UNIVERSIDAD
DE GRANADA

HERRAMIENTAS PARA GARANTIZAR JUSTICIA EN APRENDIZAJE AUTOMÁTICO

DANIEL BOLAÑOS MARTÍNEZ

Trabajo Fin de Grado

Doble Grado en Ingeniería Informática y Matemáticas

Tutores

Jorge Casillas Barranquero

Pedro González Rodelas

FACULTAD DE CIENCIAS

E.T.S. INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Granada, a 21 de noviembre de 2021

Herramientas para garantizar justicia en aprendizaje automático

Daniel Bolaños Martínez

Daniel Bolaños Martínez. *Herramientas para garantizar justicia en aprendizaje automático*.

Trabajo de fin de Grado. Curso académico 2021-2022.

Responsables de tutorización

Jorge Casillas Barranquero
Ciencias de la Computación e Inteligencia Artificial

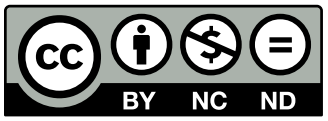
Pedro González Rodelas
Matemática Aplicada

Escuela Técnica Superior de Ingeniería Informática y Telecomunicación

Facultad de Ciencias

Doble Grado en Ingeniería Informática y Matemáticas

Universidad de Granada



The source code of this text and developed programs are available in the Github repository danibolanos/TFG-Ensuring_Fairness_in_ML

Declaración de originalidad

D. Daniel Bolaños Martínez

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2021-2022, es original, entendida esta, en el sentido de que no ha utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 21 de noviembre de 2021

Fdo: Daniel Bolaños Martínez

A mis padres y hermano, por aguantarme durante las etapas más difíciles de la época universitaria. A mis tutores, por sus consejos y brindarme total libertad en el desarrollo de este proyecto. A mi abuela, por su apoyo durante estos tres últimos meses. A mis compañeros: Jorge y Sofía por su constante ayuda altruista. Ismael, Javier y José por ser buenos amigos y mejores personas. Y finalmente a Migue, por actuar como mi hermano en la carrera.

RESUMEN

Los modelos de aprendizaje automático tienen un impacto cada vez mayor en el mundo actual, siendo utilizados para asistir y a veces sustituir a los humanos en muchos entornos. Estos modelos, a menudo funcionan aprendiendo sobre decisiones históricas tomadas por distintos grupos sociales con diferentes tasas de error en su clasificación, por lo que surge la necesidad de vigilar y controlar el sesgo involuntario de los modelos predictivos contra estos grupos de población desfavorecidos.

La mayoría de las herramientas utilizadas para mitigar el sesgo entre los grupos no privilegiados suelen depender del modelo y la métrica utilizada, aumentando así los requerimientos de los procesos en las técnicas de aprendizaje automático empleadas. Buscaremos entonces un concepto de equidad que nos permita eliminar en mayor medida la desigualdad de rendimiento entre grupos, pero que a su vez no aumente el nivel de complejidad del modelo.

En este trabajo, formalizaremos matemáticamente las definiciones de diferentes medidas de justicia y equidad; presentaremos sus propiedades, limitaciones e incompatibilidades e indagaremos en las posibles opciones para mejorar los resultados obtenidos mediante un proceso de aprendizaje automático en los términos planteados previamente: equidad por desconocimiento, paridad estadística/demográfica, medidas causales, equidad individual, entre otras.

Nos enfocaremos en desarrollar un marco para modelar la equidad usando herramientas de inferencia causal tomando como base la teoría de probabilidad y estadística. Discutiendo el concepto de equidad contrafactual que plantea que una decisión es justa para un individuo si es la misma en el mundo real y en un mundo "contrafactual" en el que el individuo perteneciese a un grupo demográfico diferente.

Finalmente haremos un análisis de las utilidades que ofrece el software Aequitas para garantizar justicia en aprendizaje automático. Realizaremos algunas pruebas para la evaluación de las medidas de equidad definidas y lo complementaremos con la resolución de un problema del mundo real utilizando herramientas de equidad contrafactual en Python. En este contexto, mostraremos herramientas de presentación y visualización de resultados que faciliten una explicación de la causa del sesgo y ayuden a los usuarios a tomar decisiones en el mundo real.

Palabras clave: medidas de equidad, mitigación del sesgo, impacto dispar, teorema de imposibilidad, equidad contrafactual, Aequitas.

ABSTRACT

Nowadays, machine learning algorithms are used in different fields with a great impact on society, such as: the process of granting bank loans (Fuster et al. [2018]), hiring staff for a job (Miller [2015]) or deciding on a criminal justice conviction (Angwin et al. [2016]). These application examples are susceptible to discrimination, which is prohibited by international law (Title VII of the Civil Rights Act: Equal Employment Opportunities).

The causes of disparities in machine learning systems are essentially exists for historical reasons. Some of the possible causes (Barocas and Selbst [2016]) are:

- Machine learning systems preserve existing discrimination in old data due to human bias. For example, if a hiring system uses the decisions made by a manager rather than his or her actual capabilities as predictive labels when selecting applicants for the position, in most cases, high-performing candidates could be rejected.
- The number of examples and the information provided by their attributes are usually smaller for the minority group, so they are less likely to be correctly modelled with respect to individuals in the majority group.
- Even if the sensitive attributes are not used in the model training, there may be attributes derived from them which, if included, will preserve the bias in the model ensemble. Sometimes it is very difficult to determine if a relevant attribute is correlated with the sensitive attributes and whether or not to include it in the training process.
- If there is an initial bias, it is likely to worsen over time. For example, the police crime register only records crimes observed by the police. The police department will tend to send more officers to places where a higher crime rate has been detected initially and therefore crimes will be more likely to be detected in those regions.

These problems have led to the development of much research on fairness in the field of machine learning, focusing on how discrimination arises, how it can be measured and how it can be mitigated. The aim of fairness will be to design algorithms that make fair predictions, avoiding disadvantaging a certain group of the population.

Despite these advances, in some settings, discrimination in modelling remains difficult to address, let alone understand, mainly due to:

- Applications generally act as black box models, in which we have restricted access to the classifier due to privacy issues, or intellectual property rights of the data used (Diakopoulos [2014]).
- Models are deployed on a population where the distribution of the data does not reflect the patterns contained in the training dataset due to, for example, a change in the population distribution (Sugiyama et al. [2017]).

Fairness in machine learning aims to study and mitigate discrimination in algorithmic decision-making processes. There are currently three approaches to bias mitigation: pre-processing methods, optimisation methods and post-processing methods. In addition to these, there are a large number of fairness criteria emerging from the social science literature that are applied to machine learning.

We will make a first approach to the concepts of equity, trying to answer the following two questions:

- **Parity or preference?** - If we are looking for fairness to achieve parity between individuals in the group or if we want to satisfy preferences within the group.
- **Treatment or impact?** - If we try to maintain fairness during data processing, or on the other hand, in the results produced by the model (impact).

In response to the above questions (Gajane and Pechenizkiy [2018]), the following equity criteria emerge:

- Unawareness.
- Individual fairness.
- Group fairness.
- Causal measures.
- Preference-based fairness.

We will need to study how the previous concepts are formalised in the field of machine learning and we will present its formalisations in different branches of computer engineering and mathematics. From a practical point of view, it is important to understand these fairness criteria and their implications by carrying out a theoretical and empirical analysis based on notions from the social science literature. This analysis is intended to help determine the goodness of existing formalisations and to be able to assess the advantages and disadvantages of each criterion with the aim of improving or constructing new fairness formalisations in the future.

As the main objective of our work, we suggest an exhaustive literature review of the bias mitigation methods and fairness measures presented in different articles such as Gajane and Pechenizkiy [2018] or Verma and Rubin [2018]. We propose to detail the mathematical tools that will be useful for the formalisation of fairness measures, with special emphasis on causal inference.

In the practice, we will perform an analysis of the fairness research software Aequitas (Saleiro et al. [2019]) on real-world examples that may have problems of bias among certain demographic groups. We will use the software on a dataset and will evaluate its results. We will also replicate an example based on the notion of counterfactual fairness (Kusner et al. [2018]) using the Python programming language.

The main contributions of this work are:

- Formalise the different families of equity measures and analyse the most relevant bias mitigation algorithms in this field.
- Present an alternative proof of the impossibility theorem to show the incompatibilities between different group fairness criteria.
- Provide an empirical analysis, using the Aequitas software, for the evaluations of the theoretically studied fairness measures.
- Replicate a real-world example of a causal model, in the Python programming language, based on the concept of counterfactual fairness, available at the link below[1].
- Contrast the results obtained by applying counterfactual fairness techniques with the different notions of equity considered.

In general terms, the work developed has the following structure:

- **PART I - Basic mathematical principles.** Analysis of the basic mathematical principles used in the course of the work. In Chapter 2 we will define results related to probability theory. In Chapter 3 we will present some examples of probability distributions. Finally, Chapter 4 will introduce some concepts of graph theory that will serve as a basis for the causal inference proposed later in the project.

- **PART II - Fairness in machine learning.** We will introduce the tools needed to formalise the different measures of fairness and a review of the most popular bias mitigation algorithms. In Chapter 5 we will introduce machine learning by presenting its properties and main evaluation metrics. In Chapter 6 we will formalise the most popular fairness measures in the literature and establish a subdivision into families. Finally, in Chapter 7 we will classify the different bias mitigation algorithms and present some examples of each type.

- **PART III - Basics of counterfactual fairness.** We will present the concept of counterfactual fairness, providing the mathematical basis on which it is based and the motivations for its construction. In Chapter 8 we will introduce causal inference and define its main tools for working in this field. In Chapter 9 we will provide an alternative proof of the impossibility theorem of fairness. Finally, in Chapter 10 we will present the concept of causal measures and define counterfactual fairness as a subclass of them based on the notion of counterfactual.

1 https://github.com/danibolanos/TFG-Ensuring_Fairness_in_ML.git

- **PART IV - Experimental analysis.** Python experiment about counterfactual fairness based on the problem proposed in Kusner et al. [2018]. In Chapter 11 we will describe the problem and examine the different designs of the proposed causal models. Finally, in Chapter 12 we will analyse the code of the proposed solution, the conditions of the experiment and discuss the results obtained. We will also include a tutorial to reproduce the experiment.
- **PART V - Conclusions and further directions.** In Chapter 13 we will analyse the conclusions reached from the project and in Chapter 14 we will present possible directions for future development.
- **BIBLIOGRAPHY.** Compilation of the different literature sources referred to throughout the work.
- **APPENDIXES.** We will analyse the Aequitas software and we will also make an estimate of the planning and cost of the project.

This work has shown us that the notions of unawareness and individual fairness, which might seem sufficient to avoid bias in a dataset, have many disadvantages when applied in practice. Moreover, the main drawbacks with the concepts of fairness group are the incompatibility between its three most popular measures, the relaxation of several of its notions and its observational nature.

Causal measures emerge as a solution to the problems described above, and counterfactual fairness is defined as a subclass of them. However, counterfactual fairness, although it has many advantages, also has some obvious problems as it implies some previous mathematical knowledge about causal inference. Nevertheless, this concept in general, manages to eliminate the existing bias and offers visual tools for the correct interpretation of the data by a user without extensive programming skills.

Finally, the use of software tools for bias detection such as Aequitas on several datasets, presents us with the need to use optimisation algorithms such as the one replicated in the experiment to work with a dataset that could integrate unfair treatments on specific groups of the population. Furthermore, it points out the obvious lack of interaction and visualisation tools for results that could be used by experts from social and technological fields in the same way.

Keywords: fairness metrics, bias mitigation, disparate impact, impossibility theorem, counterfactual fairness, Aequitas.

ÍNDICE GENERAL

1

INTRODUCCIÓN

1.1 CONTEXTO HISTÓRICO

Actualmente, los algoritmos de aprendizaje automático se utilizan en ámbitos diversos con un gran impacto en la sociedad, como son: el proceso de concesión de créditos bancarios (Fuster et al. [2018]), selección de personal para un puesto de trabajo (Miller [2015]) o decisión de una condena en justicia penal (Angwin et al. [2016]). Estos ejemplos de aplicación son propensos a la discriminación, la cual está prohibida por la legislación internacional (Title VII of the Civil Rights Act: Equal Employment Opportunities).

Las causas de las disparidades en los sistemas de aprendizaje automático existen en los conjuntos de datos de entrenamiento debido a razones históricas. Algunas de las posibles causas (Barocas and Selbst [2016]) son:

- Los sistemas de aprendizaje automático mantienen la discriminación existente en los datos antiguos debido al sesgo humano. Por ejemplo, si un sistema de contratación a la hora de seleccionar a los aspirantes al cargo utiliza como etiquetas de predicción las decisiones tomadas por un directivo en lugar de sus capacidades reales, en la mayoría de los casos se podrían rechazar candidatos con un alto nivel de rendimiento.
- La cantidad de ejemplos y la información que ofrecen sus características normalmente son menores para el grupo minoritario, por lo que es menos probable que se modelen correctamente con respecto a los individuos del grupo mayoritario. Esto tendrá inconvenientes a la hora de predecir sobre nuevos ejemplos pertenecientes al grupo desfavorecido.
- Aunque los atributos sensibles no se utilicen en el entrenamiento del modelo, podrán existir atributos derivados de éstos, los cuales si se incluyen, mantendrán el sesgo en el conjunto. A veces, es muy difícil determinar si un atributo relevante está correlacionado con los atributos sensibles y si debemos incluirlo o no en el proceso de entrenamiento.
- Si ya existe un sesgo inicial, probablemente se agravará con el tiempo. Por ejemplo, en el registro policial de delitos solo constan delitos observados por la policía. El departamento de policía tenderá entonces a enviar más agentes a lugares

donde se ha detectado una mayor tasa de delincuencia en un inicio y por tanto, será más probable que se detecten delitos en esas regiones.

Estos problemas han llevado al desarrollo de numerosas investigaciones sobre la equidad en el ámbito del aprendizaje automático, enfocadas en cómo surge la discriminación, cómo puede medirse y cómo puede mitigarse. El objetivo de la equidad será por tanto, diseñar algoritmos que hagan predicciones justas, evitando perjudicar a un determinado grupo de la población.

A pesar de estos avances, en algunos escenarios, la discriminación en los modelos sigue siendo difícil de abordar, y mucho más de entender debido principalmente a:

- Las aplicaciones generalmente actúan como modelos de caja negra, en las cuales tenemos acceso restringido al clasificador por cuestiones de privacidad, o derechos de propiedad intelectual de los datos utilizados (Diakopoulos [2014]).
- Los modelos se despliegan sobre una población dónde la distribución de los datos no refleja los patrones contenidos en el conjunto de entrenamiento debido, por ejemplo, a un cambio en la distribución de la población de interés (Sugiyama et al. [2017]).

En este trabajo, estudiaremos cómo se formaliza el concepto de equidad en el ámbito del aprendizaje automático y presentaremos estas formalizaciones en las diferentes ramas de la ingeniería informática y las matemáticas. Desde el punto de vista práctico, es importante estudiar estos criterios de equidad y sus implicaciones realizando un análisis teórico y empírico a partir de las nociones de la literatura de las ciencias sociales. Este análisis tendrá la intención de ayudar a determinar la bondad de las formalizaciones existentes y poder valorar las ventajas e inconvenientes de cada criterio con el objetivo de mejorar o construir nuevas formalizaciones de equidad en un futuro.

1.2 ANÁLISIS DEL PROBLEMA

La equidad en el aprendizaje automático, tiene como objetivo estudiar y mitigar la discriminación en los procesos de toma de decisiones algorítmicas. Actualmente podemos encontrar tres enfoques dentro de la mitigación de los sesgos. Los métodos de preprocesamiento que intentan corregir los sesgos de los datos introducidos en el modelo, los métodos de procesamiento interno u optimización que tratan de realizar la corrección de los sesgos producidos durante el proceso de aprendizaje y finalmente, los métodos de posprocesamiento que tienen como objetivo corregir el resultado de un modelo sesgado.

A estos enfoques se le suman la gran cantidad de criterios de justicia que surgen a partir de la literatura de las ciencias sociales aplicadas al aprendizaje automático. Podemos hacer una primera aproximación a estas familias y por ende a la formalización de los conceptos de equidad, intentando dar respuesta a las siguientes dos preguntas:

- **¿Paridad o preferencia?** - Si buscamos equidad para lograr una paridad entre los individuos del grupo o en cambio queremos satisfacer unas preferencias dentro del mismo.
- **¿Tratamiento o impacto?** - Si tratamos de mantener la equidad durante el tratamiento de los datos, o por el contrario, en los resultados producidos por el modelo (impacto).

Dando respuesta a las preguntas anteriores, surgen los siguientes criterios de equidad, que se resumen en la Tabla 1 (Gajane and Pechenizkiy [2018]).

	Paridad	**Preferencia**
Tratamiento	Equidad por desconocimiento Medidas causales	Tratamiento preferente
Impacto	Equidad de grupo Equidad individual	Impacto preferente

Tabla 1: Formalización de los criterios de equidad.

Debido a la gran cantidad de opciones que se nos presentan, nos gustaría saber qué nociones de equidad tienen menos limitaciones en la práctica, y qué método de mitigación de sesgo aporta menos complejidad en los algoritmos de predicción. De esta forma, podremos conocer las virtudes de cada familia de equidad y ser capaces de desarrollar nuevos conceptos o adaptar cada uno al caso de estudio que mejores resultados nos proporcione.

1.3 OBJETIVOS DEL TRABAJO

Nuestro objetivo principal será realizar un análisis teórico de las nociones de equidad y de los algoritmos construidos en base a ellas. Estudiaremos qué criterios aportan más beneficios sobre un modelo de caja negra del que únicamente podamos conocer los datos antes o después de ser procesados, e indagaremos sobre los métodos de optimización de las nociones de justicia con el propósito de satisfacer la equidad en el modelo, minimizando la pérdida de rendimiento del mismo.

Como resultado de los objetivos planteados, sugerimos una revisión bibliográfica exhaustiva de los métodos de mitigación del sesgo y de las medidas de equidad que se nos presentan en diferentes artículos como Gajane and Pechenizkiy [2018] o Verma and Rubin [2018]. Nos proponemos detallar las herramientas matemáticas que nos serán útiles para la formalización de las medidas de equidad, haciendo especial hincapié en la inferencia causal como base sobre la que se construye el modelo de equidad contrafactual que desarrollaremos en este trabajo.

En la práctica, realizaremos un análisis de la herramienta de software Aequitas (Saleiro et al. [2019]) dedicada al estudio de la equidad sobre ejemplos del mundo real que pudiesen tener problemas de sesgo entre determinados grupos demográficos. Utilizaremos el software sobre un conjunto de datos de prueba y evaluaremos sus resultados. Además replicaremos un ejemplo basado en la noción de equidad contrafactual (Kusner et al. [2018]) utilizando el lenguaje de programación Python. Extraeremos las conclusiones desde la perspectiva causal y los compararemos con los resultados obtenidos con Aequitas para el mismo conjunto de datos. Finalmente, aportaremos algunas gráficas para facilitar la interpretación de los resultados a otros investigadores o científicos de datos interesados en el tema

1.4 CONTRIBUCIONES

En resumen, las contribuciones principales de este trabajo son:

- Formalizar las distintas familias de medidas de equidad y analizar los algoritmos de mitigación de sesgo de mayor interés en este ámbito.
- Facilitar una demostración alternativa del teorema de imposibilidad, para mostrar las incompatibilidades entre los diferentes criterios de equidad de grupo.
- Proporcionar un análisis empírico, utilizando el software Aequitas, para las evaluaciones de las medidas de justicia estudiadas teóricamente.
- Replicar un ejemplo del mundo real sobre un modelo causal, en el lenguaje de programación Python, basándonos en el concepto de equidad contrafactual, disponible en el enlace siguiente[1].
- Contrastar los resultados obtenidos al aplicar las técnicas de equidad contrafactual con las diferentes nociones de justicia consideradas.

1.5 ESQUEMA GENERAL

En líneas generales, el trabajo desarrollado tiene la siguiente estructura:

- **PARTE I - Principios matemáticos básicos.** En esta parte se realizará un análisis de los fundamentos matemáticos básicos utilizados a lo largo del trabajo con el objetivo de aportar una base al lector sin conocimientos previos en la materia. En el Capítulo 2 definiremos resultados relativos a la teoría de probabilidad. En el Capítulo 3 presentaremos algunos ejemplos de distribuciones de probabilidad que nos serán útiles en el proyecto. Finalmente en el Capítulo 4 se introducirán algunos conceptos sobre teoría de grafos que servirán como base de la inferencia causal propuesta más adelante.

1 https://github.com/danibolanos/TFG-Ensuring_Fairness_in_ML.git

- **PARTE II - Justicia en aprendizaje automático.** En esta parte se introducirán las herramientas necesarias para formalizar las diferentes medidas de equidad de la literatura de justicia en aprendizaje automático y se realizará un análisis de los algoritmos más populares relacionados con el tema. En el Capítulo 5 haremos una introducción al aprendizaje automático presentando sus propiedades, técnicas de actuación y sus principales métricas de evaluación. En el Capítulo 6 formalizaremos las medidas de equidad más populares de la literatura y estableceremos una subdivisión en familias. Finalmente en el Capítulo 7 clasificaremos los diferentes algoritmos de mitigación de sesgo y presentaremos algunos ejemplos de cada tipo.
- **PARTE III - Fundamentos de la equidad contrafactual.** Esta parte tendrá como objetivo presentar el concepto de equidad contrafactual, aportando tanto la base matemática en la que se fundamenta, como las razones que motivan su construcción. En el Capítulo 8 introduciremos la inferencia causal y definiremos los modelos y grafos causales como herramientas útiles para trabajar en este campo. En el Capítulo 9 proporcionaremos una demostración alternativa del teorema de imposibilidad de la equidad, la cual será una de las razones principales que motiven la utilización de la equidad contrafactual frente al resto de medidas propuestas. Finalmente en el Capítulo 10 presentaremos el concepto de medidas causales, y definiremos la equidad contrafactual como una subclase de las mismas, a partir de la noción de contrafactual.
- **PARTE IV - Análisis experimental.** En esta parte se elaborará un ejemplo práctico en Python sobre la equidad contrafactual basado en el problema propuesto en Kusner et al. [2018]. En el Capítulo 11 describiremos el problema, y analizaremos los diferentes diseños de los modelos causales propuestos. Finalmente en el Capítulo 12 analizaremos el código de la solución propuesta, las condiciones de la experimentación y discutiremos los resultados obtenidos. Incluiremos también un tutorial para reproducir el experimento.
- **PARTE V - Conclusiones y vías futuras.** En esta parte se analizarán las conclusiones generales extraídas del proyecto y se presentarán ideas futuras que puedan ser desarrolladas a partir de las conclusiones y el contexto inicial del trabajo. En el Capítulo 13 haremos la conclusión del proyecto y en el Capítulo 14 analizaremos las vías futuras.
- **BIBLIOGRAFÍA.** Recopilación de las distintas fuentes bibliográficas referidas a lo largo del trabajo.
- **APÉNDICES.** En el Apéndice A introduciremos la herramienta Aequitas y comprobaremos su funcionamiento para las diferentes medidas de equidad de grupo estudiadas sobre varios conjuntos de datos. En el Apéndice B realizaremos una estimación de la planificación y coste del proyecto simulando su valoración en el ámbito laboral.

Parte I

PRINCIPIOS MATEMÁTICOS BÁSICOS

Definiciones y resultados relativos a teoría de probabilidad, estadística y teoría de grafos.

2

TEORÍA DE LA PROBABILIDAD

En este capítulo incluiremos definiciones y resultados previos de la teoría de probabilidad que utilizaremos a lo largo del desarrollo del trabajo. La fuente principal utilizada en este capítulo parte del trabajo contenido en Dembo [2014].

2.1 ESPACIO DE PROBABILIDAD

Construiremos la teoría asumiendo que existe un conjunto no vacío Ω que representa al conjunto de todos los posibles resultados de un experimento. Llamaremos *suceso* a cualquier subconjunto de Ω.

Definición 1 (σ-álgebra)**.** Sea $\mathcal{P}(\Omega)$ el conjunto de partes de Ω. Llamaremos *σ-álgebra* a $\mathcal{A} \subset \mathcal{P}(\Omega)$ que satisfaga:

- $\mathcal{A}$ contiene al conjunto vacío: $\emptyset \in \mathcal{A}$.
- $\mathcal{A}$ es cerrado bajo complementarios: si $A \in \mathcal{A}$, entonces $\Omega \setminus A \in \mathcal{A}$.
- $\mathcal{A}$ es cerrado bajo uniones numerables: si $A_i \in \mathcal{A}$ para todo $i \in \mathbb{N}$ y $B = \cup_{i\in\mathbb{N}} A_i$, entonces $B \in \mathcal{A}$.

De las propiedades anteriores deducimos que $\Omega \in \mathcal{A}$ y que $\mathcal{A}$ también es cerrado bajo intersecciones numerables.

Definición 2 (Medida de probabilidad)**.** Una *medida de probabilidad* P sobre un espacio de medida $(\Omega, \mathcal{A})$ es una función $P\colon \mathcal{A} \to [0,1]$ que verifica:

- $P(\Omega) = 1$.
- Si $A \subset \Omega$, entonces $P(A) \geq 0$.
- P es σ-aditiva, es decir: dada $\{A_i\}_{i\in\mathbb{N}}$ una sucesión de conjuntos disjuntos dos a dos en $\mathcal{A}$, entonces

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(A_i).$$

Denotaremos por *suceso seguro* al suceso que siempre va a ocurrir. A partir de la primera condición sabemos que el suceso seguro tiene la máxima probabilidad posible. La segunda condición garantiza la no negatividad de la probabilidad. Por último, la tercera condición implica que dado un conjunto de sucesos disjuntos dos a dos, la probabilidad de que ocurra cualquiera de ellos es igual a la suma de las probabilidades de cada uno.

Proposición 1. Toda medida de probabilidad, P, cumple:
- $P(\emptyset) = 0$.
- Dados $A, B \in \mathcal{A}$, entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Definición 3 (Espacio de probabilidad)**.** Definimos como *espacio de medida* a la tupla $(\Omega, \mathcal{A}, \mu)$ donde $\mu \colon \mathcal{A} \to \mathbb{R}_0^+$ es una medida en $(\Omega, \mathcal{A})$ y *espacio de probabilidad* a la tupla formada por $(\Omega, \mathcal{A}, P)$ donde P una medida de probabilidad en $(\Omega, \mathcal{A})$.

