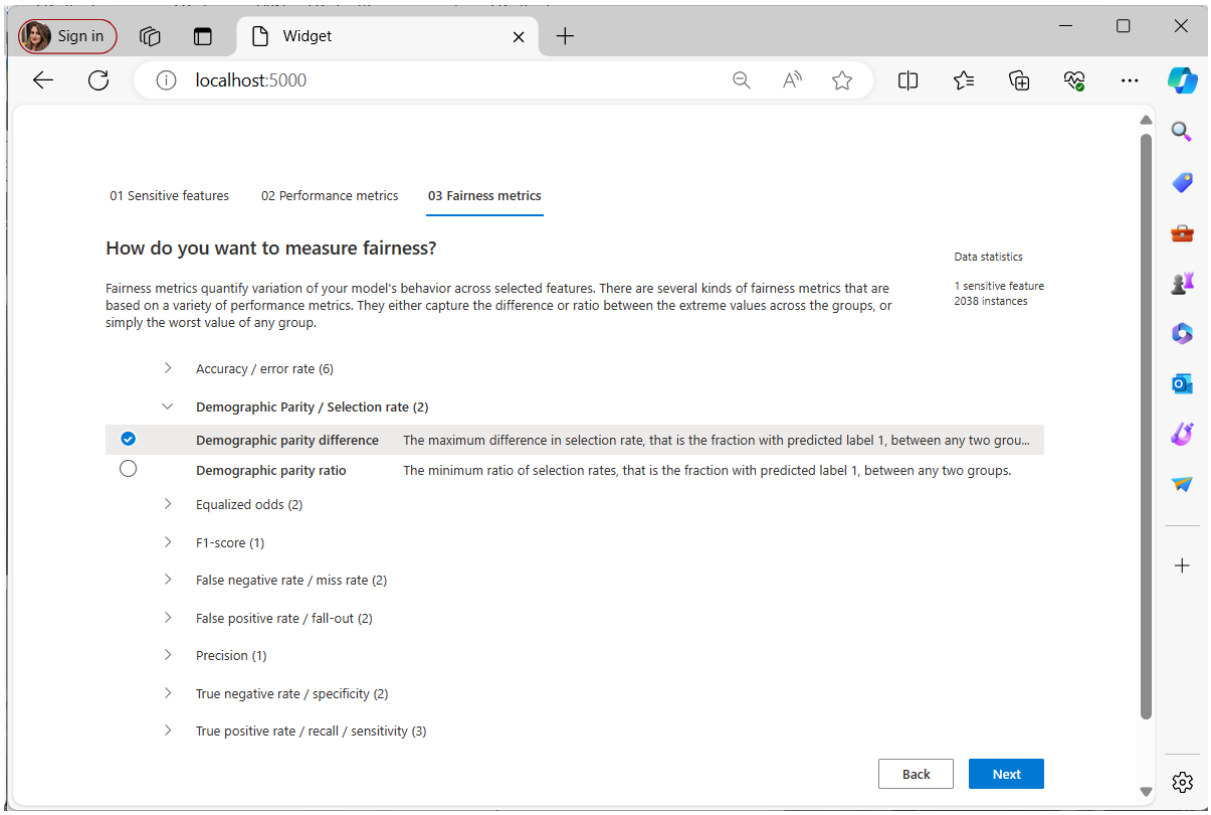
Zhou, Z. H. (2021). *Machine Learning*. Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3

8 Anexos

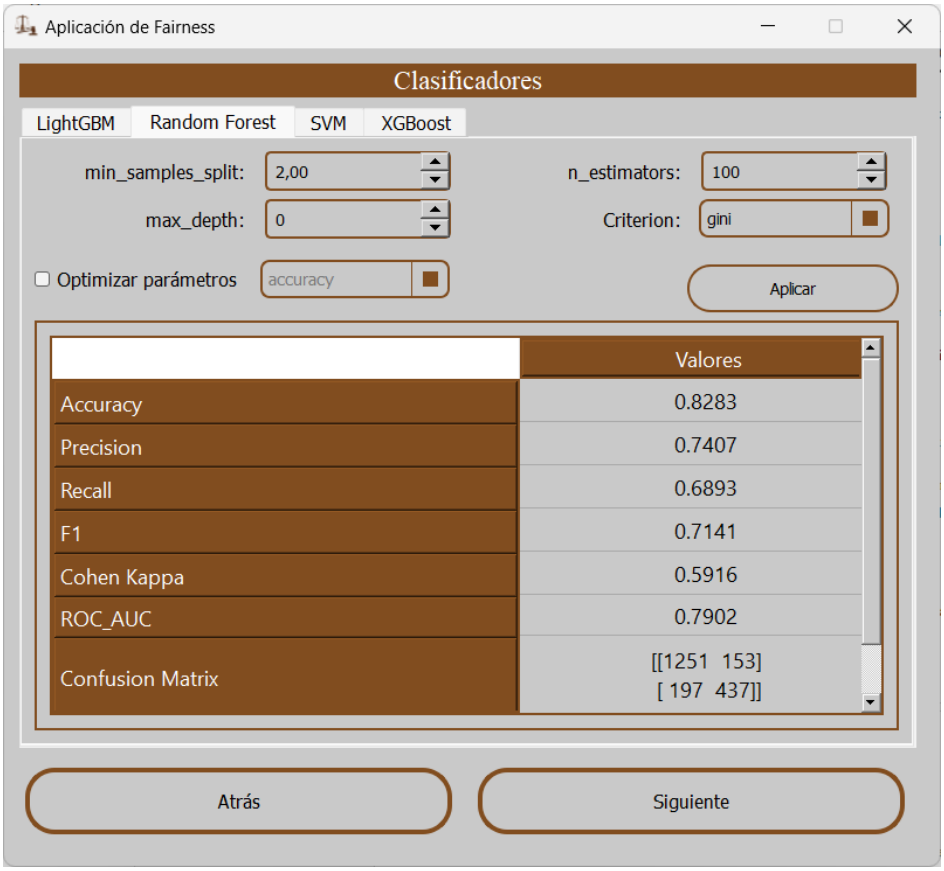
A.1 Selección de la métrica del clasificador



A.2 Selección de la métrica de fairness



A.3 Ventana de resultados de las métricas en Random Forest para el *dataset* Reclamo de seguro de coches. Experimento 1.



A.4 Ventana de selección de atributos para el *dataset* Reclamo de seguro de coches. Experimento 3