2.2 VARIABLES ALEATORIAS

Una *variable aleatoria* es una función que asigna un valor, normalmente numérico, al resultado de un experimento aleatorio. Dada una variable aleatoria no es posible saber su valor exacto al ser medida, aunque sí conocemos una distribución de probabilidad para describir la probabilidad de que se den los diferentes valores. A continuación, formalizaremos el concepto de variable aleatoria.

Definición 4 (Función medible)**.** Sean $(\Omega_1, \mathcal{A})$ y $(\Omega_2, \mathcal{S})$ dos espacios de medida. Una *función medible* n-dimensional es una función $X \colon \Omega_1 \to \Omega_2$ que verifica:

$$X^{-1}(S) = \{\omega \in \Omega \ : \ X(\omega) \in S\} \in \mathcal{A}, \text{ para todo } S \in \mathcal{S}.$$

Definición 5 (Variable aleatoria)**.** Sea $(\Omega_1, \mathcal{A}, P)$ un espacio de probabilidad y $(\Omega_2, \mathcal{S})$ un espacio de medida. Una *variable aleatoria* $\mathbf{X} = (X_1, \ldots, X_n)$ es una función medible $\mathbf{X} \colon \Omega_1 \to \Omega_2$ del espacio de probabilidad al espacio de medida.

Diremos que la variable aleatoria es *unidimensional* si $n = 1$ y *multivariante* cuando $n > 1$. Cuando tengamos una variable aleatoria multivariante $\mathbf{X} = (X_1, \ldots, X_n)$, llamaremos a $\mathbf{X}$ variable aleatoria conjunta o *vector aleatorio* y a cada X_i con $i = 1, \ldots, n$ variable aleatoria marginal.

Definición 6 (Probabilidad inducida)**.** Sea $(\Omega_1, \mathcal{A}, P)$ un espacio de probabilidad y $(\Omega_2, \mathcal{S})$ un espacio de medida. La *probabilidad inducida* por una variable aleatoria X viene dada por la función:

$$P_X(S) = P(X^{-1}(S)), \text{ para todo } S \in \mathcal{S}.$$

Ejemplo 1. Consideramos el lanzamiento de una moneda. Los posibles resultados del experimento serán cara o cruz, los cuales serán nuestros sucesos aleatorios. Definiremos nuestra variable aleatoria como:

$$X = \begin{cases} 0, & \text{si sale cara,} \\ 1, & \text{si sale cruz.} \end{cases}$$

Definición 7 (Función de distribución)**.** La *función de distribución* acumulada de una variable aleatoria X es una función $F\colon \mathbb{R} \to [0,1]$ definida como:

$$F(x) = P(X \leq x).$$

Proposición 2. La función de distribución acumulada F asociada a la variable aleatoria X satisface las siguientes propiedades:

- $\lim_{x \to +\infty} F(x) = 1$.
- $\lim_{x \to -\infty} F(x) = 0$.
- Es creciente, es decir, si $x_1 \leq x_2$, entonces $F(x_1) \leq F(x_2)$.
- Es continua por la derecha, es decir, $\lim_{x \to a^+} F(x) = F(a^+)$.

Si la imagen de la variable aleatoria X es numerable, diremos que la variable aleatoria es *discreta* y viene descrita por la función de probabilidad p que devuelve la probabilidad de X de ser igual a cierto valor x.

Si la imagen de la variable aleatoria X es infinita no numerable, diremos que la variable aleatoria es *continua* y viene descrita por la función de densidad f que caracteriza la posibilidad relativa de que X tome un valor cercano a x.

Definición 8 (Función de probabilidad)**.** Sea X una variable aleatoria discreta con posibles valores $\{x_1, \ldots, x_n\}$ su *función de probabilidad* se define como

$$f(x) = \begin{cases} P(X = x), & \text{si } x \in \{x_1, \ldots, x_n\}, \\ 0, & \text{en otro caso.} \end{cases}$$

Definición 9 (Función de densidad)**.** Sea X una variable aleatoria continua se dice que la función integrable no-negativa f es su *función de densidad* si para todo $x \in \mathbb{R}$,

$$P(X \leq x) = \int_{-\infty}^{x} f(u)\, du.$$

2.2.1 *Distribución conjunta*

Una *distribución conjunta* es la distribución de probabilidad de la intersección de las realizaciones de dos o más variables aleatorias cualesquiera. A continuación, definiremos algunos conceptos que ya discutimos para una única variable aleatoria para el caso multivariante.

Definición 10 (Función de distribución conjunta)**.** La *función de distribución conjunta* de un vector aleatorio $\mathbf{X} = (X_1, \ldots, X_n)$ es una función $F_{\mathbf{X}} \colon \mathbb{R}^n \to [0, 1]$ definida como:

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Definición 11 (Función de probabilidad conjunta)**.** Sea un vector aleatorio discreto $\mathbf{X} = (X_1, \ldots, X_n)$ con posibles valores en el conjunto producto

$$\mathcal{X} = \Big\{ \{x_{11}, \ldots, x_{1n}\} \times \cdots \times \{x_{n1}, \ldots, x_{nn}\} \Big\}$$

su *función de probabilidad conjunta* se define como

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1, \ldots, X_n = x_n), & \text{si } (x_1, \ldots, x_n) \in \mathcal{X}, \\ 0, & \text{en otro caso.} \end{cases}$$

Definición 12 (Función de densidad conjunta)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)$ un vector aleatorio continuo se dice que la función integrable no-negativa $f_{\mathbf{X}}$ es su *función de densidad conjunta* si para todo $(x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(u_1, \ldots, u_n)\, du_1 \cdots du_n.$$

2.2.2 *Distribución marginal*

La *distribución marginal* de un subconjunto de un vector aleatorio es la distribución de probabilidad de las variables contenidas en el subconjunto. Procederemos a definir algunos conceptos que serán útiles cuando necesitemos calcular la distribución para alguna componente de $\mathbf{X}$.

Definición 13 (Función de distribución marginal)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)$ un vector aleatorio, la *función de distribución marginal* de X_1 se define como:

$$F_{X_1}(x_1) = \lim_{x_2 \to +\infty} \cdots \lim_{x_n \to +\infty} F_{\mathbf{X}}(x_1, \ldots, x_n).$$

Definición 14 (Función de probabilidad marginal)**.** Sea un vector aleatorio discreto $\mathbf{X} = (X_1, \ldots, X_n)$, la *función de probabilidad marginal* de X_1 se define como:

$$f_{X_1}(x_1) = \sum_{x_2} \cdots \sum_{x_n} f_{\mathbf{X}}(x_1, \ldots, x_n),$$

y la *función de probabilidad marginal* del subconjunto (X_1, X_2) viene dada por:

$$f_{X_1,X_2}(x_1, x_2) = \sum_{x_3} \cdots \sum_{x_n} f_{\mathbf{X}}(x_1, \ldots, x_n).$$

Definición 15 (Función de densidad marginal)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)$ un vector aleatorio continuo, la *función de densidad marginal* de X_1 se define como:

$$f_{X_1}(x_1) = \int_{x_2} \cdots \int_{x_n} f_{\mathbf{X}}(x_1, \ldots, x_n)\, dx_2 \cdots dx_n,$$

y la *función de densidad marginal* del subconjunto (X_1, X_2) viene dada por:

$$f_{X_1,X_2}(x_1, x_2) = \int_{x_3} \cdots \int_{x_n} f_{\mathbf{X}}(x_1, \ldots, x_n)\, dx_3 \cdots dx_n.$$

2.3 PROBABILIDAD CONDICIONAL

Llamamos *probabilidad condicional* a la probabilidad de que ocurra un suceso A, sabiendo que también sucede otro suceso B. Algunos resultados relacionados y que usaremos en el trabajo son el Teorema de probabilidad total y el Teorema de Bayes.

Definición 16 (Probabilidad condicional)**.** Para cualesquiera dos sucesos $A, B \in \mathcal{A}$ tales que $P(B) > 0$. Definimos la *probabilidad condicional* de A sobre B como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Teorema 1 (Teorema de la probabilidad total)**.** Sea $\{A_i : i = 1, \ldots, n\}$ una partición sobre Ω y sea $B \in \mathcal{A}$ un suceso arbitrario del que se conocen las probabilidades condicionales $P(B \mid A_i)$, entonces la probabilidad del suceso B viene dada por

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B \mid A_i).$$

Demostración. Utilizando que $\{A_i \in \mathcal{A} : i = 1, \ldots, n\}$ es una partición del espacio muestral Ω y por tanto cumple:

- $\Omega = \bigcup_{i=1}^{n} A_i$.

- $A_i \cap A_j = \emptyset$, para todo $A_i \neq A_j$.

Podemos escribir el suceso B como

$$B = \bigcup_{i=1}^{n} B \cap A_i.$$

Como los conjuntos A_i son disjuntos dos a dos, entonces los conjuntos $B \cap A_i$ también lo son. En consecuencia

$$P(B) = P(B \cap A_1) + \cdots + P(B \cap A_n).$$

Por último, como $B, A_i \in \mathcal{A}$ usando la Definición 16, tenemos

$$P(B) = P(B \mid A_1)P(A_1) + \cdots + P(B \mid A_n)P(A_n) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i).$$

□

Teorema 2 (Teorema de Bayes)**.** Para cualesquiera dos sucesos $A, B \in \mathcal{A}$ tales que $P(B) > 0$, se tiene

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Demostración. Utilizando la Definición 16, tenemos que $P(B \mid A)P(A) = P(B \cap A)$ y sabiendo que $P(B \cap A) = P(A \cap B)$ concluimos:

$$\frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \cap A)}{P(B)} = P(A \mid B).$$

□

Notación 1. A partir de este punto, denotaremos $P(A \cap B)$ como $P(A, B)$.

2.3.1 *Distribución condicional*

Una *distribución condicional* se define como la distribución de una de las variables condicionada a cada valor de una o más variables aleatorias. Definiremos el concepto de función de probabilidad y función de densidad condicional, que nos serán útiles en el futuro.

Definición 17 (Función de probabilidad condicional)**.** Sean X, Y variables aleatorias discretas con función de probabilidad conjunta $f_{X,Y}$ y sea $f_Y(y)$ la función de probabilidad marginal de Y. Llamaremos *función de probabilidad condicional* de X dado $Y = y$, a la función definida como:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad \text{con } f_Y(y) \neq 0.$$

Comentario 1. La *función de densidad condicional* se define de la misma forma para X, Y variables aleatorias continuas, tomando $f_{X,Y}$ como la función de densidad conjunta y f_Y como la función de densidad marginal.

2.4 INDEPENDENCIA

Dos variables aleatorias son *independientes* entre sí cuando la probabilidad de cada variable no está influida por la ocurrencia de la otra. A continuación formalizaremos esta idea.

Definición 18 (Variables aleatorias independientes)**.** Sean X, Y variables aleatorias, diremos que son *independientes* si, y solo si,

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Cuando esto ocurra, lo denotaremos como $X \perp Y$.

Definición 19 (Variables aleatorias independientes condicionalmente)**.** Sean X, Y, Z variables aleatorias, entonces X e Y son *condicionalmente independientes* dado Z si, y solo si,

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z),$$

o equivalentemente,

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z).$$

En este caso, lo denotaremos como $X \perp Y \mid Z$.

Definición 20 (Variables aleatorias independientes e idénticamente distribuidas)**.** Sea un vector aleatorio $\mathbf{X} = (X_1, \ldots, X_n)$ diremos que sus componentes son *independientes e idénticamente distribuidas* (i.i.d) si, y solo si, son *independientes*:

$$F_{\mathbf{X}}(x) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \text{ para todo } x_1, \ldots, x_n \in \mathbb{R}$$

y son *idénticamente distribuidas*:

$$F_{X_1}(x) = F_{X_i}(x), \text{ para todo } i \in \{2, \ldots, n\}, \text{ para todo } x \in \mathbb{R}.$$

3

DISTRIBUCIONES DE PROBABILIDAD

En este capítulo recordaremos algunos ejemplos de distribuciones de probabilidad que usaremos durante el desarrollo del trabajo y definiremos algunos conceptos matemáticos estadísticos previos a la construcción de las distribuciones. Utilizaremos como referencias bibliográficas a Cramer [2004] y Dembo [2014].

3.1 ESPERANZA DE UNA VARIABLE ALEATORIA

Estamos listos para introducir el concepto de *esperanza matemática* de una variable aleatoria X. La esperanza representa el valor promedio de los valores que toma la variable.

Definición 21 (Esperanza de una variable aleatoria)**.** Sea X una variable aleatoria unidimensional no negativa en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$. Definimos su *esperanza* como:

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, dP(\omega).$$

Se denota por μ_X o μ dependiendo de si queremos destacar o no cual es la variable aleatoria a la que se refiere.

Si X es una variable aleatoria discreta y toma valores en el conjunto $\mathcal{X}$, entonces su esperanza se define como:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x P_X(x).$$

donde x es cada posible resultado del experimento y $P_X(x)$ la probabilidad inducida por X de obtener el resultado x.

Si X es continua y $f(x)$ es su función de densidad, entonces su esperanza se define como:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x)\, dx.$$

Proposición 3. Dadas X, Y variables aleatorias y $a, b \in \mathbb{R}$. $\mathbb{E}[X]$ es un operador lineal, es decir:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Demostración. Consecuencia de la linealidad de la integral de Lebesgue. □

Proposición 4. Sean X una variable aleatoria con función de densidad f_X y $g\colon \mathbb{R} \to \mathbb{R}$ una función integrable de Lebesgue, entonces se cumplen las siguientes propiedades:

- $g(X)$ es una variable aleatoria y su esperanza viene dada por:

$$\mathbb{E}[g(X)] = \int_\Omega g(x) f_X(x)\, dx.$$

  - Si X es discreta, $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) P_X(x)$.
  - Si X es continua, $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x)\, dx$.
- Sean $X_1, \ldots, X_n$ variables aleatorias independientes, entonces

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

A continuación, definimos la *esperanza condicional* de una variable aleatoria como el valor esperado de dicha variable respecto a una distribución de probabilidad condicional.

Definición 22 (Esperanza condicional)**.** Sean X, Y variables aleatorias discretas y $f_{X|Y}$ su función de probabilidad condicional, definimos la *esperanza condicional* de X dado $Y = y$ como:

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x f_{X|Y}(x \mid y),$$

en el caso continuo, sea $f_{X|Y}$ la función de densidad condicional, la esperanza condicional se calcula como:

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y)\, dy.$$

3.2 MOMENTOS DE UNA VARIABLE ALEATORIA

A partir de la definición de esperanza de una variable aleatoria, podemos construir el concepto de *momentos* de una variable aleatoria, que nos permiten extraer información relevante de la distribución desconocida.

Definición 23 (Momento no centrado de una variable aleatoria)**.** Sea X una variable aleatoria y $k \in \mathbb{N}$. Definimos el *momento no centrado de orden k* como:

$$\mu_k = \mathbb{E}[X^k],$$

siempre que exista dicha esperanza.

Comentario 2. En el caso $k = 1$, obtenemos la esperanza matemática de la variable aleatoria X, a la cual también denominamos *media* μ_X o simplemente μ.

Definición 24 (Momento centrado de una variable aleatoria)**.** Sea X una variable aleatoria, $k \in \mathbb{N}$ y $c \in \mathbb{R}$ definimos el *momento centrado en c de orden k* como:

$$\mu_k = \mathbb{E}[(X - c)^k].$$

Definición 25 (Varianza)**.** La *varianza* de una variable aleatoria se define como:

$$\mathrm{Var}(X) = \sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

A la raíz cuadrada positiva de la varianza la denotaremos como *desviación estándar* σ_X o simplemente σ.

Proposición 5. Sean X una variable aleatoria y $a, b \in \mathbb{R}$, entonces se cumplen las siguientes propiedades:

- $\mathrm{Var}(X) = \mathbb{E}[X]^2 - \mathbb{E}[X^2]$.
- $\mathrm{Var}(b) = 0$.
- $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$.
- Sea X, Y variables aleatorias independientes, entonces:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proposición 6. Dadas X, Y variables aleatorias y $a, b \in \mathbb{R}$. $\mathrm{Var}(X)$ es una operación lineal, es decir:

$$\mathrm{Var}(aX + b) = \mathbb{E}[(aX + b) - \mathbb{E}[aX + b]^2] = a^2\mathbb{E}[(X - \mu_X)^2] = a^2\mathrm{Var}(X).$$

Demostración. Consecuencia de la linealidad de la esperanza. □

En el mundo real, cuando aplicamos estos conceptos, lo haremos sobre múltiples características observables, es decir, sobre vectores aleatorios como escribiremos a continuación.

Definición 26 (Esperanza de un vector aleatorio)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)^T$ un vector aleatorio. Se define la *esperanza* de $\boldsymbol{\mu}_\mathbf{X}$ como:

$$\boldsymbol{\mu}_\mathbf{X} = \mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix},$$

siempre que existan las esperanzas unidimensionales.

Para generalizar la varianza de una variable aleatoria, construiremos una matriz multidimensional con la que aparece el concepto de covarianza.

Definición 27 (Matriz de covarianzas)**.** Sea $\mathbf{X} = (X_1, \ldots, X_n)^T$ un vector aleatorio. Se define, la *matriz de covarianzas* de $\mathbf{X}$ como:

$$\boldsymbol{\Sigma}_\mathbf{X} = \mathrm{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})^T] = \begin{bmatrix} \sigma_{11} & \ldots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \ldots & \sigma_{nn} \end{bmatrix},$$

donde $\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ji}$ es la covarianza de las variables aleatorias X_i, X_j. Podremos definirla cuando existan todas las covarianzas.

3.3 EJEMPLOS DE DISTRIBUCIONES

La *distribución de probabilidad* de una variable aleatoria es una función que hace corresponder a cada suceso definido sobre la variable, la probabilidad de que dicho suceso ocurra. Describiremos algunas de estas funciones que nos serán de interés a lo largo del desarrollo del trabajo.

Definición 28 (Moda)**.** La *moda* de una distribución es el valor donde su función de probabilidad alcanza su máximo. Es el valor que aparece con mayor frecuencia en un conjunto de datos.

Las distribuciones pueden ser *unimodales*, *bimodales* o *multimodales* dependiendo de si tienen un solo valor de moda, dos o más, respectivamente.

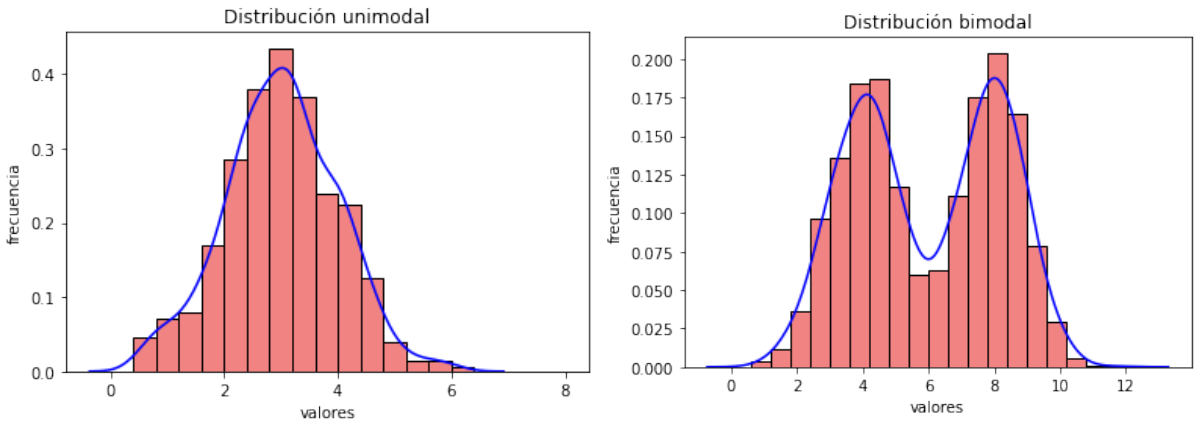


Figura 1: Ejemplos de distribuciones de probabilidad.

Distribución Bernoulli

La *distribución Bernoulli* es una distribución de probabilidad aplicada a una variable aleatoria discreta, la cual solo puede tomar dos resultados mutuamente excluyentes (éxito o fracaso).

Definición 29 (Distribución Bernoulli)**.** Una variable aleatoria unidimensional X sigue una *distribución Bernoulli* de parámetro p si su función de probabilidad viene dada por:

$$P(X = x) = \begin{cases} p, & \text{si } x = 1, \\ 1 - p, & \text{si } x = 0. \end{cases}$$

Escribiremos la distribución como $X \sim Bernoulli(p)$, donde el parámetro $p \in (0, 1)$ indica la probabilidad de éxito y $(1 - p)$ la probabilidad de fracaso del experimento.

Proposición 7 (Propiedades)**.** Si $X \sim Bernoulli(p)$, entonces la variable aleatoria X satisface las siguientes propiedades:

- $\mathbb{E}[X] = p$.
- $\text{Var}(X) = p(1 - p)$.

Distribución de Poisson

La *distribución de Poisson* es una distribución de probabilidad discreta que modela el número de sucesos raros que ocurren en un determinado periodo de tiempo.

Definición 30 (Distribución de Poisson)**.** Una variable aleatoria unidimensional X sigue una *distribución de Poisson* de parámetro λ si su función de probabilidad viene dada por:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

donde $x = 0, 1, \dots$ es el número de ocurrencias del evento o fenómeno.

Escribiremos la distribución como $X \sim Poisson(\lambda)$, donde el parámetro $\lambda > 0$ representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado.

Proposición 8. Si $X \sim Poisson(\lambda)$, entonces la variable aleatoria X satisface las siguientes propiedades:

1. $\mathbb{E}[X] = \lambda$.
2. $\mathrm{Var}(X) = \lambda$.
3. La moda de X es $\lfloor \lambda \rfloor$ (el mayor entero menor que λ).

Demostración. Demostraremos la propiedad 1 utilizando la definición de $\mathbb{E}[X]$.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=0}^{\infty} x \left(\frac{\lambda^x e^{-\lambda}}{x!} \right) \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&\overset{(*)}{=} \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda.
\end{aligned}$$

donde en $(*)$ hemos cambiado $(x-1)$ por y.

Para la propiedad 2, sabiendo que

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2. \tag{1}$$

Procederemos con el cálculo de $\mathbb{E}[X(X-1)]$.

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)\left(\frac{\lambda^x e^{-\lambda}}{x!}\right)\\ &= \lambda^2 e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!}\\ &\overset{(*)}{=} \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}\\ &= \lambda^2 e^{-\lambda}e^{\lambda}\\ &= \lambda^2.\end{aligned}$$

donde en $(*)$ hemos cambiado $(x-2)$ por y.

Sustituyendo en la Ecuación (1) concluimos que, $\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$. □

Proposición 9. Si $X \sim Poisson(\lambda)$, donde X es una variable aleatoria que representa el número de sucesos raros en una unidad de tiempo e Y es una variable aleatoria que representa el número de dichos sucesos raros en un tiempo t, se tiene que

$$Y \sim Poisson(t\lambda).$$

Ejemplo 2. En un instituto, el número medio de suspensos por clase es de 2,4. Es decir, si X es el número de suspensos por clase, entonces

$$X \sim Poisson(2{,}4).$$

¿Cuál es la probabilidad de que en una clase no haya suspensos?

$$P(X=0) = \frac{2{,}4^0 e^{-2{,}4}}{0!} = e^{-2{,}4} = 0{,}09.$$

¿Cuál es la probabilidad de que en 3 clases haya exactamente 6 suspensos?

Sea Y el número de suspensos en 3 clases. Sabemos que:

$$Y \sim Poisson(2{,}4 \cdot 3) = Poisson(7{,}2)$$

$$P(Y=6) = \frac{7{,}2^6 e^{-7{,}2}}{6!} = e^{-3{,}4} = 0{,}14.$$

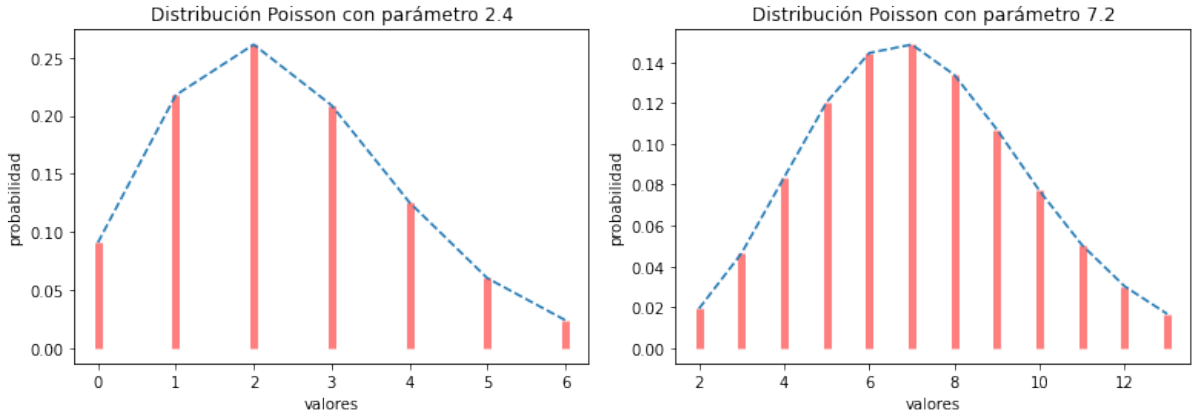


Figura 2: Distribuciones calculadas en el Ejemplo 2.

Distribución Normal

La *distribución normal* o *distribución de Gauss* se utiliza para representar variables aleatorias de valor real cuyas distribuciones son desconocidas.

Definición 31 (Distribución normal)**.** Una variable aleatoria unidimensional X sigue una *distribución normal* o *gaussiana* de parámetros μ, σ, si su función de densidad $f\colon \mathbb{R} \to \mathbb{R}$ viene dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},$$

donde $\mu, \sigma \in \mathbb{R}$.

Donde escribiremos la distribución como $X \sim \mathcal{N}(\mu, \sigma^2)$, donde recordemos el parámetro μ se refiere a la media y σ a la desviación estándar de la variable aleatoria.

Proposición 10. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces la variable aleatoria X satisface las siguientes propiedades :

- La distribución es simétrica respecto a μ.
- $P(\mu - \sigma < X < \mu + \sigma) \approx 0{,}683$.
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0{,}955$.
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0{,}997$.
- La moda de X coincide con μ.

Como consecuencia de la primera propiedad de la Proposición 10, podemos relacionar todas las variables aleatorias normales con la distribución $\mathcal{N}(0, 1)$.

Proposición 11 (Estandarización de variables aleatorias normales)**.** Sea X una variable aleatoria tal que $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1),$$

donde $\mathcal{N}(0, 1)$ es la distribución normal estándar.

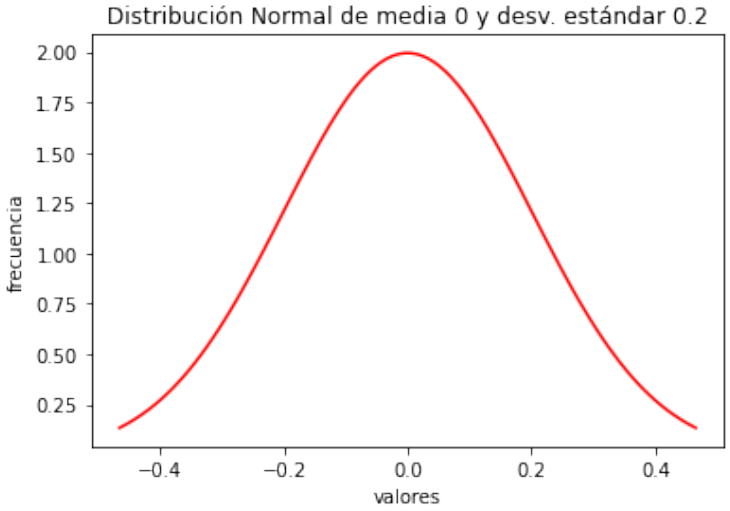


Figura 3: Ejemplo de una distribución normal.

Definición 32 (Media muestral)**.** Sean $X_1, \ldots, X_n$ variables aleatorias obtenidas a partir de X y que siguen su misma distribución, se define la *media muestral* como:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

La importancia de la distribución normal reside principalmente en el Teorema central del límite que trata sobre la distribución de la media muestral para variables aleatorias independientes e idénticamente distribuidas (i.i.d) y garantiza una distribución aproximadamente normal cuando n es lo suficientemente grande.

Teorema 3 (Teorema central del límite)**.** Sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d con media μ y desviación estándar σ^2 (ambas finitas). Con n suficientemente grande, se tiene que

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1).$$

Finalmente, podemos relacionar las dos distribuciones estudiadas aproximando la distribución de Poisson mediante la distribución normal estándar a partir de la siguiente relación:

Proposición 12. Sea X una variable aleatoria tal que $X \sim Poisson(\lambda)$ con λ suficientemente grande, entonces:

$$\frac{X-\lambda}{\sqrt{\lambda}} \sim \mathcal{N}(0,1).$$

Como una extensión del caso unidimensional, aparece el caso de distribución normal multivariante que definiremos a continuación.

Definición 33 (Distribución normal multivariante)**.** Sea un vector aleatorio $\mathbf{X} = (X_1, \ldots, X_n)^T$ diremos que sigue una *distribución normal multivariante* de parámetros $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ si su función de densidad $f\colon \mathbb{R}^N \to \mathbb{R}$ viene dada por:

$$f(x) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})},$$

donde $\boldsymbol{\mu} \in \mathbb{R}^N$ y $\boldsymbol{\Sigma} \in \mathcal{M}(\mathbb{R})$.

Escribiremos la distribución como $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde recordemos el parámetro $\boldsymbol{\mu}$ se refiere al vector de las medias de la distribución y $\boldsymbol{\Sigma}$ a la matriz de covarianzas.

Distribución Uniforme

Definición 34 (Distribución uniforme discreta)**.** Una variable aleatoria discreta X con posibles valores $\{x_1, \ldots, x_n\}$ diremos que sigue una *distribución uniforme* si:

$$P(X = x_i) = \frac{1}{n}, \text{ para todo } i = 1, \ldots, n.$$

Definición 35 (Distribución uniforme continua)**.** Una variable aleatoria continua X sigue una *distribución uniforme* en el intervalo (a, b) si su función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{si } x \in (a,b), \\ 0, & \text{si } x \notin (a,b). \end{cases}$$

Escribiremos la distribución como $X \sim U(a, b)$, donde la variable aleatoria queda definida por los extremos del intervalo, es decir, a y b son sus parámetros.

Proposición 13 (Propiedades)**.** Si $X \sim U(a, b)$, entonces la variable aleatoria X satisface las siguientes propiedades:

- $\mathbb{E}[X] = \frac{a+b}{2}$.
- $\text{Var}(X) = \frac{(b-a)^2}{12}$.
- La moda de X es cualquier valor en (a, b).

Distribución Gamma

Definición 36 (Distribución gamma)**.** Una variable aleatoria continua X sigue una *distribución gamma* de parámetros α y λ si su función de densidad viene dada por:

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}.$$

donde Γ es la *función gamma* definida como $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\, dx$.

Escribiremos la distribución como $X \sim \Gamma(\alpha, \lambda)$, donde la variable aleatoria queda definida por los parámetros $\alpha, \lambda > 0$.

Proposición 14 (Propiedades)**.** Si $X \sim \Gamma(\alpha, \lambda)$, entonces la variable aleatoria X satisface las siguientes propiedades:

- $\mathbb{E}[X] = \frac{\alpha}{\lambda}$.
- $\text{Var}(X) = \frac{\alpha}{\lambda^2}$.

Definición 37 (Distribución gamma inversa)**.** Una variable aleatoria continua X sigue una *distribución gamma inversa* de parámetros α y λ si su función de densidad viene dada por:

$$f(x) = \frac{\lambda^\alpha x^{-(\alpha+1)} e^{\frac{-\lambda}{x}}}{\Gamma(\alpha)}.$$

4

TEORÍA DE GRAFOS

En este capítulo incluiremos algunos conceptos fundamentales, que nos serán útiles en la construcción de los modelos causales, y que funcionan como base del concepto de equidad contrafactual que trataremos a lo largo del trabajo. La referencia básica de este capítulo es el libro de Godsil and Royle [2001].

4.1 GRAFOS, NODOS Y ARISTAS

Definición 38 (Par ordenado)**.** Un *par ordenado* es una pareja de elementos, en la que los elementos vienen distinguidos por su orden. El par ordenado donde el primer elemento es a y el segundo b se denota como (a, b). Si el orden no es relevante, lo llamaremos *par no ordenado* y lo denotaremos $\{a, b\}$. En este caso, es claro que $\{a, b\} = \{b, a\}$.

Definición 39 (Grafo)**.** Un *grafo* $G = (V, E)$ es un conjunto no vacío de vértices o nodos V y aristas $E \subset V \times V$ entre ellos. Si E consta de pares ordenados de vértices lo llamaremos *grafo dirigido*, si en otro caso E consta de pares no ordenados, lo llamaremos *grafo no dirigido*.

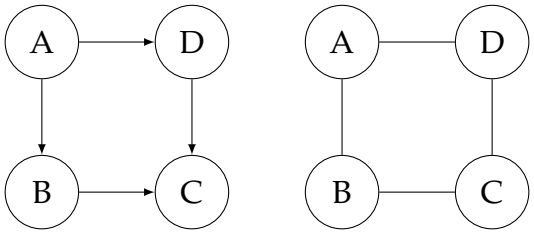


Figura 4: Ejemplo de un grafo dirigido y un grafo no dirigido, respectivamente.

Definición 40 (Camino)**.** En un grafo dirigido $G = (V, E)$, un *camino* $A \to B$ es una secuencia de vértices $\{A = v_0, v_1, \ldots, v_{n-1}, v_n = B\}$ donde $(v_i, v_{i+1}) \in E$ para todo $i \in \{0, \ldots, n-1\}$. Si G es un grafo no dirigido, $A \to B$ es un *camino* si $\{v_i, v_{i+1}\} \in E$ para todo $i \in \{0, \ldots, n-1\}$.

Definición 41 (Grafo acíclico dirigido)**.** Un *grafo acíclico dirigido* es un grafo dirigido que no tiene ciclos, es decir, que para cada nodo $v \in V$, no existe ningún camino que empiece y termine en v. Si un grafo no es acíclico lo llamaremos *ciclo*.

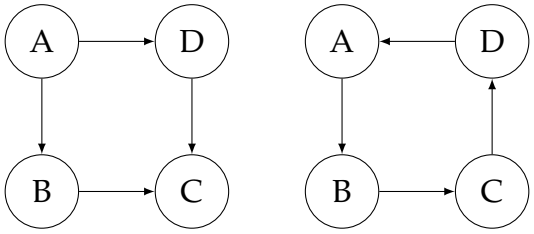


Figura 5: Ejemplo de un grafo acíclico dirigido y un ciclo dirigido, respectivamente.

Ejemplo 3. Podemos ver que el grafo de la izquierda de la Figura 5 es acíclico puesto que sea $v \in \{A, B, C, D\}$ no existe ningún camino que revisite v. Por otro lado, el grafo de la derecha es un ciclo, ya que el camino $A \to B \to C \to D \to A$ es directo y empieza y termina en A.

Definición 42 (Bucle)**.** Sea $G = (V, E)$ un grafo y $v \in V$ un vértice del mismo. Diremos que v tiene un *bucle* si $(v, v) \in E$.

Definición 43 (Grafo simple)**.** Diremos que G *grafo simple* si no posee bucles ni ninguna de sus aristas relacionan al mismo par de vértices. Llamaremos a G *multigrafo* si, y solo si, no es un grafo simple.

4.2 ESTRUCTURA DE UN GRAFO

Ahora definiremos algunas relaciones entre nodos en un grafo acíclico dirigido.

Definición 44 (Ancestros y descendientes de un nodo)**.** Sean A, B dos vértices de un grafo dirigido G. Si $A \to B$ es un camino dirigido y $B \not\to A$ (no existe un camino dirigido de B a A), entonces diremos que A es el *ancestro* de B y B es *descendiente* de A.

Ejemplo 4. En el grafo de la derecha de la Figura 5, el nodo A es un ancestro de B, D y C. Por otro lado, el nodo C es un descendiente de A, B y D.

Definición 45 (Padres e hijos de un nodo)**.** Particularizando la Definición 44, definimos a los *padres* de un nodo A como el conjunto de nodos $pa(A)$ tal que existe una arista dirigida para cada nodo de $pa(A)$ hacia A, la noción inversa de padres define el concepto de *hijos* a los que denotaremos como $hi(A)$.

Definición 46 (Vecinos de un nodo)**.** Sea un grafo G, definimos los *vecinos* de un nodo como todos los nodos directamente conectados a él. Denotaremos al conjunto de vecinos de un nodo A como $ve(A)$.

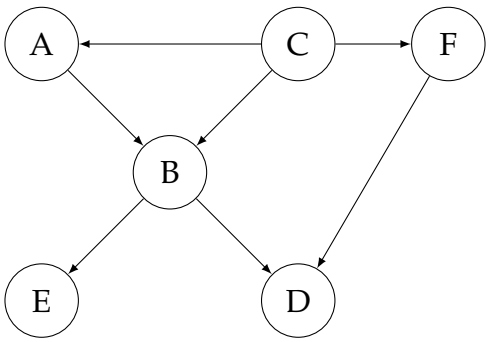


Figura 6: Ejemplo de grafo acíclico dirigido.

Ejemplo 5. Para comprender las definiciones anteriores, definiremos los conjuntos de padres, hijos y vecinos del nodo B para la Figura 6.

$$pa(B) = \{A, C\}, \quad hi(B) = \{E, D\}, \quad ve(B) = \{A, C, E, D\}.$$

Parte II

JUSTICIA EN APRENDIZAJE AUTOMÁTICO

Conceptos básicos del aprendizaje automático, formalización de las medidas de equidad y definición de los principales algoritmos de mitigación del sesgo.

5

CONCEPTOS BÁSICOS DEL APRENDIZAJE AUTOMÁTICO

En este capítulo presentaremos algunos conceptos básicos del aprendizaje automático. Definiremos el concepto de aprendizaje supervisado, discutiremos el proceso de creación de modelos predictivos a partir de un conjunto de datos dado y comentaremos las medidas básicas de evaluación de un modelo de clasificación.

5.1 ¿QUÉ ES EL APRENDIZAJE AUTOMÁTICO?

El *aprendizaje automático* o *machine learning* se encarga de extraer patrones significativos de un conjunto de datos con el objetivo de inferir en la distribución estadística subyacente (Bishop [2006]). Para ello, durante la fase de entrenamiento, aprendemos un modelo a partir de un conjunto de datos de interés seleccionado previamente. Una vez entrenado el modelo, podremos predecir o tomar decisiones sin que el modelo haya sido específicamente programado para esa tarea. Los patrones generales obtenidos a partir del entrenamiento del modelo podrán aplicarse posteriormente a datos no vistos y seguir obteniendo resultados de utilidad.

A grandes rasgos, consideramos tres tipos de algoritmos de aprendizaje automático: los de aprendizaje supervisado que actúan sobre datos etiquetados, los de aprendizaje no supervisado que actúan sobre datos no etiquetados y los de aprendizaje por refuerzo que actúan en un entorno de ensayo-error. Para nuestro trabajo, nos centraremos en el entorno de *aprendizaje supervisado*.

5.1.1 *Aprendizaje supervisado*

Definiremos los componentes del aprendizaje a partir de un ejemplo real: la aprobación de un crédito bancario. En principio, el banco no conoce ninguna fórmula ideal que pueda comunicarle cuando debe aprobar un crédito. El banco utilizará los registros de los clientes anteriores para aprender sobre ellos y encontrar una buena fórmula para la aprobación de las nuevas peticiones de créditos. Cada registro de clientes tiene información relativa al mismo, como pueden ser salario anual, años de residencia, préstamos pendientes, etc. También registra si la aprobación del crédito para ese cliente fue una buena idea, es decir, si le proporcionó o no beneficios a la

entidad financiera. Estos datos serán los que guíen la construcción de una fórmula de éxito para la aprobación del crédito que podrá utilizarse con futuros solicitantes.

A continuación, formalizaremos los principales componentes de este problema de aprendizaje. Tenemos el vector de entrada $\mathbf{x}$ que contiene la información del cliente que se utilizará para tomar la decisión del crédito, la *función objetivo desconocida* $f\colon \mathcal{X} \to \mathcal{Y}$ (fórmula ideal para la aprobación del crédito), donde $\mathcal{X}$ es el conjunto de todas las posibles características e $\mathcal{Y}$ el conjunto de todos los posibles resultados (en este caso, una decisión binaria si/no). Tenemos un *conjunto de datos* consistente en pares de entrada-salida $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ donde cada y_i viene dado por una función desconocida $y_i = f(\mathbf{x}_i)$ con $i = 1, \ldots, n$ (los valores de entrada corresponden a los antiguos clientes y el resultado será la decisión de crédito correcta para ellos en retrospectiva).

Finalmente, contamos con el *algoritmo de aprendizaje* $\mathcal{A}$ que utiliza el conjunto de datos $\mathcal{D}$ para elegir una fórmula $g\colon \mathcal{X} \to \mathcal{Y}$ que mejor aproxime a la función ideal f. El algoritmo elegirá g de entre un conjunto de fórmulas candidatas consideradas, al que denominamos *conjunto de hipótesis* $\mathcal{H}$. Por ejemplo, $\mathcal{H}$ podría ser el conjunto de todas las fórmulas lineales de las que el algoritmo elegiría la que mejor ajuste linealmente a los datos.

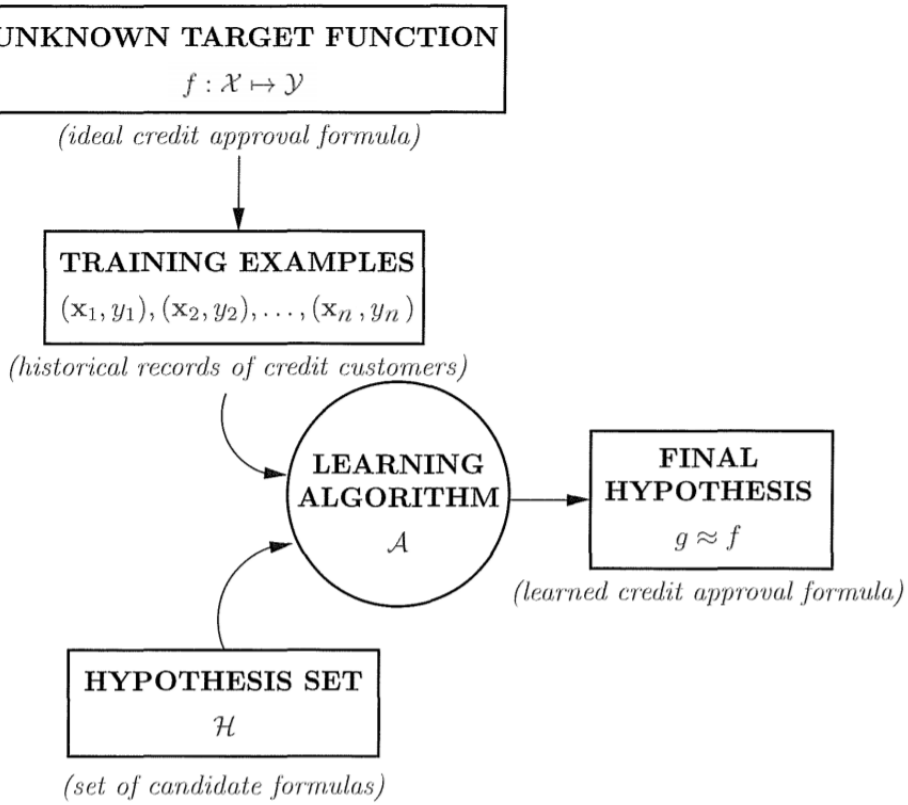


Figura 7: Configuración básica del problema de aprendizaje. (Abu-Mostafa et al. [2012])

Cuando un nuevo cliente solicite un crédito, el banco basará su decisión en g (la hipótesis que produjo el algoritmo de aprendizaje) y no en f (la función objetivo ideal) que aún sigue siendo desconocida. La decisión será buena solo en la medida en que g replique f. Para ello, el algoritmo elegirá la función g que mejor se ajuste a f en los ejemplos de entrenamiento de clientes anteriores, con la esperanza de que siga coincidiendo con f en los nuevos clientes.

5.2 PROPIEDADES DEL MODELO DE APRENDIZAJE

Definiremos algunos conceptos importantes que surgen junto a algunas cuestiones planteadas una vez creado el modelo de aprendizaje (Barocas et al. [2019]). Suponemos un conjunto de datos etiquetado $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Normalmente los datos $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ se extraen de forma independiente e idénticamente distribuida de una población $(\mathcal{X}, \mathcal{Y})$.

Definición 47 (Clasificador arbitrario)**.** Un *clasificador arbitrario* se define como una aplicación $g\colon \mathcal{X} \to \mathcal{Y}$ del conjunto de características $\mathcal{X}$ al conjunto de resultados $\mathcal{Y}$, de forma que $g(\mathbf{x})$ es el resultado predicho para el individuo $\mathbf{x}$.

Definición 48 (Función de pérdida)**.** Una *función de pérdida* es una función definida como $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ que asigna un valor real no-negativo $\ell(y', y)$ que denota el coste del valor de la predicción y' cuando la etiqueta real es y.

Definición 49 (Riesgo empírico)**.** El *riesgo empírico* de un clasificador arbitrario g con respecto a un conjunto de datos $\mathcal{D}$ dada una función de pérdida ℓ se define como:

$$R_{\mathcal{D}}(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(\mathbf{x}_i), y_i).$$

Definición 50 (Minimización del riesgo empírico)**.** La *minimización del riesgo empírico* es el problema de optimización que busca encontrar un clasificador g en una familia de funciones $\mathcal{H}$ tal que minimice el riesgo empírico,

$$\arg\min_{g \in \mathcal{H}} R_{\mathcal{D}}(g).$$

La introducción del concepto de minimización del riesgo empírico genera diversas dudas que intentaremos abordar a lo largo de este capítulo y que pueden resumirse en las siguientes tres preguntas:

1. ¿Cuál es la clase de funciones $\mathcal{H}$ que deberíamos escoger?
2. ¿Cómo resolver de manera eficiente el problema de optimización resultante?

3. ¿El clasificador encontrado tendrá el mismo rendimiento sobre los ejemplos del conjunto de entrenamiento que sobre un conjunto de datos de prueba?

Estas cuestiones están relacionadas entre sí y dan lugar a los conceptos de *representación*, *optimización* y *generalización* respectivamente.

Representación

Normalmente las aplicaciones que utilizan datos con un número relativamente pequeño de características suelen usar *modelos de predicción lineales*. Por otro lado, si los datos de entrenamiento del modelo incluyen imágenes o audio, se suelen aplicar *modelos no lineales*. Las redes neuronales por ejemplo, aplican una secuencia de transformaciones de cada tipo para obtener mejores resultados sobre el conjunto de datos de entrada.

Lo más relevante de la representación para este trabajo es conocer que la mecánica de entrenamiento de un modelo sigue siendo la misma y el modelo utilizado rara vez importa en cuestiones relativas a la equidad.

Optimización

Si nuestro objetivo es minimizar la exactitud de un clasificador, sería evidente pensar en resolver directamente el problema de minimización del riesgo empírico con respecto a la siguiente función de pérdida:

$$\ell(y', y) = \begin{cases} 1, & \text{si } y \neq y', \\ 0, & \text{si } y = y'. \end{cases}$$

El problema de usar esta función es que es difícil de optimizar. Los gradientes de la función de pérdida 0-1 toman el valor cero en todo su dominio, por lo que no podemos esperar que los métodos basados en el gradiente optimicen directamente este tipo de pérdida.

A continuación, veremos una serie de métodos de optimización diferentes que, en determinadas circunstancias, encuentran un mínimo global o local del objetivo de riesgo empírico y aproximan en cierta medida la pérdida 0-1.

- **Pérdida al cuadrado** (*squared loss*) dada por $\frac{1}{2}(y - y')^2$. La minimización empírica del riesgo con esta función de pérdida equivale a la regresión lineal por mínimos cuadrados.
- **Pérdida de bisagra** (*hinge loss*) se expresa como el máx$\{1 - yy', 0\}$. Los algoritmos SVM se refieren a la minimización empírica del riesgo con esta función junto con la regularización ℓ_2.

- **Pérdida logística** (*logistic loss*) definida como:

$$\begin{cases} -\log(\sigma(y')), & \text{si } y = 1, \\ -\log(1-\sigma(y')), & \text{si } y = -1. \end{cases}$$

  Donde $\sigma(y') = 1/(1+e^{-y'})$ es la función logística. La minimización empírica del riesgo con esta función de pérdida equivale a la regresión logística.

La elección de la función de pérdida se realizará comparando los rendimientos de las diferentes aproximaciones mediante prueba y error, eligiendo la que mejor funcione en cada caso.

Generalización

La generalización en el aprendizaje automático hace referencia a cómo de bueno es un modelo predictivo que etiqueta correctamente datos con los que ha entrenado previamente realizando la misma tarea sobre un nuevo conjunto de datos que sigue la misma distribución de la que se extrajeron los datos de entrenamiento.

No obstante, incluso los modelos más avanzados suelen funcionar peor cuando los datos de prueba se extraen de una distribución que difiere ligeramente de la seguida por los datos de entrenamiento. Un ejemplo de ello fue el caso de la creación de un nuevo conjunto de pruebas para la base de datos ImageNet (Recht et al. [2019]).

5.3 CREACIÓN DE MODELOS DE APRENDIZAJE

Dado el esquema de configuración del problema de aprendizaje supervisado de la Figura 7 discutiremos la creación de un modelo simple de aprendizaje (Abu-Mostafa et al. [2012]). Sea $\mathcal{X} = \mathbb{R}^d$ el espacio de entrada, donde $\mathbb{R}^d$ es el espacio euclídeo d-dimensional y sea $\mathcal{Y} = \{-1, 1\}$ el espacio de salida denotando una decisión binaria (si/no). En el ejemplo de concesión de crédito, las diferentes coordenadas del vector $\mathbf{x} \in \mathcal{X}$ corresponden a los datos relativos del individuo que solicita el crédito. La salida binaria y corresponde a la aprobación o denegación del préstamo. Especificamos el conjunto de hipótesis $\mathcal{H}$ mediante una forma funcional común a todas las hipótesis $h \in \mathcal{H}$. La forma funcional $h(\mathbf{x})$ elegida, asigna pesos diferentes a cada coordenada del vector $\mathbf{x}$, reflejando su importancia en la decisión del problema. Las coordenadas ponderadas se combinan para formar una puntuación y el resultado se compara con un valor umbral previamente establecido. Si el solicitante supera el umbral, el crédito es aprobado, si no, es denegado:

$$\text{Aprobar crédito si } \sum_{i=1}^{d} w_i x_i > \text{ umbral,}$$

$$\text{Denegar crédito si } \sum_{i=1}^{d} w_i x_i < \text{ umbral.}$$

Esta fórmula se puede escribir de forma más compacta como

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) + b\right), \tag{2}$$

donde $x_1, \ldots, x_d$ son los componentes del vector $\mathbf{x}$; $h(\mathbf{x}) = 1$ significa la aprobación del crédito y $h(\mathbf{x}) = -1$ significa la denegación del crédito; $\text{sign}(s) = 1$ si $s > 0$ y $\text{sign}(s) = -1$ si $s < 0$. Los pesos $w_1, \ldots, w_d$ y el umbral viene determinado en términos del sesgo b, el crédito se aprueba si $\sum_{i=1}^{d} w_i x_i > -b$.

Uno de los ejemplos más comunes en la literatura del aprendizaje automático es el del *perceptrón* introducido por el neurocientífico e informático teórico Rosenblatt [1957]. Este algoritmo de aprendizaje intenta encontrar $\mathcal{H}$ buscando los pesos y el sesgo que funcionen bien en el conjunto de datos. Algunos de los pesos $w_1, \ldots, w_d$ podrían acabar siendo negativos, teniendo un efecto adverso en la aprobación del crédito. Por ejemplo, el peso del campo relativo a deudas pendientes debería ser negativo, ya que una mayor deuda no es buena señal para la aprobación de un crédito. El valor del sesgo b podría terminar siendo muy grande o muy pequeño, reflejando lo tolerante o estricto que debe ser el banco a la hora de conceder créditos. La elección óptima de los pesos y el sesgo define la hipótesis final $g \in \mathcal{H}$ que produce el algoritmo.

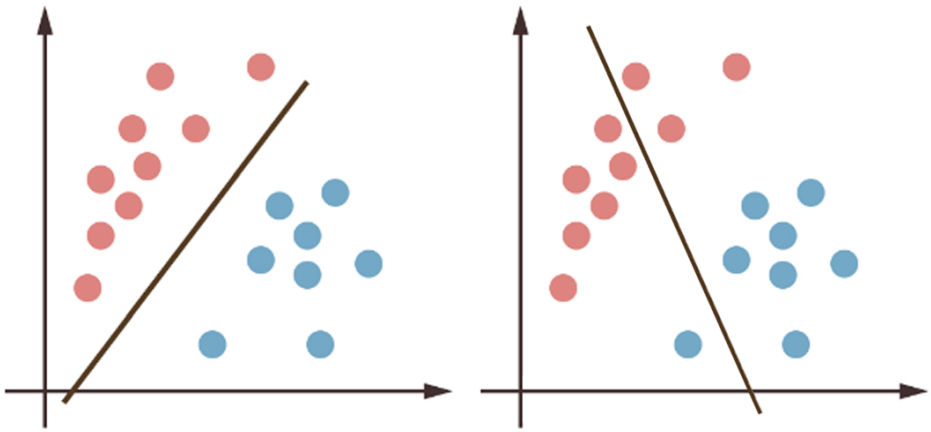


Figura 8: Clasificación de datos linealmente separables en un espacio bidimensional.

La Figura 8 ilustra dos ejemplos de clasificación de un perceptrón en un espacio de entrada bidimensional ($d = 2$). En el dibujo de la izquierda los ejemplos están perfectamente clasificados mientras que en el de la derecha hay algunos ejemplos mal clasificados. Valores diferentes para los parámetros b, w_1, w_2 dan lugar a diferentes rectas $w_1 x_1 + w_2 x_2 + b = 0$. Si el conjunto de datos es linealmente separable existirá una elección de parámetros que clasifica todos los ejemplos correctamente.

5.3.1 *Ejemplo: Perceptrón*

Introduciremos el algoritmo de del perceptrón como un ejemplo simple de un modelo de aprendizaje. Para simplificar la notación de la fórmula del perceptrón, trataremos el sesgo b como un peso $w_0 = b$ y lo añadiremos como una coordenada más al vector de pesos $\mathbf{w} = (w_0, w_1, \ldots, w_d)^T$. Además añadimos una coordenada $x_0 = 1$ al vector $\mathbf{x} = (x_0, x_1, \ldots, x_d)^T$. Observar que tratamos a $\mathbf{x}$ y $\mathbf{w}$ como vectores columna. Formalmente denotaremos el espacio de entrada como:

$$\mathcal{X} = \{1\} \times \mathbb{R}^d = \{\mathbf{x} = (x_0, x_1, \ldots, x_d)^T \ : \ x_0 = 1, \ x_1, \ldots, x_d \in \mathbb{R}\}.$$

Denotando $\mathbf{w}^T\mathbf{x} = \sum_{i=0}^{d} w_i x_i$, podemos reescribir la Ecuación 2 como:

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x}).$$

El algoritmo determinará el valor de $\mathbf{w}$ basado en los datos. Supondremos que el conjunto de datos es linealmente separable, lo que significa que podemos encontrar un vector $\mathbf{w}$ tal que $h(\mathbf{x})$ consigue una decisión correcta $h(\mathbf{x}_n) = y_n$ para todos los ejemplos del conjunto de datos de entrenamiento.

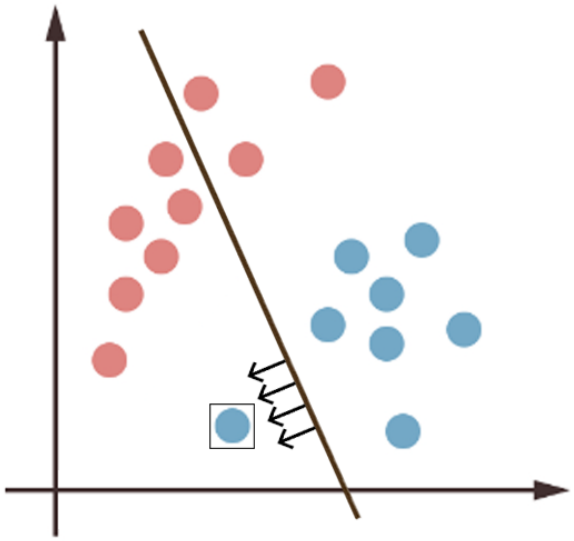


Figura 9: Esquema de actualización del algoritmo del perceptrón.

Nuestro algoritmo de aprendizaje encontrará este w usando un simple método iterativo, que funciona de la siguiente forma:

En cada paso $t = 0, 1 \ldots$, hay un valor actual del vector de pesos $\mathbf{w}_t$, seleccionamos aleatoriamente un índice $i \in \{1, \ldots, n\}$ correspondiente a un ejemplo actualmente mal clasificado, lo denota como $(\mathbf{x}_t, y_t)$ y lo usa para actualizar el valor de $\mathbf{w}_t$. Como el

ejemplo está mal clasificado, tenemos que $y_t \neq \text{sign}(\mathbf{w}_t^T \mathbf{x}_t)$. La regla de actualización viene dada por la siguiente expresión:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_t \mathbf{x}_t.$$

Esta regla mueve la frontera con el objetivo de cambiar la dirección en la clasificación correcta de $\mathbf{x}_t$, como se puede ver en la Figura 9. El algoritmo continua con las sucesivas iteraciones hasta que no haya ejemplos mal clasificados en el conjunto de datos.

Aunque la regla de actualización solo considera un único ejemplo del conjunto de entrenamiento y podría desordenar la clasificación para otros ejemplos no implicados en la iteración actual, el algoritmo garantiza una solución óptima (véase Teorema 1 en Collins [2012]). El resultado se mantiene de forma independiente al ejemplo que elijamos y a la inicialización del vector de pesos al comienzo del algoritmo. Por simplicidad, elegiremos un ejemplo mal clasificado aleatorio e inicializaremos w_0 como un vector de ceros.

Otra definición del algoritmo

Podemos definir el algoritmo de perceptrón como una instancia de la minimización del riesgo empírico (Barocas et al. [2019]). Por la descripción del algoritmo, buscamos un separador lineal y por tanto, nuestro conjunto de hipótesis corresponde con la clase de funciones lineales

$$\mathcal{H} = \{f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle \ : \ \mathbf{w} \in \mathbb{R}^d\}.$$

Utilizaremos el método de gradiente estocástico como base del algoritmo, ya que es un método de optimización que elige un ejemplo aleatorio en cada paso y realiza una actualización de los parámetros del modelo. Se define con la regla siguiente:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(f(\mathbf{x}_t), y_t).$$

Donde $\nabla \ell(f(\mathbf{x}_t), y_t)$ es el gradiente de la función de pérdida con respecto a los parámetros del modelo $\mathbf{w}_t$ para un ejemplo $(\mathbf{x}_t, y_t)$. El escalar $\eta > 0$, es un parámetro denominado *tamaño de paso*, normalmente referido a una constante pequeña.

Consideramos ahora la función de pérdida

$$\ell(y, \langle \mathbf{w}, \mathbf{x} \rangle) = \text{máx}(1 - y\langle \mathbf{w}, \mathbf{x} \rangle, 0),$$

donde su gradiente se define como:

$$\nabla \ell(y, \langle \mathbf{w}, \mathbf{x} \rangle) = \begin{cases} -y\mathbf{x}, & \text{si } y\langle \mathbf{w}, \mathbf{x} \rangle < 1, \\ 0, & \text{si } y\langle \mathbf{w}, \mathbf{x} \rangle > 1. \end{cases}$$

La expresión anterior define una parte de la regla de actualización del perceptrón, la otra parte la deduciremos al añadir la penalización $\frac{\alpha}{2}\|\mathbf{w}\|^2$ (donde $\|\cdot\|$ denota la norma euclídea) a la función de pérdida para impedir que los pesos tomen valores por encima de un umbral establecido. Esta penalización se denomina *regularización* ℓ_2 o *regularización de Tíjonov* y su propósito es fomentar la generalización.

Sumando las dos funciones de pérdida, obtenemos una expresión del riesgo empírico regularizado por ℓ_2 para la función de pérdida de bisagra (*hinge loss*).

$$R_{\mathcal{D}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \text{máx}(1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle, 0) + \frac{\alpha}{2}\|\mathbf{w}\|^2 .$$

Definiremos el algoritmo perceptrón como la solución a este problema de minimización del riesgo empírico utilizando el método del gradiente estocástico.

5.3.2 *Regresión lineal*

Volviendo al ejemplo del apartado 5.1.1, recordemos que el banco tiene un registro de clientes con variables que pueden ser usadas para aprender un clasificador lineal de decisión para la aprobación del crédito. En este caso, en lugar de limitarnos a tomar una decisión binaria (aprobar o no el crédito), en el caso de aprobación, podríamos querer establecer un umbral en el valor del crédito concedido. Esta tarea, puede ser automatizada haciendo uso del aprendizaje por *regresión* (Abu-Mostafa et al. [2012]).

El banco parte de un conjunto de datos $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, donde $\mathbf{x}_n$ es la información del cliente e y_n es el límite de crédito establecido por uno de los expertos del banco. Ahora y_n será un número real en lugar de un valor binario. El banco querrá usar un modelo de aprendizaje para encontrar una hipótesis g que replique la actuación humana en el umbral de concesión del crédito. En este caso, no buscaremos una función determinista $y = f(x)$, sino que supondremos que la etiqueta y_n proviene de una distribución $P(y \mid \mathbf{x})$. No obstante, la naturaleza del problema sigue siendo la misma, tenemos una distribución desconocida $P(\mathbf{x}, y)$ que genera cada $(\mathbf{x}_n, y_n)$ y queremos encontrar una hipótesis g que minimice el error entre $g(\mathbf{x})$ e y con respecto a esa distribución.

Algoritmo de regresión lineal

El algoritmo se basa en la minimización del riesgo empírico para la pérdida al cuadrado entre $h(\mathbf{x})$ e y. El problema es equivalente a minimizar la siguiente función:

$$R_{\mathcal{D}}(h) = \frac{1}{n}\sum_{i=1}^{n}(h(\mathbf{x}_i) - y_i)^2.$$

En regresión lineal, h se expresa como combinación lineal de las componentes de $\mathbf{x}$:

$$h(\mathbf{x}) = \sum_{i=0}^{d} w_i x_i = \mathbf{w}^T \mathbf{x},$$

donde $x_0 = 1$ y $\mathbf{x} \in \{1\} \times \mathbb{R}^d$ y $\mathbf{w} \in \mathbb{R}^{d+1}$. Para el caso lineal se suele definir una matriz de representación para $R_{\mathcal{D}}(h)$. La matriz de datos $X \in \mathbb{R}^{N \times (d+1)}$ tiene como filas los vectores $\mathbf{x}_n$ e $\mathbf{y} \in \mathbb{R}^n$ como vector columna cuyas componentes son los valores objetivo y_n. Podemos expresar el error en función de $\mathbf{w}$, X e $\mathbf{y}$ como:

$$\begin{aligned} R_{\mathcal{D}}(\mathbf{w}) &= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 \\ &= \frac{1}{n} \|X\mathbf{w} - \mathbf{y}\|^2 \\ &= \frac{1}{n} (\mathbf{w}^T X^T X \mathbf{w} - 2\mathbf{w}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}). \end{aligned} \tag{3}$$

Nuestro problema se reduce a minimizar la función dada por:

$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} R_{\mathcal{D}}(\mathbf{w}). \tag{4}$$

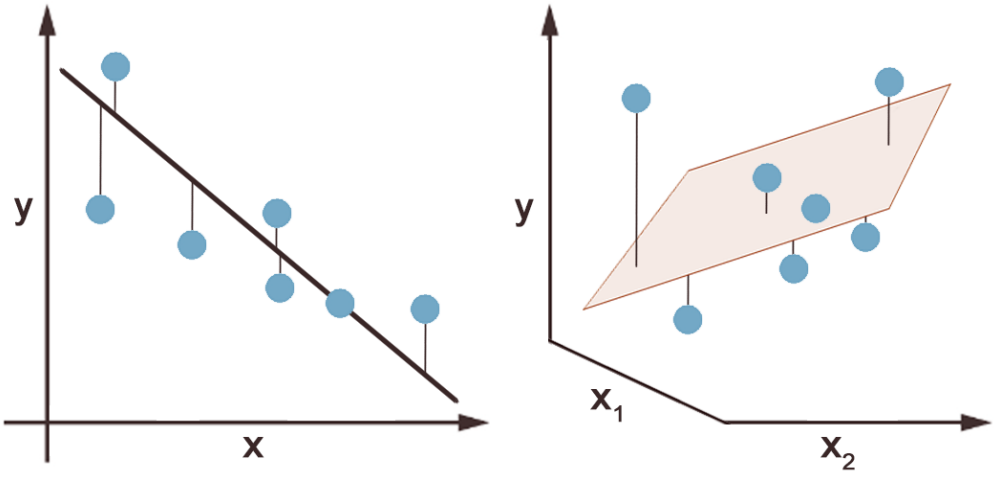


Figura 10: Problema de regresión lineal en una y dos dimensiones respectivamente.

La Figura 10 ilustra la solución para los casos unidimensional y bidimensional respectivamente. La Ecuación 3 implica que $R_{\mathcal{D}}(\mathbf{w})$ es diferenciable, por lo que podemos encontrar el mínimo de esta función igualando su gradiente a cero.

$$\nabla R_{\mathcal{D}}(\mathbf{w}) = \frac{2}{n} (X^T X \mathbf{w} - X^T \mathbf{y}) = 0.$$

Para encontrar la solución se debe cumplir que $X^T X \mathbf{w} = X^T \mathbf{y}$. Si $X^T X$ es regular, $\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$ es la única solución óptima para la Ecuación 4. En otro caso, existirá una solución óptima, pero no será única.

5.4 EVALUACIÓN EN APRENDIZAJE AUTOMÁTICO

Un modelo en aprendizaje automático funciona bien, cuando puede predecir resultados correctos a partir de un conjunto de datos de entrada no conocido previamente. La manera de medir este fenómeno, no es única, por lo que existen una gran cantidad de *métricas de rendimiento* en el ámbito del *machine learning*. La elección de la métrica dependerá del problema específico, su dominio y de sus restricciones al mundo real.

Medidas para clasificación

Una matriz de confusión define un modelo de predicción a partir de las dimensiones de los *valores reales de las etiquetas* y de sus posibles *predicciones*, resumiendo el número de predicciones correctas e incorrectas por clase.

		Etiqueta real	
		$y = 1$	$y = -1$
Predicción	$\hat{y} = 1$	Verdadero Positivo (TP)	Falso Positivo (FP) (Error tipo I)
	$\hat{y} = -1$	Falso Negativo (FN) (Error tipo II)	Verdadero Negativo (TN)

Tabla 2: Matriz de confusión: ilustra la relación entre la etiqueta real y la predicción.

Una matriz de confusión binaria como la de la Tabla 2, nos ofrece información sobre el número de *verdaderos positivos* (TP) ($y = 1 \wedge \hat{y} = 1$), *falsos positivos* (FP) ($y = -1 \wedge \hat{y} = 1$), *falsos negativos* (FN) ($y = 1 \wedge \hat{y} = -1$) y *verdaderos negativos* (TN) ($y = -1 \wedge \hat{y} = -1$). A partir de ellos, podemos obtener el número total de positivos predichos (TP+FP), el número total de negativos predichos (TN+FN), el número total de etiquetados positivos (TP+FN), y el número total de etiquetados negativos (TN+FP).

Además, podemos definir las siguientes métricas de clasificación avanzadas construidas a partir de combinaciones lineales de las básicas:

- **Tasa de verdaderos positivos** (*Recall*),

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

- **Tasa de falsos negativos**,

$$\text{FNR} = 1 - \text{TPR} = \frac{\text{FN}}{\text{TP+FN}}.$$

- **Tasa de verdaderos negativos** (*Specificity*),

$$\text{TNR} = \frac{\text{TN}}{\text{TN+FP}}.$$

- **Tasa de falsos positivos**,

$$\text{FPR} = 1 - \text{TNR} = \frac{\text{FP}}{\text{TN+FP}}.$$

- **Valor positivo predictivo** (*Precision*),

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}}.$$

- **Tasa de falso descubrimiento**,

$$\text{FDR} = 1 - \text{PPV} = \frac{\text{FP}}{\text{TP+FP}}.$$

- **Valor negativo predictivo**,

$$\text{NPV} = \frac{\text{TN}}{\text{TN+FN}}.$$

- **Tasa de falsa omisión**,

$$\text{FOR} = 1 - \text{NPV} = \frac{\text{FN}}{\text{TN+FN}}.$$

- **Exactitud** (*Accuracy*),

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}.$$

<table>
<tr><td></td><td></td><td colspan="2">Etiqueta real</td><td></td><td></td></tr>
<tr><td></td><td></td><td>$y = 1$</td><td>$y = -1$</td><td></td><td></td></tr>
<tr><td rowspan="2">Predicción</td><td>$\hat{y} = 1$</td><td>Verdadero Positivo (TP)</td><td>Falso Positivo (FP)</td><td>Precision $\frac{TP}{TP+FP}$</td><td>Tasa de Falso Descubrimiento $\frac{FP}{TP+FP}$</td></tr>
<tr><td>$\hat{y} = -1$</td><td>Falso Negativo (FN)</td><td>Verdadero Negativo (TN)</td><td>Tasa de Falsa Omisión $\frac{FN}{TN+FN}$</td><td>Valor Negativo Predictivo $\frac{TN}{TN+FN}$</td></tr>
<tr><td></td><td></td><td>Recall $\frac{TP}{TP+FN}$</td><td>Tasa de Falsos Positivos $\frac{FP}{TN+FP}$</td><td>Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$</td><td></td></tr>
<tr><td></td><td></td><td>Tasa de Falsos Negativos $\frac{FN}{TP+FN}$</td><td>Specificity $\frac{TN}{TN+FP}$</td><td></td><td></td></tr>
</table>

Tabla 3: Matriz de confusión con métricas de evaluación avanzadas.

Normalmente mediremos el rendimiento de un modelo con una de las métricas mencionadas previamente o, en su defecto, con una combinación de ambas. Una de las medidas de rendimiento más usadas es el *valor positivo predictivo* o *precisión*, que mide el porcentaje de predicciones correctas realizas. Sin embargo, existen problemas derivados de utilizar la precisión a la hora de medir el rendimiento cuando abordamos problemas que contienen desequilibrio entre las clases del conjunto de datos. Por ejemplo, si solo 5 de cada 100 muestras es positiva, un modelo trivial que siempre prediga la clase negativa tendrá una precisión del 95 %.

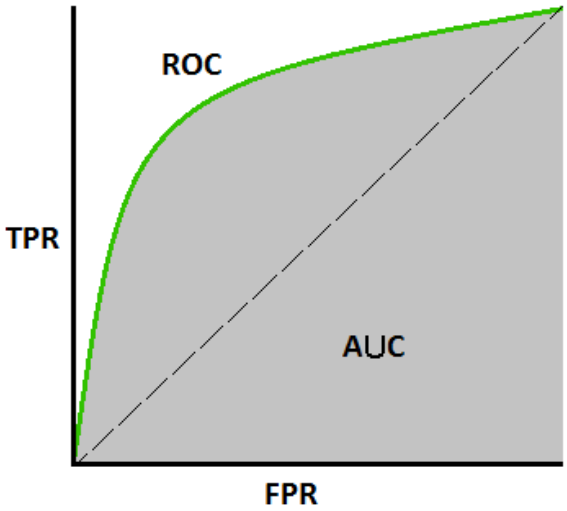


Figura 11: Ejemplo de un gráfico de curva ROC.

Otras métricas comúnmente utilizadas son la F_1-*score* la cual podemos calcularr como $\text{F}_1\text{-score} = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}}$ y el *área bajo la curva ROC* (AUC). La curva ROC se representa en un gráfico bidimensional, en el que la tasa de verdaderos positivos (TPR) se dibuja en el eje vertical y la tasa de falsos positivos (FPR) en el eje horizontal. Como podemos ver en la Figura 11, diferentes umbrales de clasificación corresponden a diferentes puntos en el espacio ROC.

El área bajo la curva ROC (AUC) se interpreta como la probabilidad de que el modelo clasifique más alto un ejemplo positivo aleatorio más que uno negativo también aleatorio. Al ser una probabilidad, el AUC oscilará entre los valores 0 y 1. Un modelo cuyas predicciones sean un 100 % incorrectas tendrá un AUC de 0, mientras que otro cuyas predicciones sean un 100 % correctas tendrá un AUC de 1. También es importante saber que el AUC es invariable con respecto a la escala y con respecto al umbral de clasificación por lo que es una buena métrica de evaluación para cualquier tipo de problemas en este ámbito.

Medidas para regresión

Definiremos el *error cuadrático medio* (ECM) y la raíz del *error cuadrático medio* (RMSE) como medidas de uso frecuente para evaluar la precisión de un modelo de regresión.

Proposición 15. Sea $\hat{\mathbf{y}}$ un vector de n predicciones e $\mathbf{y}$ el vector con las etiquetas reales de las mismas, entonces podemos calcular las medidas de de evaluación anteriores como:

$$\text{ECM} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2.$$

$$\text{RMSE} = \sqrt{\text{ECM}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2}.$$

Finalmente definiremos otro posible criterio de evaluación para un modelo estadístico cuyo principal propósito sea predecir futuros resultados o probar una hipótesis.

Definición 51 (Coeficiente de determinación)**.** El *coeficiente de determinación* (R^2) es un estadístico que determina la calidad de un modelo para predecir resultados y medir la variación que puede ser explicada por el mismo. En regresión lineal, sean X, Y dos variables aleatorias, el coeficiente R^2 se calcula como:

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}.$$

Proposición 16. Sea $\hat{\mathbf{y}}$ un vector de n predicciones e $\bar{\mathbf{y}}$ la media del vector con las etiquetas reales, podemos calcular su coeficiente de determinación a partir de la siguiente expresión:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^2}{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})^2}.$$

6

FORMALIZACIÓN DE LAS MEDIDAS DE EQUIDAD

En este capítulo presentaremos el concepto de equidad, realizaremos un análisis de las distintas nociones de equidad conocidas y formalizaremos sus definiciones.

6.1 ¿QUÉ ES LA EQUIDAD?

Con el aumento de los métodos de toma de decisiones automatizadas en la actualidad, la necesidad de satisfacer *equidad* en los modelos de *machine learning* ha cobrado importancia. Por ello, cabe hacerse las siguientes preguntas: ¿Qué es la equidad? ¿Cómo podemos medirla? ¿Y, cómo podemos fomentarla en nuestros algoritmos? En esta sección se intentará dar respuesta a estas preguntas.

En la Sección 5.1, formalizamos el proceso de aprendizaje automático sobre un ejemplo de aprobación de créditos bancarios. Hemos visto que existe un proceso de aprendizaje sobre registros históricos de datos que nos aportan información a la hora de predecir nuevos ejemplos, pero: ¿Existirán atributos que discriminen a un grupo determinado de la población?, ¿Dos clientes con características similares recibirán la misma predicción? Estas preguntas son las motivaron el concepto de equidad.

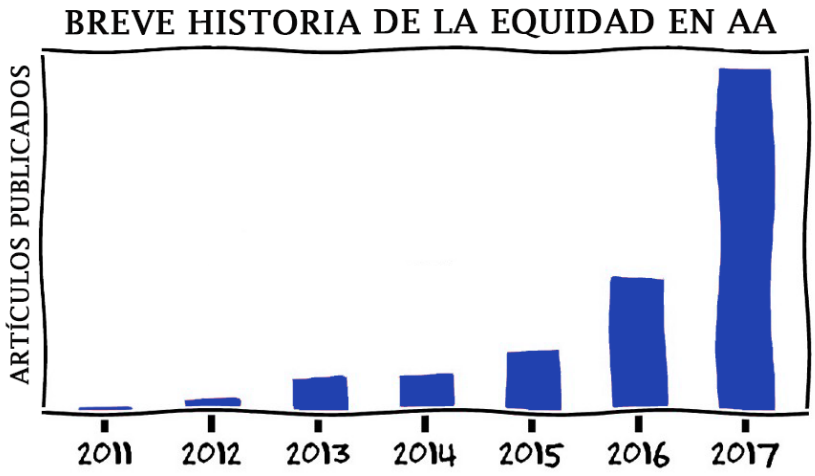


Figura 12: Incremento de las publicaciones sobre equidad entre 2011 y 2017.

Podemos dividir los trabajos sobre equidad en *machine learning* en, detectar el sesgo y discriminación en los modelos (Dwork et al. [2011]) y mitigar el sesgo algorítmico (Corbett-Davies et al. [2017]). Para estas tareas, al igual que en cualquier ciencia experimental, debemos ser capaces de medir el concepto partiendo de una definición teórica. La equidad es un concepto inherentemente subjetivo y que depende en gran medida del ámbito en el que lo apliquemos. Por lo tanto, a partir de conceptos de la literatura de la ciencias sociales, se han ido proponiendo diferentes medidas y formalizando estos conceptos para que puedan aplicarse al aprendizaje automático.

La primera idea es buscar apoyo legal y comprobar si existe alguna definición que pueda utilizarse para formular la equidad matemáticamente. Existen leyes en la mayoría de países desarrollados, que prohíben el trato desigual entre personas en función de atributos sensibles tales como el sexo o la raza (Title VII of the Civil Rights Act: Equal Employment Opportunities). Estas leyes suelen evaluar la imparcialidad de un proceso de toma de decisiones utilizando dos nociones distintas (Barocas and Selbst [2016]): el tratamiento dispar y el impacto dispar. Un proceso de toma de decisiones sufrirá un trato dispar si basamos su juicio en el atributo sensible del sujeto, y tendrá un impacto dispar si sus resultados perjudican (o benefician) de forma desigual a personas con valores de atributos sensibles diferentes.

6.1.1 *Principales familias de las medidas de equidad*

Los conceptos anteriores son demasiado abstractos como para tener una formulación cuantitativa directa por lo que, poco a poco, se han ido añadiendo numerosas definiciones de equidad a la literatura del aprendizaje automático. Sin embargo, la Tabla 1 recoge los criterios más importantes, los cuales han sido previamente recopilados en trabajos como Gajane and Pechenizkiy [2018] o Verma and Rubin [2018].

En este capítulo, nos centraremos en las medidas de equidad relativas a *impacto y tratamiento dispar*, estableciendo a su vez una división más específica en algunas de ellas. Los conceptos que trataremos a lo largo de este capítulo son:

- **Equidad por desconocimiento**.
- **Equidad individual**.
- **Equidad de grupo**: formalizaremos la paridad demográfica, el criterio de probabilidades igualadas y la tasa de paridad predictiva.
- **Medidas causales**: analizaremos su base matemática y desarrollaremos un ejemplo práctico de su aplicación en la Parte IV de este trabajo.

La equidad basada en *preferencias* está fuera del marco de este trabajo, por lo que no será explicada en profundidad. Para el lector interesado en este criterio, puede consultar el artículo Zafar et al. [2017c].

6.1.2 *Medición de la parcialidad y la equidad*

Consideramos una tarea de la clasificación estándar en la que el objetivo es predecir una variable de resultado binaria $y \in \mathcal{Y}$, utilizando un vector de variables de entrada $\mathbf{x} \in \mathcal{X}$ que sigue una distribución de probabilidad $P_{\mathbf{x}}$.

Sea un clasificador arbitrario $g : \mathcal{X} \to \mathcal{Y}$ con $\mathcal{X} = \mathbb{R}^d$, y donde $\mathcal{Y} = [0, 1]$ si produce una probabilidad predicha (por ejemplo, regresión logística), o $\mathcal{Y} = \{-1, 1\}$ si el clasificador produce un resultado predicho (por ejemplo, SVM). A lo largo del proyecto, normalmente asumiremos que el espacio de salida trabaja sobre decisiones binarias (si/no), es decir, $\mathcal{Y} = \{-1, 1\}$.

Muchos de los problemas en los que aparece el concepto de equidad pueden formularse como problemas estadísticos de evaluación de riesgos en los que asignamos una puntuación de valor real $s \in [0, 1]$ a cada individuo del conjunto de datos y se toma una decisión $\hat{y}$ basada en la puntuación, normalmente seleccionando un número predefinido de entidades que deben clasificarse como positivas.

Las principales definiciones que usaremos en este capítulo son las siguientes:

- **Vector de características -** $\mathbf{x} \in \mathcal{X}$ es un vector de características reales que identifican a un individuo.
- **Puntuación -** $s \in [0, 1]$ es una puntuación de valor real asignada a cada entidad por el clasificador.
- **Predicción -** $\hat{y} \in \mathcal{Y}$ es una predicción binaria asignada a un individuo determinado, basada en el umbral de la puntuación.
- **Etiqueta real -** $y \in \{-1, 1\}$ es la etiqueta binaria que representa el valor real de un individuo en concreto.

6.2 EQUIDAD POR DESCONOCIMIENTO

La *equidad por desconocimiento* se basa en eliminar los atributos sensibles para todos los individuos en el proceso de predicción. Algunos clasificadores propuestos en la literatura de aprendizaje automático satisfacen esta medida (Dwork et al. [2011]) debido a que es intuitiva y muy fácil de aplicar.

Notación 2. Sea $\Delta = \{\pi(\mathbf{x}) \colon \mathbb{R}^d \to \mathbb{R}^m \ : \ \mathbf{x} \in \mathcal{X},\ 1 \leq m \leq d\}$ el conjunto formado por las proyecciones del conjunto de individuos a todas las posibles dimensiones menores o iguales que la dimensión de $\mathcal{X}$. Sea $\mathcal{A} \subset \Delta$ un subconjunto que contiene todos los individuos de $\mathcal{X}$ a los que se ha aplicado una proyección π sobre ciertas características. Notaremos como $\mathcal{A}_i$ al elemento de $\mathcal{A}$ que contiene las características del individuo $\mathbf{x}_i \in \mathcal{X}$ que cierta proyección π ha seleccionado.

Definición 52 (Equidad por desconocimiento)**.** Sea $\mathcal{A} \subset \Delta$ donde $\mathcal{A}_i$ contiene las proyecciones sobre los atributos sensibles para el individuo $\mathbf{x}_i$ y sean $g\colon \mathcal{X} \to \mathcal{Y}$ y $h\colon \mathcal{X}\backslash\mathcal{A} \to \mathcal{Y}$ dos clasificadores arbitrarios. Diremos que g logra *equidad por desconocimiento* si, y solo si,

$$g(\mathbf{x}_i) = h(\mathbf{x}_i\backslash\mathcal{A}_i), \text{ para todo } \mathbf{x}_i \in \mathcal{X}.$$

Uno de los principales problemas de la equidad por desconocimiento es que no da una condición suficiente para evitar la discriminación, ya que puede haber muchas características altamente correlacionadas (por ejemplo, la zona de residencia) que funcionen como sustitutos del atributo sensible (por ejemplo, la raza). Por lo tanto, no bastaría con eliminar el atributo sensible para eliminar las disparidades. Además, se han documentado diversos ejemplos de equidad por desconocimiento para la raza; en ámbitos como educación, concesión de préstamos o justicia penal y se ha demostrado que, a largo plazo, el enfoque ciego de la raza es menos eficaz que el enfoque consciente de la misma (Fryer et al. [2008]).

Las críticas anteriores cuestionan la idoneidad de la equidad por desconocimiento en los dominios en los que, los atributos sensibles pueden deducirse a partir de los atributos no sensibles disponibles, y tenemos conocimiento de la existencia de barreras estructurales, que obstaculizan a los grupos desfavorecidos a partir de encuestas verosímiles sobre los grupos demográficos.

Ejemplo 6. Supongamos un modelo utilizado para aprobar o denegar créditos bancarios. Por sesgos históricos, sabemos que uno de los atributos sensibles en la concesión de préstamos, es la raza. Procedemos entonces a eliminar esta información de todos los individuos en el modelo de predicción. El problema surge cuando notamos que el código postal, otro atributo presente en el vector de características, está altamente correlacionado con la raza y por tanto, las decisiones basadas en este serán racialmente discriminatorias. En consecuencia, el criterio de equidad por desconocimiento en este caso concreto, sería insuficiente.

Esta práctica se conoce como *redlining*, cuyo término fue acuñado en la década de 1960, debido a la práctica de negar bienes y servicios a las minorías mediante la *redlining* de barrios específicos en un mapa (Custers et al. [2012]).

6.3 EQUIDAD INDIVIDUAL

La *equidad individual* se basa en métricas de similitud sobre los atributos, y establece que individuos similares deben recibir predicciones similares independientemente del atributo sensible (Dwork et al. [2011]). Además, la equidad individual es más precisa que la equidad de grupo, ya que impone restricciones en el tratamiento específico para cada par de individuos.

Definición 53 (Equidad individual)**.** Sea $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario, $D\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ una medida de distancia sobre el espacio de clasificación resultante y $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ una medida de distancia sobre los individuos, se dice que g cumple con la *equidad individual* si, y solo si,

$$D(g(\mathbf{x}_i), g(\mathbf{x}_j)) \leq d(\mathbf{x}_i, \mathbf{x}_j), \text{ para todo } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}.$$

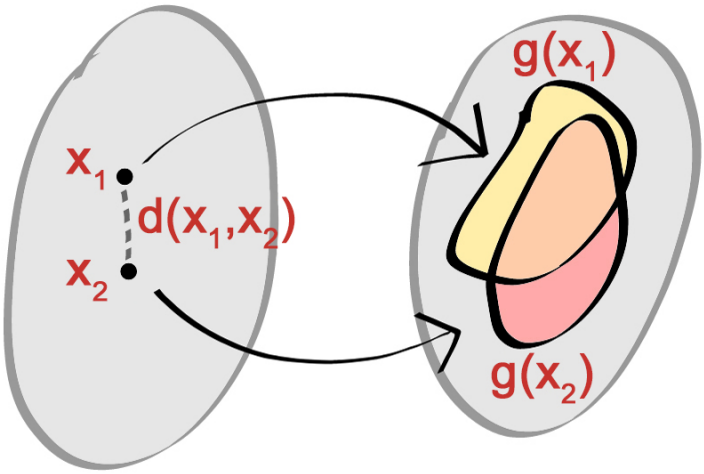


Figura 13: Ilustración de la noción de equidad individual.

La definición de equidad individual, también es conocida como *propiedad* (D, d)*-Lipschitz*. Cualquier clasificador que satisfaga esta propiedad también verificará la paridad demográfica con un cierto sesgo (véase Lema 3.1 en Dwork et al. [2011]).

En la literatura de las ciencias sociales, esta formalización equivale al individualismo igualitario, conocido por ser el principio formal de justicia. Esta noción responsabiliza a la *métrica de la distancia* de garantizar la justicia del clasificador. Si la métrica de la distancia utiliza los atributos sensibles directa o indirectamente para calcular la distancia entre dos individuos, un clasificador que satisfaga la Definición 53 podría seguir causando impacto dispar. Por tanto, la equidad individual, no podría considerarse adecuada para dominios en los que no se dispone de una métrica de distancia fiable y no discriminatoria.

Ejemplo 7. Imaginemos tres candidatos a un puesto de trabajo, A, B y C. A tiene únicamente el título de graduado y un año de experiencia laboral relacionada. B tiene un máster y un año de experiencia laboral relacionada. C tiene un doble grado pero no tiene experiencia laboral. En principio, no podemos disponer del rendimiento de los tres individuos ya que no podemos contratar a todos. Entonces: ¿Está A más cerca de B que de C? Si es así, ¿Por cuánto? La cosa se complica aún más cuando entran en juego los atributos sensibles. ¿Cómo deberíamos cuantificar la diferencia de pertenencia a un grupo con nuestra métrica?

En este ejemplo podemos observar los problemas de la equidad individual comentados anteriormente, y cómo dependen directamente de la construcción de la función métrica de distancia entre individuos.

6.4 EQUIDAD DE GRUPO

La *equidad de grupo* mide el impacto dispar entre grupos desfavorecidos y privilegiados, como podrían ser grupos de diferentes razas, edad o género.

Supongamos que en $\mathcal{X}$ se definen todas los posibles valores para las características {raza, género, salario, trabajo, edad}, denotaremos por $\mathcal{A}$ el conjunto de atributos sensibles sobre $\mathcal{X}$ considerando por ejemplo, en este caso,

$$\mathcal{A}=\{\text{raza, género, edad}\}.$$

Consideremos un *atributo multivaluado* $a = \{a_1, \ldots, a_n\} \in \mathcal{A}$, por ejemplo,

$$\text{raza}=\{\text{hispánico, caucásico, afroamericano, otro}\}.$$

Definimos un *grupo* $G(a_i)$ como un conjunto de entidades que tienen en común un valor específico del atributo a, por ejemplo raza=hispánico corresponde a todos los individuos de raza hispánica del conjunto de datos.

Teniendo en cuenta todos los grupos definidos por el atributo a, las predicciones $\hat{y}$ y la etiqueta real y para cada entidad de cada grupo, podemos hablar ahora de las métricas de grupo. Las principales definiciones sobre grupos para evaluar el sesgo y la equidad son las siguientes:

- **Atributo -** $a = \{a_1, \ldots, a_n\}$ es un atributo multivaluado, por ejemplo,

  $$\text{raza}=\{\text{hispánico, caucásico, afroamericano, otro}\}.$$

- **Grupo -** $G(a_i)$ es un grupo de todas las entidades que comparten el mismo valor de atributo $a = a_i$, por ejemplo, raza=hispánico.
- **Grupo de objetivo -** $G(a_o)$ es un grupo utilizado como objetivo del cálculo de las métricas de disparidad.
- **Grupo de referencia -** $G(a_r)$ es el grupo que se utiliza como referencia para calcular las métricas de disparidad. Suele fijarse siguiendo un criterio determinado.
- **Etiquetado positivo** - LP_G número de entidades etiquetadas como positivas dentro de un grupo.
- **Etiquetado negativo -** LN_G número de entidades etiquetadas como negativas dentro de un grupo.

Notación 3. De aquí en adelante, utilizaremos la siguiente notación:

- Y es una variable aleatoria binaria, que representa la etiqueta real de un individuo de $\mathcal{X}$.
- $\hat{Y}$ es una variable aleatoria binaria, que representa el resultado de la predicción de un clasificador $g\colon \mathcal{X} \to \mathcal{Y}$ para un individuo de $\mathcal{X}$.
- A es una variable aleatoria binaria, que representa si un individuo de $\mathcal{X}$ pertenece al grupo objetivo (a_o) o de referencia (a_r).

Métricas de grupos de distribución

Concretaremos *métricas de decisión* a nivel de grupo, centradas en la distribución de los individuos entre los grupos del conjunto seleccionado para la intervención, las cuales no precisan del valor de la etiqueta Y. Definimos las métricas de distribución de los grupos como sigue:

- **Positivos predichos -** PP_G número de entidades dentro de un grupo donde la decisión es positiva, es decir, $\hat{Y} = 1$.
- **Negativos predichos -** PN_G número de entidades dentro de un grupo cuya decisión es negativa, es decir, $\hat{Y} = -1$.
- **Total de predicciones positivas -** número total de entidades predichas como positivas en los grupos definidos por a,

$$\mathrm{K} = \sum_{i=1}^{n} \mathrm{PP}_{G(a_i)}.$$

- **Prevalencia predicha -** fracción de entidades dentro de un grupo que se predijo como positiva,

$$\mathrm{PPRev}_G = \frac{\mathrm{PP}_G}{|G|} = P(\hat{Y} = 1 \mid A = a_i).$$

- **Tasa de positivos predichos -** fracción de las entidades predichas como positivas que pertenecen a un determinado grupo,

$$\mathrm{PPR}_G = \frac{\mathrm{PP}_G}{\mathrm{K}} = P(A = a_i \mid \hat{Y} = 1).$$

Métricas de grupo basadas en la etiqueta real

A continuación, discutiremos diferentes métricas que surgen dependiendo de la coincidencia o no entre los valores de la variable de predicción $\hat{Y}$ y la etiqueta real Y. La mayoría de ellas, ya fueron presentadas en la Sección 5.4. Las métricas de grupo basadas en los errores y aciertos son las siguientes:

- **Falso positivo -** FP_G es el número de entidades del grupo con,

$$\hat{Y} = 1 \ \wedge \ Y = -1.$$

- **Falso negativo -** FN_G es el número de entidades del grupo con,

$$\hat{Y} = -1 \ \wedge \ Y = 1.$$

- **Verdadero positivo -** TP_G es el número de entidades del grupo con,

$$\hat{Y} = 1 \ \wedge \ Y = 1.$$

- **Verdadero negativo -** TN_G es el número de entidades del grupo con,

$$\hat{Y} = -1 \ \wedge \ Y = -1.$$

- **Prevalencia -** fracción de entidades dentro de un grupo cuyo resultado verdadero fue positivo,

$$\text{Prev}_G = \frac{\text{LP}_G}{|G|} = P(Y = 1 \mid A = a_i).$$

- **Tasa de falso descubrimiento -** fracción de falsos positivos de un grupo dentro de los positivos predichos del mismo,

$$\text{FDR}_G = \frac{\text{FP}_G}{\text{PP}_G} = P(Y = -1 \mid A = a_i, \hat{Y} = 1).$$

- **Tasa de falsa omisión -** fracción de falsos negativos de un grupo dentro de los negativos predichos del mismo,

$$\text{FOR}_G = \frac{\text{FN}_G}{\text{PN}_G} = P(Y = 1 \mid A = a_i, \hat{Y} = -1).$$

- **Tasa de falsos positivos -** fracción de falsos positivos de un grupo dentro de los negativos etiquetados del mismo,

$$\text{FPR}_G = \frac{\text{FP}_G}{\text{LN}_G} = P(\hat{Y} = 1 \mid A = a_i, Y = -1).$$

- **Tasa de falsos negativos -** fracción de falsos negativos de un grupo dentro de los positivos etiquetados del mismo,

$$\text{FNR}_G = \frac{\text{FN}_G}{\text{LP}_G} = P(\hat{Y} = -1 \mid A = a_i, Y = 1).$$

- **Valor negativo predictivo -** fracción de verdaderos negativos de un grupo dentro de los negativos predichos del mismo,

$$\text{NPV}_G = \frac{\text{TN}_G}{\text{PN}_G} = P(Y = -1 \mid A = a_i, \hat{Y} = -1).$$

- **Valor positivo predictivo (Precision) -** fracción de verdaderos positivos de un grupo dentro de los positivos predichos del mismo,

$$\text{PPV}_G = \frac{\text{TP}_G}{\text{PP}_G} = P(Y = 1 \mid A = a_i, \hat{Y} = 1).$$

- **Tasa de verdaderos positivos (Recall) -** fracción de verdaderos positivos de un grupo dentro de los positivos etiquetados del mismo,

$$\text{TPR}_G = \frac{\text{TP}_G}{\text{LP}_G} = P(\hat{Y} = 1 \mid A = a_i, Y = 1).$$

- **Tasa de verdaderos negativos (Specificity) -** fracción de verdaderos negativos de un grupo dentro de los negativos etiquetados del mismo,

$$\text{TNR}_G = \frac{\text{TN}_G}{\text{LN}_G} = P(\hat{Y} = -1 \mid A = a_i, Y = -1).$$

- **Exactitud (Accuracy) -** fracción de resultados verdaderos de un grupo dentro del total de casos examinados del mismo,

$$\text{Accuracy}_G = \frac{\text{TP}_G + \text{TN}_G}{\text{LP}_G + \text{LN}_G} = P(\hat{Y} = Y \mid A = a_i).$$

- **Tasa global de clasificación errónea. -** fracción de resultados falsos de un grupo dentro del total de casos examinados del mismo,

$$\text{OMR}_G = \frac{\text{FP}_G + \text{FN}_G}{\text{LP}_G + \text{LN}_G} = P(\hat{Y} \neq Y \mid A = a_i).$$

En los apartados siguientes, formalizaremos algunas de las nociones populares de equidad de grupo que podemos encontrar en la literatura. Sea un atributo sensible multivaluado $a \in \mathcal{A}$, las métricas relativas a equidad de grupo se definen como una igualdad entre las probabilidades de un *grupo objetivo* (a_o) en comparación con un *grupo de referencia* (a_r).

El grupo de referencia se suele seleccionar en base a diferentes criterios. Por ejemplo, se podría utilizar el grupo mayoritario entre los grupos definidos por A, o el enfoque tradicional de fijar un grupo históricamente favorecido, por ejemplo, en el caso de la raza, los individuos de raza caucásica.

6.4.1 *Paridad demográfica*

La *paridad demográfica*, también conocida como *paridad estadística* o *independencia*, es uno de los criterios de equidad de grupo más conocidos. Esta noción de equidad afirma que la probabilidad de ser clasificado con el resultado positivo (o negativo) debe ser independiente de que el individuo pertenezca al grupo protegido, es decir, que los datos demográficos de los individuos clasificados positivamente son idénticos a los de la población en su conjunto (Dwork et al. [2011]).

Definición 54 (Independencia en clasificación binaria)**.** Sean C, A variables aleatorias. La *independencia* entre C y A, equivale a que se cumpla la siguiente restricción:

$$P(C = c \mid A = a_r) = P(C = c \mid A = a_o).$$

Lo denotaremos como $C \perp A$.

Definición 55 (Paridad demográfica)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario. Se dice que g cumple con la *paridad demográfica* si, y solo si, $\hat{Y} \perp A$.

Relajaciones y aproximaciones

Podemos *relajar* el concepto de paridad demográfica suponiendo que $\hat{Y} = 1$ y aproximando su definición con una acotación en valor absoluto de las probabilidades a partir una constante fijada $\tau \in [0, 1]$. De esta manera, aproximamos este criterio de equidad como:

$$\left|P(\hat{Y} = 1 \mid A = a_r) - P(\hat{Y} = 1 \mid A = a_o)\right| \leq \tau.$$

Tomando un $\epsilon \in [0, 1)$ también podemos aproximar el concepto de paridad demográfica de la siguiente manera:

$$1 - \epsilon \leq \frac{P(\hat{Y} = 1 \mid A = a_o)}{P(\hat{Y} = 1 \mid A = a_r)}.$$

En algunos trabajos se suele obviar la acotación por una constante y simplemente se define la paridad demográfica utilizando la Definición 55 y asumiendo que $\hat{Y} = 1$, lo que sería equivalente a igualar las métricas PPRev entre los subgrupos.

$$\text{PPRev}_{a_r} = \text{PPRev}_{a_o} \Rightarrow P(\hat{Y} = 1 \mid A = a_r) = P(\hat{Y} = 1 \mid A = a_o).$$

Ejemplo 8. Consideremos un sistema de detección de delitos y dos grupos de igual tamaño, A y B. Suponemos que los miembros del grupo B tienen el doble probabilidades reales de cometer un delito que los individuos del grupo A. Al igualar la probabilidad de un resultado positivo, el mismo número de predicciones positivas se distribuiría entre un grupo mucho mayor de delincuentes para B que para A. Así, un delincuente del grupo B tendría menos probabilidades de serlo que un delincuente del grupo A ($\text{FNR}_A < \text{FNR}_B$). De hecho, para la misma precisión, la tasa de verdaderos positivos del grupo B sería la mitad de la del grupo A, $\frac{1}{2}\text{TPR}_A = \text{TPR}_B$, cumpliendo la paridad demográfica.

6.4.2 *Probabilidades igualadas*

Uno de los problemas de la paridad demográfica es que ignora una posible correlación entre Y y A. El criterio de las *probabilidades igualadas*, también conocido como *ratio de paridad positiva* o *separación*, tiene en cuenta la etiqueta real de cada grupo y su condicionamiento al resto de variables. Además, proporciona un incentivo para reducir los errores de manera uniforme en todos los grupos sin descartar el clasificador perfecto (que obtenga $\hat{Y} = Y$) a diferencia de la paridad estadística.

Definición 56 (Independencia condicional en clasificación binaria)**.** Sean C, Y, A variables aleatorias. La *independencia condicional* entre C y A dado Y, equivale a que se cumpla la siguiente restricción:

$$P(C = c \mid A = a_r, Y = y) = P(C = c \mid A = a_o, Y = y).$$

Lo denotaremos como $C \perp A \mid Y$.

El criterio de las probabilidades igualadas establece que $\hat{Y}$ debe ser condicionalmente independiente de A dado Y, permitiendo que el clasificador dependa de A a través de la variable objetivo (Hardt et al. [2016]). Para utilizar este criterio, es necesario conocer las etiquetas reales de cada individuo, por lo que esta medida restringe su uso para determinadas tareas en las que no conozcamos previamente el resultado de la acción sobre el individuo.

Definición 57 (Probabilidades igualadas)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario. Se dice que g cumple con el criterio de las *probabilidades igualadas* si, y solo si, $\hat{Y} \perp A \mid Y$.

Relajaciones y aproximaciones

El concepto previo depende de varias variables, por lo que normalmente en la práctica, se tiende a relajar el criterio fijando algunos valores en la definición. A partir de estas relajaciones surgen otros criterios de equidad que también son ampliamente utilizados y conocidos en la literatura.

La relajación más común surge al suponer que $\hat{Y} = 1$, en este caso, el criterio de las probabilidades igualadas equivale a igualar las métricas FPR y TPR entre los subgrupos. Esta aproximación beneficia al individuo, y equilibra la probabilidad de tener un resultado beneficioso, en todos los subgrupos de los individuos etiquetados tanto positiva como negativamente.

$$\text{FPR}_{a_r} = \text{FPR}_{a_o} \Rightarrow P(\hat{Y} = 1 \mid A = a_r, Y = -1) = P(\hat{Y} = 1 \mid A = a_o, Y = -1).$$
$$\text{TPR}_{a_r} = \text{TPR}_{a_o} \Rightarrow P(\hat{Y} = 1 \mid A = a_r, Y = 1) = P(\hat{Y} = 1 \mid A = a_o, Y = 1).$$

Definiremos el concepto de igualdad de oportunidades como la igualdad de las tasas de verdaderos positivos entre los subgrupos (Hardt et al. [2016]). En algunos trabajos, también se define este criterio de forma equivalente para las tasas de verdaderos negativos.

$$\text{TPR}_{a_r} = \text{TPR}_{a_o} \Rightarrow P(\hat{Y} = 1 \mid A = a_r, Y = 1) = P(\hat{Y} = 1 \mid A = a_o, Y = 1).$$

Definición 58 (Igualdad de oportunidades)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario. Se dice que g cumple con la *igualdad de oportunidades* si, y solo si,

$$P(\hat{Y} = 1 \mid A = a_r, Y = 1) = P(\hat{Y} = 1 \mid A = a_o, Y = 1).$$

La igualdad de oportunidades es naturalmente más débil que la Definición 57, ya que asumimos los valores de $\hat{Y} = 1$ e $Y = 1$. Este concepto, iguala la probabilidad de que los individuos etiquetados positivamente sean correctamente clasificados con el resultado positivo (beneficioso). Por ejemplo, dos individuos, un hombre y una mujer, que están cualificados para un trabajo ($Y = 1$), deberían tener la misma probabilidad de conseguir el trabajo ($\hat{Y} = 1$).

Ejemplo 9. Consideremos un sistema de contratación y dos grupos de igual tamaño, A y B. Imaginemos que en el grupo A de los 100 aspirantes al cargo, 58 están cualificados, mientras que el grupo B solo 2 de ellos son aptos para el cargo. Si la empresa decide aceptar a 30 solicitantes y satisfacer la igualdad de oportunidades, se concederán 29 ofertas al grupo A mientras que solo se concederá 1 oferta al grupo B. Si el trabajo es bien remunerado, el grupo A mejorará sus condiciones de vida y a la larga, podrá permitir una mejor educación para sus hijos, y en consecuencia una mejor cualificación de los mismos en el futuro.

En este ejemplo podemos observar que la igualdad de oportunidades no ayuda a cerrar la brecha entre los dos grupos, es más la brecha entre el grupo A y el grupo B, tenderá a ampliarse con el tiempo.

6.4.3 *Tasa de paridad predictiva*

La *tasa de paridad predictiva*, también denominada *suficiencia* surge de una motivación equivalente a la del criterio de las probabilidades igualadas. El concepto se define de igual manera haciendo uso de la independencia condicional, pero intercambiando los papeles de $\hat{Y}$ e Y.

Definición 59 (Tasa de paridad predictiva)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario. Se dice que g cumple con la *tasa de paridad predictiva* si, y solo si, $Y \perp A \mid \hat{Y}$.

Relajaciones y aproximaciones

La relajación más común surge al suponer que $\hat{Y} = Y$, en este caso, el criterio de las probabilidades igualadas equivale a igualar las métricas PPV y NPV entre los subgrupos.

$$\begin{aligned}\text{PPV}_{a_r} = \text{PPV}_{a_o} &\Rightarrow P(Y = 1 \mid A = a_r, \hat{Y} = 1) = P(Y = 1 \mid A = a_o, \hat{Y} = 1).\\ \text{NPV}_{a_r} = \text{NPV}_{a_o} &\Rightarrow P(Y = -1 \mid A = a_r, \hat{Y} = -1) = P(Y = -1 \mid A = a_o, \hat{Y} = -1).\end{aligned}$$

Las limitaciones de este concepto de equidad son semejantes a las de las probabilidades igualadas, pudiendo acrecentar las diferencias entre los grupos privilegiado y desfavorecido.

6.4.4 *Medidas basadas en la puntuación*

A diferencia de las definiciones anteriores, que se basan en los índices de clasificación binaria, algunas nociones de equidad se basan en la *puntuación* de la probabilidad predicha S y la etiqueta real Y (Mitchell et al. [2021]).

El *balance para la clase positiva* (o *negativa*) se cumple cuando la puntuación esperada para un individuo clasificado positivamente (o negativamente) es igual en todos los grupos.

Definición 60 (Balance para la clase positiva)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $S \in [0, 1]$ la puntuación asignada por el clasificador arbitrario $g\colon \mathcal{X} \to \mathcal{Y}$. Se dice que g cumple con el *balance para la clase positiva* si, y solo si,

$$\mathbb{E}[S = s \mid A = a_r, Y = 1] = \mathbb{E}[S = s \mid A = a_o, Y = 1].$$

Definición 61 (Balance para la clase negativa)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $S \in [0, 1]$ la puntuación asignada por el clasificador arbitrario $g\colon \mathcal{X} \to \mathcal{Y}$. Se dice que g cumple con el *balance para la clase negativa* si, y solo si,

$$\mathbb{E}[S = s \mid A = a_r, Y = -1] = \mathbb{E}[S = s \mid A = a_o, Y = -1].$$

Sin embargo, hay que tener en cuenta que en los problemas del mundo real es casi imposible cumplir el equilibrio para la clase negativa y para la clase positiva simultáneamente.

6.4.5 *Igualdad de las métricas de predicción*

Como hemos podido observar, en general, la mayoría de los criterios de equidad definidos surgen como una igualación de las métricas de grupo presentadas al inicio de la Sección 6.4. De esta forma podemos definir, un nuevo criterio para cada métrica de la siguiente forma.

Definición 62 (Paridad métrica)**.** Sea $A \in \mathcal{A}$ un atributo sensible multivaluado y $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario. Se dice que g cumple con la *paridad métrica* si, y solo si,

$$\text{Métrica}_{a_r} = \text{Métrica}_{a_o}.$$

Donde Métrica podrá ser cualquier métrica de grupo definida previamente.

Aunque esta definición nos permite crear una gran variedad de criterios de equidad, la formulación a partir de una igualdad sigue siendo difícil a la hora de aplicarla en la práctica.

6.4.6 *Impacto desigual*

Para facilitar la implementación práctica de las métricas anteriores, aparece el término *impacto desigual* (no confundir con el término impacto dispar definido en la Tabla 1). Esta noción está directamente relacionada con la regla $p\,\%$, que podemos encontrar en la literatura jurídica (Binns [2021]), según la cual una decisión es discriminatoria si el coeficiente del impacto desigual es inferior a un valor τ dependiente de p ($\tau = \frac{p}{100}$).

Definición 63. (Impacto desigual) Sea $A \in \mathcal{A}$ un atributo sensible multivaluado, $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario y $\tau \in [0,1]$. Se dice que g satisface el *impacto desigual* si, y solo si,

$$DI = \frac{P(\hat{Y} = 1 \mid A = a_o)}{P(\hat{Y} = 1 \mid A = a_r)} < \tau. \tag{5}$$

En el mundo real, si un clasificador aplicado a una tarea en una empresa satisface el impacto desigual, debe justificarse que su aplicación es esencial para el funcionamiento seguro y eficiente del negocio, y no existen procedimientos alternativos que sean sustancialmente igual de válidos y tengan un impacto menos adverso. La Comisión para la Igualdad de Oportunidades en el Empleo (EEOC) de Estados Unidos adopta la regla del 80 % ($\tau = 0{,}8$) para considerar si una decisión tiene impacto desigual (Adverse Impact Analysis / Four-Fifths Rule).

Construcción de otras medidas en la práctica

Si queremos evitar el impacto desigual, impondremos la desigualdad opuesta de la Ecuación (5). Estableciendo además una cota superior dada por $\frac{1}{\tau}$, obtenemos el criterio de equidad implementado por Aequitas (Saleiro et al. [2019]).

$$\tau \leq \frac{P(\hat{Y} = 1 \mid A = a_o)}{P(\hat{Y} = 1 \mid A = a_r)} \leq \frac{1}{\tau}.$$

La formulación aportada por Aequitas se basa en calcular la fracción de la métrica de grupo elegida y una vez calculada, comprobar si se encuentra dentro del rango definido. La pertenencia o no al rango será una aproximación a la igualación de las métricas de la Definición 62.

$$\tau \leq \text{Métrica de disparidad}_G \leq \frac{1}{\tau}.$$

Usamos $\tau \in (0, 1]$ para controlar el rango de valores de disparidad que pueden considerarse justos. Para aplicar la regla del 80 %, simplemente bastaría con tomar $\tau = 0{,}8$. Diremos que un clasificador será tan justo como lo permita el valor máximo del sesgo entre los grupos definidos por los atributos protegidos.

Esta formalización también puede extenderse a cualquier métrica de grupo de las explicadas en la Sección 6.4. Definiremos, por ejemplo, la tasa de falsas omisiones (FOR) como:

$$\text{Métrica FOR}_{G(a_o)} = \frac{\text{FOR}a_o}{\text{FOR}a_r} = \frac{P(Y = 1 \mid A = a_o, \hat{Y} = -1)}{P(Y = 1 \mid A = a_r, \hat{Y} = -1)}.$$

Aequitas utiliza la tasa de positivos predichos (PPR) para aproximar el concepto de paridad demográfica. Usaremos unas métricas u otras en función del impacto y el objetivo que quiera intervenir el usuario. Si las intervenciones pueden perjudicar a los individuos (punitivas), entonces querremos minimizar los falsos positivos, centrándonos en la tasa de falsos descubrimientos (FDR) o la de falsos positivos (FPR). Si por otro lado tienen como objetivo beneficiar a los individuos (asistenciales), deberíamos preocuparnos más por los falsos negativos, priorizando la tasa de falsa omisión (FOR) o la de falsos negativos (FNR).

7

ALGORITMOS DE MITIGACIÓN DE SESGO

En este capítulo discutiremos los diferentes algoritmos existentes para la mitigación del sesgo, daremos algunos ejemplos específicos de los aportados en la bibliografía y comentaremos sus características más relevantes.

7.1 MODELOS DE APRENDIZAJE JUSTOS

En la literatura, podemos encontrar una gran variedad de métodos y algoritmos que nos pueden ayudar a mejorar la equidad en un modelo de aprendizaje. Los enfoques de *mitigación del sesgo* pueden subdividirse en tres categorías: algoritmos de preprocesamiento, que intentan aprender representaciones justas de los datos; algoritmos de optimización durante el entrenamiento, que ajustan el proceso de aprendizaje para cumplir los criterios de justicia; y algoritmos de posprocesamiento, que adaptan las predicciones del modelo en función de sus resultados. Estas categorías no tienen por qué ser mutuamente excluyentes y a veces pueden tener varios métodos de actuación sobre los datos.

A lo largo de este capítulo, presentaremos las características comunes entre los algoritmos de cada uno de los tres grupos presentados y discutiremos algunos ejemplos que podemos encontrar en la literatura, profundizando en algún caso específico de cada tipo. Además, contextualizaremos en su categoría correspondiente el algoritmo de optimización de equidad contrafactual presentado por Kusner et al. [2018] sobre el que basaremos la parte experimental de nuestro trabajo.

7.1.1 *Selección de los datos del modelo*

En la Sección 5.1, observamos que para construir un modelo de aprendizaje es necesario tener un conjunto de datos $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ a partir del cual poder entrenar el modelo y extraer la información que se quiere aprender. Normalmente para entrenar los modelos, no utilizaremos el conjunto de datos al completo, sino que tomaremos un conjunto de entrenamiento X, que contiene m individuos extraídos aleatoriamente del conjunto de datos total, donde $m < n$. Cada elemento $\mathbf{x} \in X$ es un vector de longitud d donde cada componente del vector describe una característica

del individuo. Además, cada vector $\mathbf{x}$ tiene asociado un atributo sensible $a \in \{0, 1\}$, donde el valor 0 denota la pertenencia al grupo de referencia (a_r) y 1 al grupo objetivo (a_o). Denotaremos por Y al conjunto con las etiquetas reales de los individuos que están en X. De esta forma, tenemos que $X \times Y \subset \mathcal{D}$.

En la práctica, se suele utilizar el conjunto de los individuos restantes contenidos en $\mathcal{D}$ como conjunto de prueba del modelo, con el objetivo de poder comprobar su rendimiento sobre un conjunto de datos con el que no ha sido entrenado.

7.1.2 *Equilibrio entre equidad y métricas de evaluación*

Algunas medidas de evaluación como la precisión o exactitud dependen directamente del conjunto de datos, la definición de equidad utilizada y los algoritmos empleados. La equidad en la práctica, perjudica a métricas como la exactitud. Si queremos mitigar el sesgo entre grupos, debemos hacer una compensación entre la equidad y la exactitud, sacrificando esta última como se puede observar en la Figura 14.

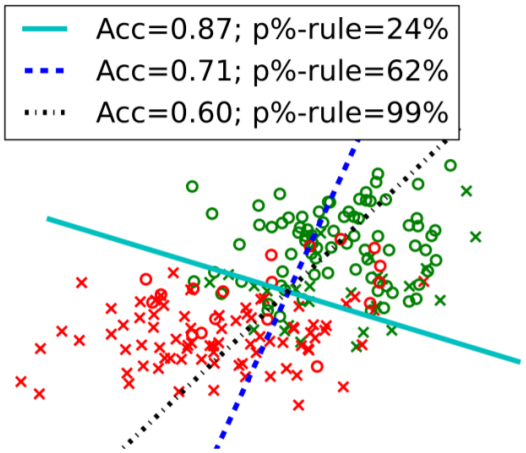


Figura 14: Exactitud vs. Independencia en un problema de clasificación (Zafar et al. [2017b]).

7.2 ALGORITMOS DE PREPROCESAMIENTO

Los *algoritmos de preprocesamiento* buscan mejorar la equidad antes de entrenar el modelo, modificando los datos de entrenamiento de forma que no presenten sesgos antes de ser procesados. El propósito de estos algoritmos es aprender una nueva representación Z que elimine la información correlacionada con el atributo sensible A y preserve, en la medida de lo posible, la información del conjunto de individuos X sin necesidad de conocer el valor de sus etiquetas Y. La tarea posterior (por ejemplo, regresión o clasificación) desempeñada por g, será independiente del método usado y podrá producir resultados que preserven diversos criterios de equidad.

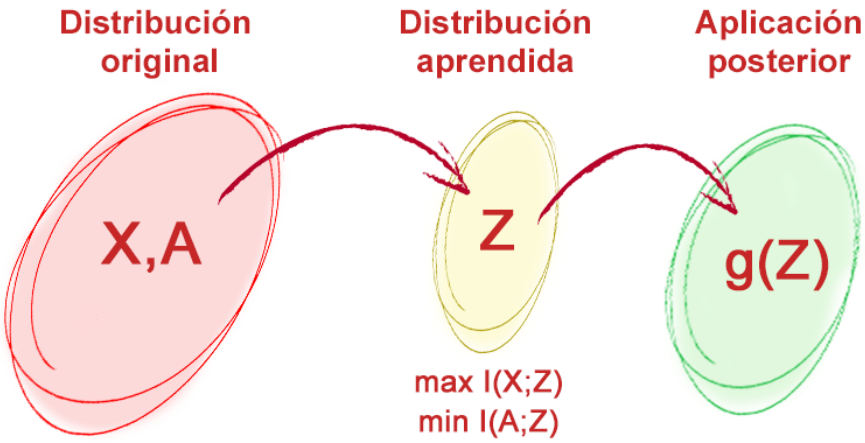


Figura 15: Etapas del proceso de preprocesamiento.

7.2.1 *Ejemplo: Aprendizaje de la representación justa*

Zemel et al. [2013] presenta este problema como el aprendizaje de un función que lleva muestras individuales a representaciones intermedias. Este trabajo tiene dos objetivos: minimizar la pérdida de información (los datos originales deberían preservar la misma cantidad de información) y maximizar la equidad (la pertenencia a un grupo protegido no debería afectar negativamente a los individuos del mismo).

Este ejemplo, podría confundirse como una aplicación de equidad por desconocimiento, pero no es así, ya que la pertenencia a grupos protegidos no es simplemente ignorada, sino que se trata activamente junto con la información redundante de estos atributos. El modelo propuesto, tiene como objetivo mantener la equidad de grupo e individual, al mismo tiempo que maximiza la exactitud.

Notación 4. Introducimos la siguiente notación que utilizaremos en este ejemplo:

- $X^+ = \{\mathbf{x}_n \in X \ : \ A = 1\} \subset X$ es el conjunto de datos de entrenamiento cuyos miembros pertenecen al grupo protegido, $X^- = \{\mathbf{x}_n \in X \ : \ A = 0\}$ denota al conjunto cuyos miembros pertenecen al grupo no protegido.
- Z es una variable aleatoria multivariante, donde cada uno de sus r valores representa un "prototipo". Asociado a cada prototipo existe un vector $\mathbf{v}_k$ en el mismo espacio que los individuos $\mathbf{x}_n$.

La idea es representar cada individuo $\mathbf{x}_n \in X$ como una combinación lineal ponderada de r prototipos para satisfacer la paridad demográfica, minimizando la pérdida de información original y maximizando la exactitud en la medida de lo posible.

Se define la probabilidad *softmax* de que un elemento sea un prototipo concreto como:

$$M_{n,k} := P(Z = k \mid \mathbf{x}_n) = \frac{\exp(-d(\mathbf{x}_n, \mathbf{v}_k))}{\sum_{j=1}^{r} \exp(-d(\mathbf{x}_n, \mathbf{v}_j))}, \text{ para todo } n, k.$$

donde d es una función de medida de distancia (por ejemplo, la distancia ℓ_2). Los autores definen la regularización ℓ_2 como una función de distancia ponderada que viene dada por,

$$d(\mathbf{x}_n, \mathbf{v}_k, \alpha) = \sum_{i=1}^{d} \alpha_i (x_{ni} - v_{ki})^2.$$

Se define $\hat{y}_n$ como la predicción para y_n calculada partiendo de la marginalización sobre el valor de Y para la predicción de cada prototipo.

$$\hat{y}_n = \sum_{k=1}^{r} M_{n,k} w_k.$$

El modelo de aprendizaje, minimiza la siguiente función de pérdida:

$$L = A_z L_z + A_x L_x + A_y L_y.$$

Donde L_z regulariza la paridad demográfica, L_x es el error de reconstrucción de la distribución y L_y cuantifica la pérdida de predicción. Los factores A_z, A_x, A_y se definen como hiperparámetros para equilibrar estas pérdidas.

$$L_z = \sum_{k=1}^{r} \left| M_k^+ - M_k^- \right|.$$

expresando $M_k^+ = \dfrac{1}{|X^+|} \displaystyle\sum_{n \in X^+} M_{n,k}$ y M_k^- se formula de forma similar a partir de X^-.

$$L_x = \sum_{n=1}^{m} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2.$$

donde $\hat{\mathbf{x}}_n$ son los nuevos valores de $\mathbf{x}_n$ para Z:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{r} M_{n,k} \mathbf{v}_k.$$

$$L_y = \sum_{n=1}^{m} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n).$$

En la fase de entrenamiento, los valores de $\mathbf{v}, \mathbf{w}, \alpha$ se optimizan conjuntamente a través del método L-BFGS (Zhu et al. [1997]), para minimizar la función objetivo L. Los valores de A_x, A_y, A_z se seleccionan a través de la afinación de los hiperparámetros, optando por los que producen mejores resultados (búsqueda en malla). Tenemos que tener en cuenta que la función objetivo no es convexa y, por tanto, no garantiza la optimización.

Ventajas e inconvenientes

Algunas de las principales ventajas de los algoritmos de preprocesamiento, ya han sido mencionadas previamente. En general, estos métodos, son muy útiles cuando el clasificador que utilizaremos para la fase de entrenamiento es un modelo de caja negra y no conocemos su actuación sobre los datos. En estos casos, al devolver una nueva distribución que no contiene correlación con los atributos sensibles, los datos podrán ser usados de forma independiente para cualquier tarea posterior sin preocuparnos por la existencia de sesgo entre grupos.

Por otro lado, solo podemos utilizar los métodos de preprocesamiento para optimizar los criterios de equidad que no requieran información sobre el valor de las etiquetas Y (por ejemplo, la paridad demográfica o la equidad individual). Como desconocemos el uso que se le va a dar a los nuevos datos, en algunos casos, podría no garantizarse la equidad en el modelo final aprendido. Además, este grupo de algoritmos suele ser el peor en términos de rendimiento entre exactitud y equidad.

Otros ejemplos en la literatura

El concepto de preprocesamiento optimizado es introducido por Calmon et al. [2017]. En este se plantea la reducción de la discriminación como una tarea de optimización convexa con el objetivo de minimizar la pérdida de exactitud (preservando la utilidad), mientras se limita por ciertas medidas de equidad de grupo e individual.

Wang et al. [2019] ofrece métodos que usan distribuciones contrafactuales para resolver el trato dispar de un clasificador de caja negra en una población de despliegue sin la necesidad de reentrenar el modelo. El método propuesto se basa en la construcción de una nueva distribución para los individuos del grupo objetivo, de manera que mejoren sus resultados en promedio.

7.3 ALGORITMOS DE OPTIMIZACIÓN DURANTE EL ENTRENAMIENTO

Los *algoritmos de optimización* alteran el entrenamiento del propio modelo. En este contexto, la mitigación del sesgo se plantea como un aprendizaje al que se le añade una restricción o un término de regularización al objetivo de optimización existente. Es decir, un modelo aprende a optimizar una función de pérdida en los datos de entrenamiento, sujeta a restricciones de equidad (por ejemplo, la distancia máxima a la paridad demográfica).

Otro enfoque diferente, sería el de la optimización de métricas de rendimiento complejas que incluyan alguna noción de equidad. Por ejemplo, introduciendo una penalización relacionada con la equidad en la función objetivo.

7.3.1 *Ejemplo: Aprendizaje en clasificación sin impacto dispar.*

Uno de los enfoques más populares es el de la optimización con restricciones; donde el objetivo es encontrar un conjunto de parámetros $\boldsymbol{\theta} \in \Theta$, que minimicen una función objetivo $l_0(\boldsymbol{\theta})$, sujeta a m restricciones funcionales $l_i(\boldsymbol{\theta})$, para todo $i \in \{1, \ldots, m\}$.

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} l_0(\boldsymbol{\theta}) \text{ donde } l_i(\boldsymbol{\theta}) \leq 0 \text{ para todo } i \in \{1, \ldots, m\}.$$

Zafar et al. [2017b] enmarca la tarea de clasificación justa, imponiendo restricciones lineales a la covarianza entre los atributos sensibles y las predicciones. Este método es adecuado para múltiples atributos sensibles y para cualquier clasificador basado en contornos convexos (por ejemplo, SVM o regresión logística). Además, se propone otra formulación similar destinada a satisfacer necesidades del mundo real, maximizando la equidad sujeta a restricciones de exactitud.

En el trabajo anterior se asume que el conjunto de datos original presenta un sesgo histórico, por lo que en Zafar et al. [2017a], se amplía este enfoque a los casos en los que tenemos acceso a los resultados reales no sesgados durante la fase de entrenamiento, y podemos saber si una decisión histórica fue correcta o incorrecta.

Notación 5. Introducimos la siguiente notación para el ejemplo propuesto:

- $\mathcal{D}'$ es el conjunto de datos de entrenamiento que se define como

$$\mathcal{D}' = \{(\mathbf{x}_i, y_i) \in X \times Y \ : \ i = 1, \ldots, m\}.$$

- $\boldsymbol{\theta}$ son los parámetros que debemos aprender.
- $L(\boldsymbol{\theta})$ es la función de pérdida convexa original.
- $d_{\boldsymbol{\theta}}(\mathbf{x})$ es la función de distancia con signo del vector de características $\mathbf{x}$ al límite de decisión del clasificador.
- $f_{\boldsymbol{\theta}}(\mathbf{x})$ es la función de clasificación, definida por

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \begin{cases} 1, & \text{si } d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0, \\ -1, & \text{en otro caso.} \end{cases}$$

Podemos añadir como restricciones al problema de optimización original, la paridad OMR (definida como una relajación para $\hat{Y} \neq Y$ del criterio de probabilidades igualadas) o la paridad FNR. Aunque, para el ejemplo, utilizaremos la paridad FPR construida a partir de la Definición 62 y que se define como:

$$P(\hat{Y} = 1 \mid A = 0, Y = -1) = P(\hat{Y} = 1 \mid A = 1, Y = -1).$$

Cabe señalar que la paridad FNR y FPR implican la paridad TPR y TNR respectivamente, y por tanto, la igualdad de oportunidades. En este ejemplo, usaremos como restricción la paridad FPR a partir de la cual surge la siguiente formulación de optimización:

$$\begin{aligned} \text{minimizar:} \quad & L(\boldsymbol{\theta}) \\ \text{sujeto a:} \quad & P(\hat{Y} = 1 \mid A = 0, Y = -1) - P(\hat{Y} = 1 \mid A = 1, Y = -1) \leq \epsilon, \\ & P(\hat{Y} = 1 \mid A = 0, Y = -1) - P(\hat{Y} = 1 \mid A = 1, Y = -1) \geq -\epsilon. \end{aligned} \tag{6}$$

La complejidad de las restricciones, hacen que el problema que plantea minimizar la función $L(\boldsymbol{\theta})$ (a priori no convexa) sea intratable a nivel computacional, al no poder utilizar los algoritmos tradicionales como el de descenso de gradiente estocástico u otros resolutores para encontrar la solución óptima al problema que se plantea.

Para subsanar esta cuestión, se presentan algunas relajaciones de las restricciones que utilizan la covarianza entre los atributos sensibles de los individuos y $d_{\boldsymbol{\theta}}(\mathbf{x})$ para detectar la relación entre el atributo protegido y las predicciones a nivel de grupo:

$$\begin{aligned} \text{Cov}(a, g_{\boldsymbol{\theta}}(y, \mathbf{x})) &= \mathbb{E}[(a - \bar{a})(g_{\boldsymbol{\theta}}(y, \mathbf{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \mathbf{x}))] \\ &\approx \frac{1}{m} \sum_{(\mathbf{x}, y, a) \in \mathcal{D}'} (a - \bar{a}) g_{\boldsymbol{\theta}}(y, \mathbf{x}), \end{aligned}$$

donde el término $\mathbb{E}[(z - \bar{z})]\bar{g}_{\boldsymbol{\theta}}(\mathbf{x})$ se anula, ya que $\mathbb{E}[(z - \bar{z})] = 0$ y la función $g_{\boldsymbol{\theta}}(y, \mathbf{x})$ se puede definir como:

$$g_{\boldsymbol{\theta}}(y, \mathbf{x}) = \text{mín}(0, \frac{1 - y}{2} y d_{\boldsymbol{\theta}}(\mathbf{x})).$$

Tras la relajación, podemos reescribir (6) como:

$$\begin{aligned} \text{minimizar:} \quad & L(\boldsymbol{\theta}) \\ \text{sujeto a:} \quad & \frac{1}{m} \sum_{(\mathbf{x}, y, a) \in \mathcal{D}'} (a - \bar{a}) g_{\boldsymbol{\theta}}(y, \mathbf{x}) \leq c, \\ & \frac{1}{m} \sum_{(\mathbf{x}, y, a) \in \mathcal{D}'} (a - \bar{a}) g_{\boldsymbol{\theta}}(y, \mathbf{x}) \geq -c, \end{aligned} \tag{7}$$

donde el umbral de covarianza $c \in \mathbb{R}^+$, controla el grado de desempeño del criterio de igualdad de oportunidades.

Esta formulación sigue siendo no convexa, por lo que a continuación convertiremos estas restricciones en un *programa convexo-cóncavo disciplinado* (DCCP), que puede resolverse de manera eficiente aprovechando los recientes avances en la programación convexa-cóncava (Shen et al. [2016]).

En primer lugar, consideramos la restricción descrita en (7), es decir:

$$\sum_{(\mathbf{x},y,a)\in\mathcal{D}'} (a-\bar{a})g_{\boldsymbol{\theta}}(y,\mathbf{x}) \sim c,$$

donde $\sim$, podría denotar '$\leq$' o '$\geq$'. Además, dejamos de lado la constante $\frac{1}{n}$ para simplificar. Como $a \in \{0,1\}$, dividimos la suma en la expresión anterior en dos términos:

$$\sum_{(\mathbf{x},y)\in\mathcal{D}'_0} (0-\bar{a})g_{\boldsymbol{\theta}}(y,\mathbf{x}) + \sum_{(\mathbf{x},y)\in\mathcal{D}'_1} (1-\bar{a})g_{\boldsymbol{\theta}}(y,\mathbf{x}) \sim c, \tag{8}$$

donde $\mathcal{D}'_0$ y $\mathcal{D}'_1$ son subconjuntos del conjunto de datos $\mathcal{D}'$ que toman valores $a=0$ y $a=1$, respectivamente. Definimos $m_0 = |\mathcal{D}'_0|$ y $m_1 = |\mathcal{D}'_1|$, entonces $\bar{z} = \frac{m_1}{m}$ y podemos reescribir (8) como:

$$-\frac{m_1}{m}\sum_{(\mathbf{x},y)\in\mathcal{D}'_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{m_0}{m}\sum_{(\mathbf{x},y)\in\mathcal{D}'_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \sim c,$$

que, dado que $g_{\boldsymbol{\theta}}$ es convexa en en $\boldsymbol{\theta}$ (por suposición), resulta en una función convexa-cóncava.

Por lo tanto, podemos reescribir el problema definido por (7) como:

$$\begin{aligned} \text{minimizar:} \quad & L(\boldsymbol{\theta}) \\ \text{sujeto a:} \quad & -\frac{m_1}{m}\sum_{(\mathbf{x},y)\in\mathcal{D}'_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) + \frac{m_0}{m}\sum_{(\mathbf{x},y)\in\mathcal{D}'_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c, \\ & -\frac{m_1}{m}\sum_{(\mathbf{x},y)\in\mathcal{D}'_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) + \frac{m_0}{m}\sum_{(\mathbf{x},y)\in\mathcal{D}'_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c, \end{aligned}$$

que es un DCCP para cualquier función de pérdida convexa $L(\boldsymbol{\theta})$, y puede ser resuelto eficientemente usando heurísticas como la propuestas en Shen et al. [2016].

Ventajas e inconvenientes

Entre las principales ventajas que tiene este grupo de algoritmos, se encuentra que al poder modificar el modelo y optimizar las métricas utilizadas, consiguen un mejor rendimiento en términos de las medidas de equidad y exactitud, variando en función del algoritmo utilizado.

En cambio, como estos métodos modifican directamente el proceso de aprendizaje, a menudo son difíciles de generalizar a diferentes modelos o métricas. Además, en el ámbito de los problemas de *machine learning* en el mundo real, no siempre tendremos acceso al modelo de clasificación, por lo que esto supone un gran inconveniente en la optimización del mismo.

Otros ejemplos en la literatura

La equidad contrafactual presentada por Kusner et al. [2018], se basa en la noción de que un resultado debería ser el mismo independientemente del grupo demográfico del individuo. Definiendo una decisión como justa, si se mantiene igual cuando se cambia el valor del atributo protegido. El artículo plantea como objetivo la mitigación del sesgo, como un problema de optimización de equidad usando como base la inferencia causal. Este enfoque funciona bien para capturar los sesgos sociales e identificar el equilibrio entre equidad y utilidad.

Russell et al. [2017] presenta un artículo donde extiende el trabajo de Kusner et al. [2018], aportando un método para ofrecer predicciones justas con respecto a varios modelos causales simultáneos.

7.4 ALGORITMOS DE POSPROCESAMIENTO

Los *algoritmos de posprocesamiento* tienen como objetivo ajustar un clasificador ya entrenado, para que cumpla con unas restricciones de equidad específicas. Esto se suele hacer mediante la calibración de umbrales, cuya idea principal es encontrar un umbral adecuado utilizando una función de puntuación para cada grupo.

7.4.1 *Ejemplo: Aprendizaje en igualdad de oportunidades*

Hardt et al. [2016] desarrolla un marco para eliminar la discriminación de forma óptima para cualquier modelo de clasificación aprendido. Los autores definen la equidad como una restricción del concepto de probabilidades igualadas. Esta técnica, conocida como *predictor derivado*, utiliza la calibración del umbral para obtener el punto de la curva ROC, que cumple con los criterios de equidad establecidos.

Se utilizan diferentes valores de umbral para los distintos subgrupos, y se buscan soluciones factibles a lo largo de la intersección de cada intervalo convexo de la curva ROC correspondiente. Dado un clasificador y las correspondientes curvas ROC para ambos grupos. Podemos encontrar el umbral basado en las curvas (Figura 16).

El criterio de probabilidades igualadas se satisface únicamente cuando las curvas ROC de los dos grupos se cruzan, como se muestra en la imagen izquierda de la Figura 16; la igualdad de oportunidades, como relajación de la noción anterior, puede satisfacerse tomando un umbral tal que las tasas de verdaderos positivos de los dos grupos sean iguales, como puede verse en el gráfico de la derecha de la misma imagen.

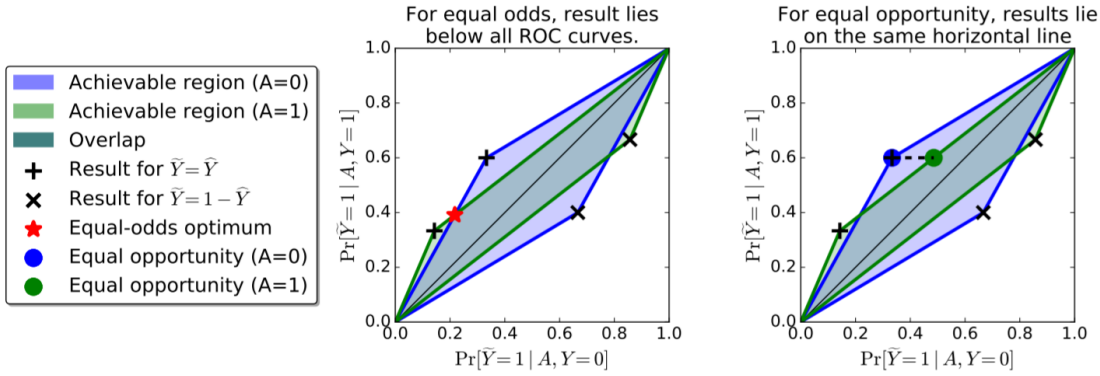


Figura 16: Búsqueda del clasificador óptimo para los criterios de probabilidades igualadas e igualdad de oportunidad, respectivamente. (Hardt et al. [2016])

Ventajas e inconvenientes

Los algoritmos de posprocesamiento comparten algunas ventajas con los de preprocesamiento, pudiéndose aplicar de forma independiente al modelo de clasificación sin necesidad de modificar su método de actuación. Además, consigue un rendimiento relativamente bueno en la optimización de la mayoría de definiciones de equidad (a excepción de la equidad contrafactual).

Sin embargo, se puede argumentar que al actuar sobre el modelo después de haberlo aprendido, este proceso es intrínsecamente subóptimo. Siendo equivalente a aprender a sabiendas un modelo sesgado y luego corregirlo, en lugar de aprender un modelo insesgado desde el principio.

Otros ejemplos en la literatura

En el artículo presentado por Woodworth et al. [2017] se amplía el trabajo de Hardt et al. [2016], demostrando que su método de posprocesamiento podría ser subóptimo en algunos casos. Se ofrece una demostración de que el problema es intratable, y se presenta un método que aproxima el criterio de equidad proporcionando un resultado estadísticamente cercano al óptimo.

Parte III

FUNDAMENTOS DE LA EQUIDAD CONTRAFACTUAL

Discusión sobre la inferencia causal, el cálculo de contrafactuales y el teorema de incompatibilidad como fundamentos matemáticos de la equidad contrafactual.

8

INFERENCIA CAUSAL

En este capítulo formalizaremos algunos conceptos básicos para desarrollar la teoría relativa a la causalidad en el ámbito de la equidad. Además, introduciremos los grafos como herramienta para describir los modelos causales explicados, así como los efectos que tienen estos modelos en las poblaciones donde se aplican.

8.1 MODELOS CAUSALES

Elegiremos los *modelos causales estructurales* aprovechando que pueden ofrecernos una base sólida para las diferentes nociones causales utilizadas en este trabajo. La forma más sencilla de conceptualizar un modelo causal estructural, es como un programa que genera una distribución a partir de *variables de ruido* independientes mediante una secuencia de instrucciones formales.

Imaginemos que en lugar de muestras de una distribución, tenemos un programa informático que genera muestras a partir de una semilla aleatoria. El código de este programa, partiría de una semilla aleatoria simple e iría construyendo muestras cada vez más complejas. Esta idea es la misma que utiliza un modelo causal estructural cambiando la sintaxis de programación por lenguaje matemático.

8.1.1 *Ejemplo: Construcción de un modelo causal*

Supongamos una población en la que un individuo hace ejercicio regularmente con una probabilidad de $\frac{1}{2}$. Con una probabilidad de $\frac{1}{3}$, el individuo tiene predisposición a desarrollar sobrepeso en ausencia de ejercicio regular. Del mismo modo, en ausencia de ejercicio, la aparición de una enfermedad cardíaca puede aparecer con una probabilidad de $\frac{1}{3}$. Denotaremos por X el indicador de ejercicio regular, por Y el de exceso de peso, y por Z el indicador de la enfermedad cardíaca. A continuación, construiremos un modelo causal estructural para generar muestras de esta población hipotética (Barocas et al. [2019]).

Algoritmo 1: Programa distribución causal 1.

Muestras de variables aleatorias independientes de Bernoulli:
$U_1 \sim Bernoulli\left(\frac{1}{2}\right)$, $U_2, U_3 \sim Bernoulli\left(\frac{1}{3}\right)$;
$X := U_1$;
$Y :=$ **if** $X = 1$ **then** 0 **else** U_2;
$Z :=$ **if** $X = 1$ **then** 0 **else** U_3;

A partir de la descripción anterior, observamos que en nuestra población el ejercicio evita tanto el sobrepeso como las enfermedades cardíacas, pero en ausencia de ejercicio, ambos son independientes. Nuestro programa genera una distribución conjunta sobre las variables aleatorias (X, Y, Z). Podemos calcular las probabilidades bajo esta distribución. Por ejemplo, la probabilidad de sufrir una enfermedad cardíaca bajo la distribución especificada por nuestro modelo es de $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. También podemos calcular la probabilidad condicional de padecer enfermedades cardíacas dado el sobrepeso. Dado el suceso $Y = 1$ podemos inferir que el individuo no hace ejercicio $X = 0$, por lo que la probabilidad de sufrir una enfermedad cardíaca debido al sobrepeso es $\frac{1}{3}$.

Formalmente, tener un programa que genere una distribución es más potente que el simple acceso al muestreo. Una de las razones es que podemos manipular el programa de la manera que queramos, mientras resulte en un programa funcional. Podríamos, por ejemplo, establecer $Y := 1$, dando lugar a una nueva distribución. El programa resultante tiene el siguiente aspecto:

Algoritmo 2: Programa distribución causal 2.

Muestras de variables aleatorias independientes de Bernoulli:
$U_1 \sim Bernoulli\left(\frac{1}{2}\right)$, $U_2, U_3 \sim Bernoulli\left(\frac{1}{3}\right)$;
$X := U_1$;
$Y := 1$;
$Z :=$ **if** $X = 1$ **then** 0 **else** U_3;

Calculando de nuevo la probabilidad de sufrir una enfermedad cardíaca sobre la nueva distribución, de nuevo obtenemos $\frac{1}{6}$. Este cálculo revela una idea importante, la sustitución $Y := 1$ no corresponde a un condicionamiento de $Y = 1$. Una se trata de una acción y la otra es una observación de la que podemos extraer conclusiones. En este ejemplo, si observamos que un individuo tiene sobrepeso, podemos inferir que tiene un mayor riesgo de enfermedad cardíaca. Sin embargo, esto no significa que la reducción del peso corporal evite las enfermedades cardíacas. En cambio, la intervención $Y := 1$ crea un nuevo modelo en el que todos los individuos de la población tienen sobrepeso con todo lo que ello conlleva.

A continuación, profundizaremos un poco más en este punto considerando otra población hipotética, especificada por el siguiente programa:

Algoritmo 3: Programa distribución causal 3.

Muestras de variables aleatorias independientes de Bernoulli:
$U_1 \sim Bernoulli\left(\frac{1}{2}\right)$, $U_2, U_3 \sim Bernoulli\left(\frac{1}{3}\right)$;
$Y := U_2$;
$X :=$ **if** $Y = 0$ **then** 0 **else** U_1;
$Z :=$ **if** $X = 1$ **then** 0 **else** U_3;

En esta población, la única razón por la que los individuos eligen hacer ejercicio con cierta probabilidad es el sobrepeso. Por otro lado, las enfermedades del corazón se desarrollan en ausencia de ejercicio. La sustitución $Y := 1$ en este modelo conduce a un aumento de la probabilidad de hacer ejercicio y por tanto, a una disminución de la probabilidad de sufrir una enfermedad cardíaca. El condicionamiento de $Y = 1$ también tiene el mismo efecto y en ambos casos, la probabilidad de sufrir un problema cardíaco es de $\frac{1}{6}$.

8.1.2 *Formalización de los modelos causales estructurales*

Los modelos causales estructurales nos proporcionan un cálculo preciso para razonar sobre el efecto de las acciones hipotéticas. Formalmente, un modelo causal estructural es una secuencia de asignaciones que generan una distribución conjunta a partir de variables de ruido independientes. A continuación ofreceremos la definición de modelo causal estructural presentada por Pearl [2000].

Definición 64 (Modelo causal estructural)**.** Un *modelo causal estructural* M se define como una tupla (U, V, F) de conjuntos tales que:

- U es un conjunto de variables aleatorias de ruido, las cuales deben ser conjuntamente independientes. Corresponden a factores no causados por ninguna variable del conjunto V de variables observadas.
- F es un conjunto de funciones $\{f_1, \ldots, f_n\}$, una para cada $V_i \in V$, tal que,

$$V_i = f_i(pa_i, U_i), \text{ para todo } i = 1, \ldots, n.$$

donde $pa_i \subseteq V \setminus \{V_i\}$ y $U_i \subseteq U$. Estas ecuaciones también son conocidas como *ecuaciones estructurales*.

El modelo es causal en el sentido de que, dada una distribución de probabilidad $P(U)$ sobre las variables de ruido U, podemos derivar la distribución de un subconjunto $W \subseteq V$ tras una intervención en $V \setminus W$. Cuando M denota un modelo causal estructural, escribiremos la probabilidad de un evento E bajo la distribución conjunta vinculada como $P_M(E)$.

Para familiarizarnos con la notación, supongamos que M denota el modelo causal estructural del Apartado 8.1.1, entonces la probabilidad de sufrir una enfermedad cardíaca en este modelo será $P_M(Z) = \frac{1}{6}$.

8.2 GRAFOS CAUSALES

La notación pa_i utilizada en la Definición 64, se refiere al subconjunto de variables $pa(V_i)$ que contiene los padres del nodo V_i. Esta notación viene motivada por la suposición de que el modelo se factoriza como un grafo dirigido, el cual en este trabajo, restringiremos al caso acíclico (DAG). A este grafo lo llamaremos: el *grafo o diagrama causal* correspondiente al modelo causal estructural especificado.

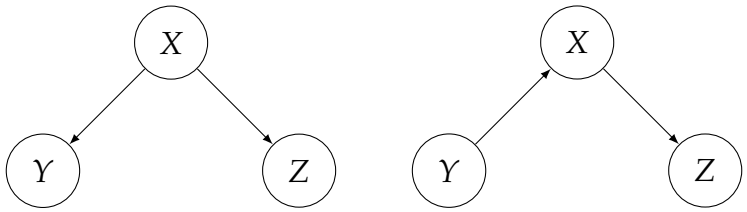


Figura 17: Grafos causales de los modelos descritos por los Programas 1 y 3, respectivamente.

Los diagramas causales se utilizan cuando las asignaciones exactas en un modelo causal estructural son secundarias, y lo que es realmente relevante son los caminos presentes y ausentes entre nodos. Los grafos también nos permiten aprovechar el lenguaje de teoría de grafos para discutir nociones causales. En particular, los grafos causales nos ayudan a distinguir la causa y el efecto (de tipo directo o indirecto), en función de si un nodo es ancestro o descendiente de otro.

8.2.1 *Forks*

Definición 65 (*Fork*)**.** Sea G un grafo acíclico dirigido, U el camino entre dos nodos y $A \in U$. Llamaremos *fork* al nodo A si $(A, B) \in E$, para todo $B \in ve(A) \cap U$.

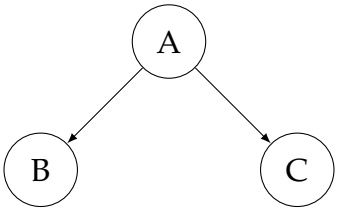


Figura 18: Ejemplo de *fork*.

En la Figura 18 el nodo A es un ejemplo de *fork* o dicho de otro modo, el nodo A es la causa común de los nodos B y C. En el grafo causal resultante de la distribución del Programa 1, el nodo X también es un ejemplo de *fork* ($Y \leftarrow X \rightarrow Z$). En ese caso la variable indicadora de ejercicio regular X influía, tanto en el aumento de peso Y, como en el riesgo de enfermedad Z. Sin embargo, como ya discutimos en el Apartado 8.1.1, las variables Y y Z no están correlacionadas positivamente. Llegamos a la conclusión de que el nodo *fork*, tiene un efecto de confusión que conduce a un desacuerdo entre el cálculo de las probabilidades condicionales y las intervenciones.

Ejemplo 10. En un conocido estudio médico, un presunto efecto beneficioso de la terapia de sustitución hormonal para reducir las enfermedades cardiovasculares, desapareció tras identificar el estatus socioeconómico como variable de confusión (Humphrey et al. [2002]). Los ejemplos de confusión suponen una amenaza para la validez de las conclusiones extraídas de los datos en problemas del mundo real.

8.2.2 *Colliders*

Definición 66 (*Collider*)**.** Sea G un grafo acíclico dirigido, U el camino entre dos nodos y $A \in U$. Llamaremos *collider* al nodo a A si $(B, A) \in E$, para todo $B \in ve(A) \cap U$.

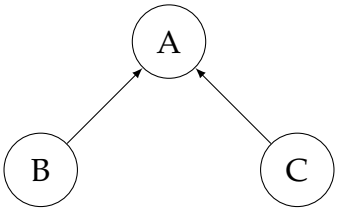


Figura 19: Ejemplo de *collider*.

En la Figura 19 el nodo A es un ejemplo de *collider*. Cabe destacar que los *colliders*, no dan lugar a situaciones en las que se pueda dar confusión. De hecho en la figura anterior, la relación entre B y C no es confusa, lo que significa que podemos sustituir las intervenciones por probabilidades condicionales. Sin embargo, condicionar un *collider* podría crear una correlación entre B y C, un fenómeno al que denominaremos sesgo de *collider*.

Ejemplo 11. En el ámbito sanitario, dos enfermedades independientes pueden correlacionarse negativamente cuando se analizan pacientes hospitalizados. La razón es que cuando cualquiera de las dos enfermedades (B o C) es suficiente para el ingreso en el hospital (indicado por la variable A), observar que un paciente tiene una enfermedad hace que la otra sea estadísticamente menos probable. A esto es lo que se le conoce como paradoja de Berkson (Berkson [2014]).

8.2.3 *Mediador*

En la definición de *fork*, no tenemos una relación directa entre los nodos B y C. Si queremos un efecto total de B sobre C estableceremos esta relación causal a través de A. En este caso, A no será un factor confusión y recibirá el nombre de *mediador*.

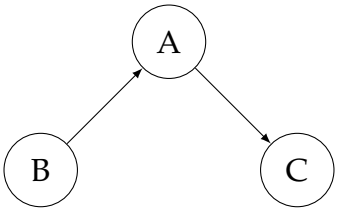


Figura 20: Ejemplo de mediador.

En el grafo causal asociado a la distribución generada por el Programa 3, el nodo X es un ejemplo de mediador ($Y \to X \to Z$). La noción de mediador es especialmente relevante para el tema del análisis de la discriminación, ya que establece una relación directa entre las diferentes variables que definen a un individuo. Esta relación causal nos servirá como herramienta para extraer conclusiones sobre las causas de segregación entre grupos.

8.3 INTERVENCIÓN Y CONFUSIÓN

Los modelos causales estructurales nos proporcionan una herramienta de formalización del efecto de acciones e intervenciones sobre la población donde se aplican. Como hemos visto previamente, para modelar estos efectos simplemente necesitamos la capacidad de realizar sustituciones.

8.3.1 *Operadores para realizar actuaciones en el modelo*

A partir de los ejemplos propuestos en el Apartado 8.1.1, hemos observado que fijar una variable por sustitución puede corresponder o no a una probabilidad condicional. Esto refuerza nuestra intuición de que una observación no es una acción. En cambio, una sustitución sí es una acción, ya que al sustituir un valor estamos rompiendo el curso natural de la acción captada por nuestro modelo.

Definición 67 (Intervención)**.** Dado un modelo causal estructural M, se define una *intervención* sobre una variable observada X, como la sustitución de la ecuación $X := f(pa, U)$ por la ecuación $X := x$ para un valor x constante.

Denotaremos el modelo resultante por $M' = M[X := x]$, para indicar la modificación que realizamos sobre el modelo original M. Bajo esta asignación mantenemos X constante, eliminando la influencia de sus nodos padres, y por tanto de cualquier otra variable del modelo. Por otra parte, los nodos hijos de X recibirán un valor constante x cuando consulten el valor de su padre.

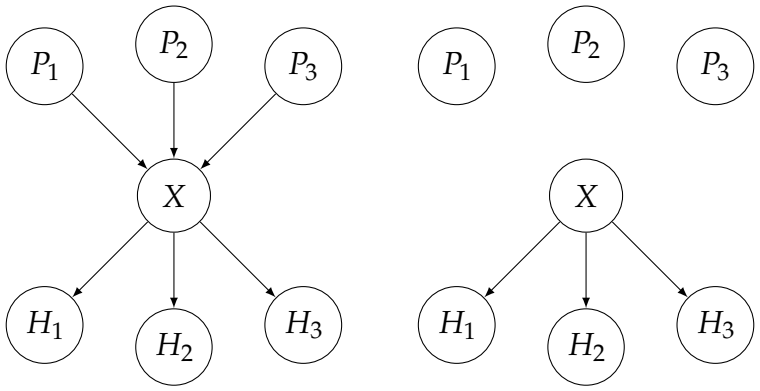


Figura 21: Grafo causal antes y después de la sustitución.

El operador de asignación también se denomina *operador do*, para destacar que corresponde a la realización de una acción o intervención. La notación que usaremos para calcular las probabilidades dado un evento cualquiera E después de aplicar el *operador do*, será $P_{M[X:=x]}(E)$. También podemos utilizar otra notación equivalente, que aproxima el concepto de probabilidad condicional, y que se define como:

$$P(E \mid \text{do}(X := x)) = P_{M[X:=x]}(E).$$

8.3.2 *Confusión entre dos variables*

Las cuestiones importantes en inferencia causal, están vinculadas a cuándo podemos reescribir una *operación do* en términos de probabilidades condicionales. Cuando esto sea posible, podremos estimar el efecto de la operación a partir de las probabilidades condicionales estimadas de los datos.

Sea una variable Y, sobre la que actúa una variable X, nos interesará que exista una equivalencia entre el efecto causal de la acción y la probabilidad condicional correspondiente, es decir, que se cumpla la siguiente igualdad:

$$P(Y = y \mid \text{do}(X := x)) = P(Y = y \mid X = x).$$

En general, esto no es cierto. Al fin y al cabo, la diferencia entre la observación (probabilidad condicional) y la acción (intervención) es la principal motivación de la inferencia causal.

Definición 68 (Confusión)**.** Sean Y una variable aleatoria sobre la que actúa otra variable X, diremos que son *confusas* si, y solo si,

$$P(Y = y \mid \text{do}(X := x)) \neq P(Y = y \mid X = x).$$

Cuando tenemos dos variables aleatorias confusas, podemos estimar el efecto de una intervención en términos de probabilidades condicionales a partir de la denominada *fórmula de ajuste*.

Proposición 17. Sean X, Y dos variables confusas, podemos aproximar el efecto causal de una intervención dada a partir de probabilidades condicionales como:

$$P(Y = y \mid \text{do}(X := x)) = \sum_z P(Y = y \mid X = x, PA = z)P(PA = z),$$

donde PA indica el conjunto $pa(X)$.

Dependiendo de la estructura del grafo, podremos eliminar o no la confusión entre dos variables utilizando la fórmula de ajuste sobre un nodo u otro. Si el grafo tiene una estructura de *fork* (por ejemplo, $B \leftarrow A \rightarrow C$), eliminaremos la confusión entre los nodos B y C, condicionando A. En cualquier otro caso (mediador o *collider*), ajustar una variable tendría consecuencias opuestas a las que buscamos.

Criterio de backdoor

El tratamiento de la confusión a partir de la fórmula de ajuste, puede ser una tarea complicada cuando la cantidad de nodos en el grafo aumente considerablemente. Para detectar las variables sobre las que deberemos condicionar, aparece el *criterio de backdoor* (Pearl [2000]). Este método parte de la idea de seleccionar un conjunto de variables, que bloqueen todos los *caminos de backdoor* entre los dos nodos sobre los que queremos eliminar la confusión.

Definición 69 (Camino de *backdoor*)**.** Un *camino de backdoor* entre dos nodos A y B, es cualquier camino que empiece con una arista de la forma "$\leftarrow$" hacia A.

Definición 70 (Conjunto de *backdoor*)**.** Un *conjunto de backdoor*, es una secuencia de variables o nodos contenida en un camino de *backdoor*.

Para aplicar el criterio de *backdoor*, primero seleccionaremos un conjunto de *backdoor* del grafo. Si el conjunto está formado por una secuencia de nodos relacionados únicamente por aristas de tipo "$\rightarrow$", podremos eliminar la confusión entre las variables, aplicando la fórmula de ajuste sobre un nodo central de la cadena. Por otro lado, si el camino contiene un *collider* o un descendiente de este, la confusión es inevitable, ya

que bloqueando el camino podríamos impedir que la información fluyera a través de los nodos.

Ejemplo 12. Sea un grafo causal dado por la secuencia $A \leftarrow C \rightarrow D \rightarrow E \rightarrow B$, nuestro objetivo será eliminar la confusión entre las variables A y B. Es evidente que la cadena $A \leftarrow C \rightarrow D \rightarrow E \rightarrow B$ es un camino de *backdoor*. A continuación, seleccionamos un conjunto de *backdoor* entre ambos nodos, por ejemplo $C \rightarrow D \rightarrow E$. En vista de la forma de la secuencia anterior, podemos eliminar la confusión entre A y B aplicando la fórmula de ajuste sobre el nodo D.

Confusión no observada

La fórmula de ajuste presentada en la Proposición 17, podría sugerir que siempre podemos eliminar el sesgo de confusión condicionando a los nodos padres. Sin embargo, esto no se cumple cuando aparecen *factores de confusión no observados*. En la práctica, a menudo hay variables que son difíciles de medir o que no fueron resgistradas. Podemos incluir estos nodos no observados en un grafo, indicando su influencia con líneas discontinuas.

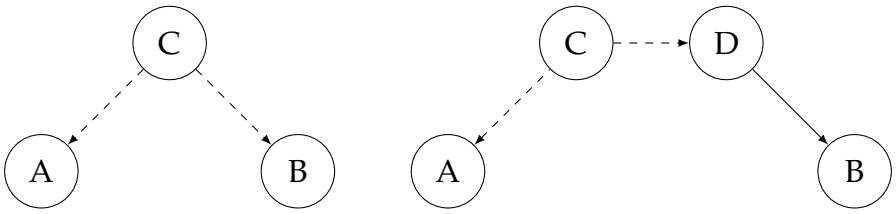


Figura 22: Ejemplos de confusión no observada.

La Figura 22 muestra dos casos de confusión no observada. En el primer ejemplo, el efecto causal de A sobre B no es identificable. En el segundo caso, podemos eliminar la confusión entre A y B a partir del criterio de *backdoor*. Sea $C \rightarrow D \rightarrow B$ un conjunto de *backdoor*, podemos eliminar la confusión entre las variables A y B ajustando la fórmula sobre la variable D aunque C no se observe.

Cabe destacar que podemos combatir la confusión no observada aumentando el número de variables consideradas, pero esto aumentaría progresivamente la complejidad de nuestro modelo causal. En la práctica, es habitual controlar el mayor número posible de variables con el objetivo de eliminar el sesgo de confusión. Sin embargo, como hemos visto, el control de mediadores y *colliders* podría ser problemático en la resolución de nuestro problema.

9

TEOREMA DE IMPOSIBILIDAD DE LA EQUIDAD

En este capítulo se propondrá una demostración alternativa del teorema de imposibilidad de equidad. Este enunciado surge como una formalización de la *incompatibilidad* entre los criterios de paridad demográfica, probabilidades igualadas y tasa de paridad predictiva.

9.1 CARACTERIZACIÓN DEL TEOREMA

La mayoría de los criterios de equidad definidos en el Capítulo 6, se construyen a partir de restricciones no triviales de la distribución de probabilidad conjunta. Por ello, es lógico pensar que la imposición de varios de ellos de forma simultánea, restringirían el espacio de búsqueda hasta el punto de que solo obtendríamos soluciones degeneradas.

El teorema de la imposibilidad, cuya primera aproximación fue ofrecida por Kleinberg et al. [2016], establece que no se puede satisfacer más de una medida de equidad al mismo tiempo para un clasificador bien entrenado y un atributo sensible que sea capaz de introducir un sesgo en el modelo. En nuestro caso, presentaremos una versión del teorema para tres de los criterios de equidad de grupo estudiados: paridad demográfica, tasa de paridad predictiva y probabilidades igualadas. Los enunciados de los lemas demostrados, a lo largo de este capítulo, han sido definidos en base al trabajo de Barocas et al. [2019].

9.1.1 *Paridad demográfica vs. Tasa de paridad predictiva*

Comenzamos con un lema que muestra cómo, en general, la paridad demográfica y la paridad predictiva se excluyen mutuamente. La única suposición necesaria es que el atributo sensible A y la variable Y no sean independientes, es decir, dependen una de la otra. Esto es una forma diferente de decir, que un grupo tiene mayor tasa de resultados positivos que otro, lo que es cierto en la mayoría de casos.

Lema 1. Supongamos que A e Y son variables dependientes. Entonces, la paridad demográfica ($\hat{Y} \perp A$) y la tasa de paridad predictiva ($Y \perp A \mid \hat{Y}$), no pueden verificarse simultáneamente.

Demostración. El enunciado del lema es análogo a la siguiente expresión,

$$Y \not\perp A \implies \neg(\hat{Y} \perp A \wedge Y \perp A \mid \hat{Y}).$$

Procederemos por contrarrecíproco, lo que equivale a demostrar que,

$$\hat{Y} \perp A \wedge Y \perp A \mid \hat{Y} \implies Y \perp A.$$

Si se da la independencia entre las variables A e $\hat{Y}$, entonces se cumple,

$$P(A = a, \hat{Y} = \hat{y}) = P(A = a)P(\hat{Y} = \hat{y}). \tag{9}$$

Por otro lado, sabemos que la independencia condicional dada por $Y \perp A \mid \hat{Y}$ satisface que,

$$P(Y = y, A = a \mid \hat{Y} = \hat{y}) = P(Y = y \mid \hat{Y} = \hat{y})P(A = a \mid \hat{Y} = \hat{y}). \tag{10}$$

Finalmente, aplicando las hipótesis del enunciado y usando el teorema de probabilidad total (T.P.T) sobre $P(Y = y, A = a)$, llegamos a la siguiente expresión:

$$\begin{aligned}
P(Y = y, A = a) &\overset{\text{T.P.T}}{=} \sum_{\hat{y}} P(\hat{Y} = \hat{y})P(Y = y, A = a \mid \hat{Y} = \hat{y}) \\
&\overset{(10)}{=} \sum_{\hat{y}} P(\hat{Y} = \hat{y})P(Y = y \mid \hat{Y} = \hat{y})P(A = a \mid \hat{Y} = \hat{y}) \\
&= \sum_{\hat{y}} P(\hat{Y} = \hat{y})P(Y = y \mid \hat{Y} = \hat{y})\frac{P(A = a, \hat{Y} = \hat{y})}{P(\hat{Y} = \hat{y})} \\
&\overset{(9)}{=} \sum_{\hat{y}} P(\hat{Y} = \hat{y})P(Y = y \mid \hat{Y} = \hat{y})\frac{P(A = a)P(\hat{Y} = \hat{y})}{P(\hat{Y} = \hat{y})} \\
&= \sum_{\hat{y}} P(\hat{Y} = \hat{y})P(Y = y \mid \hat{Y} = \hat{y})P(A = a) \\
&= P(A = a)\sum_{\hat{y}} P(\hat{Y} = \hat{y})P(Y = y \mid \hat{Y} = \hat{y}) \\
&\overset{\text{T.P.T}}{=} P(A = a)P(Y = y).
\end{aligned}$$

La última igualdad nos da que las variables A e Y son independientes y por tanto tenemos que $Y \perp A$. □

9.1.2 *Paridad demográfica vs. Probabilidades igualadas*

Un resultado análogo de exclusión mutua, es válido para la paridad demográfica y el criterio de probabilidades igualadas. El enunciado, en este caso, es un poco más rebuscado y requiere la suposición adicional de que la variable Y sea binaria. También necesitamos que la variable $\hat{Y}$ dependa de Y. Esta suposición es una relajación bastante suave, ya que cualquier función de clasificación útil tiene correlación con la variable Y.

Lema 2. Supongamos que Y es una variable binaria, A e Y son dependientes y además, Y también depende de $\hat{Y}$. Entonces, la paridad demográfica ($\hat{Y} \perp A$) y el criterio de las probabilidades igualadas ($\hat{Y} \perp A \mid Y$), no pueden verificarse simultáneamente.

Demostración. El enunciado del lema es equivalente a la siguiente expresión,

$$Y \not\perp A \ \wedge \ Y \not\perp \hat{Y} \quad \Longrightarrow \quad \neg(\hat{Y} \perp A \ \wedge \ \hat{Y} \perp A \mid Y).$$

Por el contrarrecíproco, deberemos demostrar que,

$$\hat{Y} \perp A \ \wedge \ \hat{Y} \perp A \mid Y \quad \Longrightarrow \quad Y \perp A \ \vee \ Y \perp \hat{Y}.$$

Si se da la independencia entre las variables A e $\hat{Y}$, entonces se cumple,

$$P(\hat{Y} = \hat{y}, A = a) = P(\hat{Y} = \hat{y})P(A = a). \tag{11}$$

Sabemos que la independencia condicional dada por $\hat{Y} \perp A \mid Y$ satisface que,

$$P(\hat{Y} = \hat{y} \mid A = a, Y = y) = P(\hat{Y} = \hat{y} \mid Y = y). \tag{12}$$

Aplicando el teorema de la probabilidad total (T.P.T) y la hipótesis de independencia condicional sobre $P(\hat{Y} = \hat{y} \mid A = a)$, obtenemos la siguiente expresión:

$$\begin{aligned} P(\hat{Y} = \hat{y} \mid A = a) &\overset{\text{T.P.T}}{=} \sum_y P(Y = y \mid A = a)P(\hat{Y} = \hat{y} \mid A = a, Y = y) \\ &\overset{(12)}{=} \sum_y P(Y = y \mid A = a)P(\hat{Y} = \hat{y} \mid Y = y). \end{aligned} \tag{13}$$

Usando la hipótesis de independencia entre las variables A e $\hat{Y}$,

$$\begin{aligned} P(\hat{Y} = \hat{y} \mid A = a) &= \frac{P(\hat{Y} = \hat{y}, A = a)}{P(A = a)} \\ &\overset{(11)}{=} \frac{P(\hat{Y} = \hat{y})P(A = a)}{P(A = a)} \\ &= P(\hat{Y} = \hat{y}). \end{aligned} \tag{14}$$

Combinando las Ecuaciones (13) y (14), llegamos a la siguiente expresión:

$$P(\hat{Y} = \hat{y}) = \sum_{y} P(Y = y \mid A = a) P(\hat{Y} = \hat{y} \mid Y = y). \tag{15}$$

Por otro lado, aplicando el teorema de la probabilidad total sobre $P(\hat{Y} = \hat{y})$, tenemos que,

$$P(\hat{Y} = \hat{y}) = \sum_{y} P(Y = y) P(\hat{Y} = \hat{y} \mid Y = y). \tag{16}$$

Combinando las Ecuaciones (15) y (16), conseguimos la expresión dada por:

$$\sum_{y} P(Y = y \mid A = a) P(\hat{Y} = \hat{y} \mid Y = y) = \sum_{y} P(Y = y) P(\hat{Y} = \hat{y} \mid Y = y). \tag{17}$$

A continuación, y para que sea más cómodo de manipular la expresión anterior definiremos la siguiente notación:

$$\begin{aligned} p &= P(Y = 0), \\ p_a &= P(Y = 0 \mid A = a), \\ \hat{y}_y &= P(\hat{Y} = \hat{y} \mid Y = y). \end{aligned}$$

Por hipótesis del Lema, Y es una variable binaria (supongamos que puede tomar los valores 0 o 1) y por tanto, podemos reescribir la Ecuación (17) como:

$$p\hat{y}_0 + (1 - p)\hat{y}_1 = p_a\hat{y}_0 + (1 - p_a)\hat{y}_1.$$

Simplificando en la ecuación anterior, tenemos que,

$$p(\hat{y}_0 - \hat{y}_1) = p_a(\hat{y}_0 - \hat{y}_1),$$

lo cual es equivalente a,

$$(p - p_a)(\hat{y}_0 - \hat{y}_1) = 0. \tag{18}$$

La igualdad de la Ecuación (18), se satisface si se da alguno de los siguientes casos:

- $p = p_a$, que es equivalente a $P(Y = 0) = P(Y = 0 \mid A = a)$, y por tanto tenemos,

  $$\begin{aligned} P(Y = 1) &= 1 - P(Y = 0) \\ &= 1 - P(Y = 0 \mid A = a) \\ &= P(Y = 1 \mid A = a), \end{aligned}$$

  donde la expresión anterior equivale a que $Y \perp A$.

- $\hat{y}_0 = \hat{y}_1$, que equivale a $P(\hat{Y} = \hat{y} \mid Y = 0) = P(\hat{Y} = \hat{y} \mid Y = 1)$, y por tanto tenemos que $\hat{Y} \perp Y$.

□

9.1.3 *Probabilidades igualadas vs. Tasa de paridad predictiva*

Por último, pasamos a la relación entre la tasa de paridad predictiva y el criterio de probabilidades igualadas. Ambas exigen una relación de independencia condicional no trivial entre las tres variables A, $\hat{Y}$ e Y. Imponer ambas simultáneamente, conduce a un espacio de soluciones degenerado, como afirma el corolario siguiente.

Corolario 1. Supongamos que todos los sucesos en la distribución conjunta $(A, \hat{Y}, Y)$ tienen probabilidad positiva, y además A depende de Y. Entonces, el criterio de probabilidades igualadas ($\hat{Y} \perp A \mid Y$) y la tasa de paridad predictiva ($Y \perp A \mid \hat{Y}$), no pueden verificarse simultáneamente.

Demostración. El enunciado del corolario es análogo a la siguiente expresión,

$$Y \not\perp A \implies \neg(\hat{Y} \perp A \mid Y \land Y \perp A \mid \hat{Y}).$$

Por el contrarrecíproco, tenemos que probar que,

$$\hat{Y} \perp A \mid Y \land Y \perp A \mid \hat{Y} \implies Y \perp A.$$

Por la propiedad simétrica de la independencia condicional, tenemos que,

$$\hat{Y} \perp A \mid Y = A \perp \hat{Y} \mid Y,$$
$$Y \perp A \mid \hat{Y} = A \perp Y \mid \hat{Y}.$$

La hipótesis dada por $A \perp \hat{Y} \mid Y$ satisface que,

$$P(A = a \mid \hat{Y} = \hat{y}, Y = y) = P(A = a \mid Y = y). \tag{19}$$

Por otro lado, sabemos que $A \perp Y \mid \hat{Y}$ cumple que,

$$P(A = a \mid Y = y, \hat{Y} = \hat{y}) = P(A = a \mid \hat{Y} = \hat{y}). \tag{20}$$

Para que tengan sentido las Ecuaciones (19) y (20), es necesaria la hipótesis de que $P(\hat{Y} = \hat{y}, Y = y) > 0$ que a su vez, es consecuencia directa de que la distribución conjunta $(A, \hat{Y}, Y)$ tenga probabilidad positiva.

Por la propiedad de unión débil de la independencia condicional, tenemos que las Ecuaciones (19) y (20) son equivalentes, y por tanto:

$$P(A = a \mid Y = y) = P(A = a \mid \hat{Y} = \hat{y}). \tag{21}$$

Usando la definición de probabilidad condicionada sobre $P(A = a \mid Y = y)$, tenemos que,

$$P(A = a \mid Y = y) = \frac{P(A = a, Y = y)}{P(Y = y)}, \tag{22}$$

queremos demostrar que $Y \perp A$, es decir, $P(A = a, Y = y) = P(A = a)P(Y = y)$, que aplicado a la Ecuación (22), equivale a probar:

$$P(A = a \mid Y = y) = P(A = a).$$

Usando el teorema de la probabilidad total (T.P.T) sobre $P(A)$, conseguimos la siguiente expresión:

$$\begin{aligned} P(A = a) &\overset{\text{T.P.T}}{=} \sum_{\hat{y}} P(\hat{Y} = \hat{y}) P(A = a \mid \hat{Y} = \hat{y}) \\ &\overset{(21)}{=} \sum_{\hat{y}} P(\hat{Y} = \hat{y}) P(A = a \mid Y = y) \\ &= P(A = a \mid Y = y) \sum_{\hat{y}} P(\hat{Y} = \hat{y}) \\ &= P(A = a \mid Y = y). \end{aligned}$$

□

Para un objetivo binario, la hipótesis de no degeneración del Corolario 1 establece que en todos los grupos, para todos los valores de predicción, tenemos casos positivos y negativos. En caso de que el clasificador sea binario, podemos debilitar la suposición exigiendo que el clasificador haga al menos una predicción falsa positiva. Lo atractivo de la afirmación resultante es, que su prueba se basa esencialmente en una relación bastante popular entre la tasa de verdaderos positivos (*Recall*) y el valor predictivo positivo (*Precision*).

Lema 3. Supongamos que $\hat{Y}$ es una variable binaria que toma valores de un clasificador con una tasa de falsos positivos no nula, y además A dependiente de Y. Entonces, el criterio de probabilidades igualadas ($\hat{Y} \perp A \mid Y$) y la tasa de paridad predictiva ($Y \perp A \mid \hat{Y}$), no pueden verificarse simultáneamente.

Demostración. (Barocas et al. [2019]). Dado que $Y \not\perp A$, existirán dos grupos, a los que llamaremos p_0 y p_1 cumpliendo que $p_0 \neq p_1$, donde $p_a = P(Y = 1 \mid A = a)$.

Supondremos que se satisface el criterio de las probabilidades igualadas. Por hipótesis, el clasificador tendrá la misma tasa para todos los grupos de falsos y verdaderos positivos (FPR,TPR > 0). Procederemos a demostrar que la tasa de paridad predictiva no se satisface en estas condiciones.

En el caso binario, la tasa de paridad predictiva implica que todos los grupos tienen el mismo de PPV (ver Apartado 6.4.3). El valor predictivo positivo en el grupo a, denotado PPV_a satisface

$$\text{PPV}_a = \frac{\text{TPR} p_a}{\text{TPR} p_a + \text{FPR}(1 - p_a)}.$$

De la expresión anterior vemos que: $\text{PPV}_0 = \text{PPV}_1$ si, y solo si, $\text{TPR} = 0$ o $\text{FPR} = 0$. Descartaremos esto último por hipótesis. Por tanto, se debe cumplir que $\text{TPR} = 0$. Sin embargo, podemos deducir que $\text{NPV}_0 \neq \text{NPV}_1$ a partir de la siguiente expresión,

$$\text{NPV}_a = \frac{(1 - \text{FPR})(1 - p_a)}{(1 - \text{TPR})p_a + (1 - \text{FPR})(1 - p_a)}.$$

Por tanto, la tasa de paridad predictiva no se satisface. □

9.1.4 *Enunciado y demostración*

Una vez demostrados los resultados previos, estamos preparados para enunciar y demostrar la versión del teorema de imposibilidad para la equidad de grupo.

Teorema 4 (Teorema de imposibilidad de la equidad)**.** Consideremos un problema de clasificación binaria con una tasa de falsos positivos no nula, donde se cumple la siguiente relación de dependencia entre las variables: $Y \not\perp A$ e $\hat{Y} \not\perp Y$. Si existe una asignación de riesgo, entonces ésta no puede satisfacer los criterios de paridad demográfica, probabilidades igualadas y paridad predictiva simultáneamente dos a dos.

Demostración. Bajo las hipótesis del teorema, aplicamos los Lemas 1, 2 y 3. □

9.2 PERSPECTIVA CAUSAL

Podemos consultar una versión diferente de la demostración del Teorema 4, desde la perspectiva de la inferencia causal, en el artículo propuesto por Saravanakumar [2021]. Además, el artículo presenta una propuesta teórica de corrección de los datos de entrenamiento, con el objetivo de mitigar el sesgo presente en el mismo y ajustarlo a las nociones de equidad actuales que pudiesen cambiar con el tiempo.

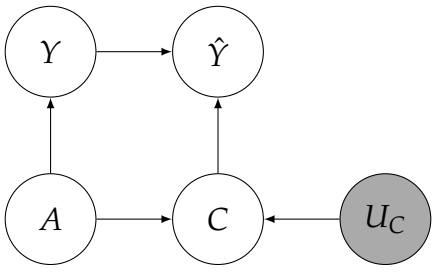


Figura 23: Diagrama causal con la variable de corrección.

El diagrama causal para el proceso de generación de datos puede representarse, como podemos observar en la Figura 23, introduciendo una variable de corrección C que determinará si la información de la etiqueta verdadera Y debe propagarse a la predicción $\hat{Y}$ o no.

La variable de corrección se define como una función dependiente del atributo sensible ($C \not\perp A$), pero que no permite el paso de la información de A a través de ella. Por otro lado, el flujo de información de Y a $\hat{Y}$ no tiene que estar necesariamente bloqueado para el grupo favorecido, esto es deseable si queremos evitar el sesgo sobre el grupo marginado.

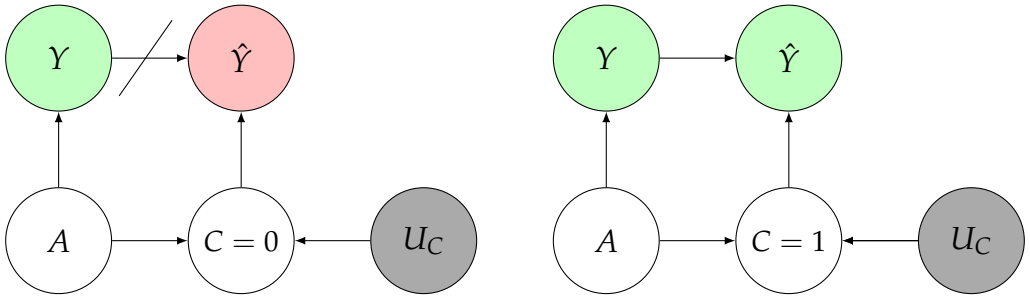


Figura 24: Efecto de la variable de corrección.

Proposición 18. Sea g un clasificador arbitrario y U_C una variable de ruido para la variable de corrección C. Redefinimos la ecuación estructural de $\hat{Y}$ como:

$$f_C = A \oplus U_C$$
$$\hat{Y} = \begin{cases} g(y), & \text{si } C = 1, \\ g(a), & \text{si } C = 0. \end{cases}$$

La Proposición 18 propone la utilización de una corrección del sesgo, eligiendo el grado de desviación de la dependencia de los datos de Y. De esta forma, la probabilidad con la que U_C toma el valor 0 o 1, puede ser interpretada como un hiperparámetro ajustable con el objetivo de reflejar cómo de injustos son los datos, y cuánta desviación respecto de los datos históricos se desea alcanzar en el clasificador desarrollado.

Dada esta nueva aproximación, las nociones de paridad demográfica y el criterio de probabilidades igualadas, podrían ser satisfechas conjuntamente ajustando el hiperparámetro U_C. Obteniendo las siguientes ecuaciones de equidad modificadas:

$$\hat{Y} \perp A \mid (C = 0)$$
$$\hat{Y} \perp A \mid Y, C.$$

10

MEDIDAS CAUSALES

En este capítulo realizaremos un estudio del modelo contrafactual, y cómo sirve de base para diferentes medidas causales, en particular para la equidad contrafactual que discutiremos más detalladamente.

10.1 CONTRAFACTUALES

Una vez definidos los modelos causales estructurales, podemos formular preguntas más delicadas que el mero efecto de una acción. En concreto, preguntas contrafactuales como: ¿Habría evitado el atasco si hubiese tomado otra ruta diferente? o ¿Me habrían concedido el préstamo si mi raza o edad fuesen distintas? Podemos dar respuesta a estas preguntas, a partir de un modelo causal estructural. Sin embargo, el procedimiento para extraer la respuesta del modelo, necesita del cálculo de *contrafactuales*.

10.1.1 *Ejemplo: Modelo de decisión contrafactual*

Supongamos un problema de decisión entre dos modelos de caja negra para resolver un problema. Denotaremos por X a la variable indicadora de cada algoritmo. Si el problema es irresoluble ($U = 1$), ninguno de los algoritmos podrá encontrar una solución. Si el problema es resoluble ($U = 0$), un algoritmo obtendrá mejores resultados que el otro. El rendimiento es independientemente de cualquiera de los dos modelos con una probabilidad de $\frac{1}{2}$. Definiremos dos variables aleatorias U_0, U_1, que nos informarán del rendimiento para los algoritmos $X = 0$ y $X = 1$, respectivamente. El modelo seleccionado entre los dos existentes, será elegido al azar por una variable U_X con probabilidad $\frac{1}{2}$. Supondremos también una variable $Y \in \{0, 1\}$, que nos dirá si el algoritmo ha encontrado una solución óptima ($Y = 0$) o no ($Y = 1$). A continuación especificaremos el modelo discutido con el siguiente programa:

Algoritmo 4: Programa distribución contrafactual.

Muestras de variables aleatorias independientes de Bernoulli:
$U, U_0, U_1, U_X \sim Bernoulli\left(\frac{1}{2}\right)$;
$X := U_X$;
$Y := X \cdot \text{máx}\{U, U_1\} + (1 - X) \cdot \text{máx}\{U, U_0\}$;

El grafo asociado al modelo anterior viene dado por:

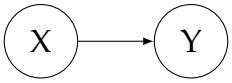


Figura 25: Grafo causal asociado al ejemplo.

Supongamos que elegimos el algoritmo $X = 1$ y observamos que no consigue una solución $Y = 1$. A continuación, nos hacemos la siguiente pregunta: ¿Habría sido mejor elegir el otro algoritmo? Para responder a ello, calcularemos la probabilidad $P_{M[X:=0]}(Y = 0)$. Dada la sustitución $X := 0$ en nuestro modelo, para que el algoritmo encontrase una solución óptima necesitamos que $\text{máx}\{U, U_0\} = 0$. Esto sólo ocurre cuando $U = 0$ (con probabilidad $\frac{1}{2}$) y $U_0 = 0$ (también con probabilidad $\frac{1}{2}$). Concluimos que $P_{M[X:=0]}(Y = 0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Aunque pudiera parecerlo, esta no es la respuesta correcta a nuestra pregunta. La razón es, que hicimos los cálculos sin considerar la decisión previa $\{X = 1, Y = 1\}$. A partir de esta observación, ciertas configuraciones de las variables de ruido (U, U_0, U_1) ya no son factibles. En concreto, si U y U_1 hubieran sido ambos cero, habríamos encontrado una solución correcta con el algoritmo $X = 1$, pero esto contradice a nuestra observación $Y = 1$. De hecho, la configuración $\{X = 1, Y = 1\}$, sólo permite los valores de la Tabla 4 para U y U_1.

U	U_1
0	1
1	0
1	1

Tabla 4: Posibles valores de las variables de ruido dada la evidencia observada.

Cada uno de estos tres casos, es igualmente probable, lo que en particular significa que el evento $U = 1$ tiene una probabilidad de $\frac{2}{3}$. Sin considerar la observación, recordemos que $U = 1$ tenía una probabilidad de $\frac{1}{2}$. Por lo que la evidencia observada $\{X = 1, Y = 1\}$, ha sesgado la distribución de la variable de ruido U hacia el valor 1. Utilizaremos la letra U' para referirnos a esta versión sesgada de U.

Podemos volver a considerar el efecto de la acción $X := 0$ sobre el resultado Y, trabajando ahora con la nueva variable U'. Para $Y = 0$, necesitamos que $\text{máx}\{U', U_0\} = 0$.

Esto significa que $U' = 0$, un suceso que ahora tiene probabilidad $\frac{1}{3}$, y $U_0 = 0$ (con probabilidad $\frac{1}{2}$, igual que antes). Por lo tanto, obtenemos que una vez actualizado el modelo, $P_{M'[X:=0]}(Y = 0) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$.

Si comparamos el nuevo valor con el resultado anterior, tenemos que la incorporación de las observaciones disponibles en nuestro cálculo, disminuyó la probabilidad de que encontrásemos la solución óptima con el otro algoritmo,

$$P_{M'[X:=0]}(Y = 0) = \frac{1}{6} < \frac{1}{4} = P_{M[X:=0]}(Y = 0).$$

La razón de ello es que el evento observado, sesga la distribución de las variables de ruido haciendo que fuese incluso más probable que en general, el problema no tuviese solución. Llamaremos al resultado que acabamos de calcular: el valor contrafactual al elegir el algoritmo alternativo, dado que el algoritmo seleccionado no encontró una solución óptima.

10.1.2 *Formalización del cálculo contrafactual*

Sea $M = (U, V, F)$ un modelo causal estructural, la especificación de F es un supuesto fuerte que permite el cálculo de valores contrafactuales. Calcularemos el valor de Y, si Z hubiera tomado el valor z, como la solución de Y para un $U := u$ dado, en el que las ecuaciones de Z se sustituyen por $Z := z$. Denotaremos la expresión anterior por $Y_{Z \leftarrow z}(U)$ (Pearl [2000]).

La *inferencia contrafactual*, especificada por un modelo causal M dada la evidencia E, equivale al cálculo de $P(Y_{Z \leftarrow z}(U) \mid E = e)$, donde $E, Z, Y \subseteq V$. Existen tres pasos esenciales en el cálculo de contrafactuales: incorporar las observaciones sesgando las variables de ruido mediante una operación de condicionamiento, realizar una *operación do* en el modelo causal después de sustituir las variables de ruido sesgadas y finalmente, calcular la distribución para una variable objetivo. Estos tres pasos se denominan *abducción, acción* y *predicción*, y se definen de la siguiente manera:

Definición 71 (Cálculo del contrafactual)**.** Dado un modelo causal estructural M, un evento observado E, una intervención $Z := z$ y una variable objetivo Y, definimos el *contrafactual* $Y_{Z \leftarrow z}(E)$ mediante los siguientes pasos:

- *Abducción*: Condicionar la distribución conjunta de $U = (U_1, \ldots, U_n)$ al suceso E. Esto da lugar a una distribución sesgada $U' = P(U \mid E = e)$.
- *Acción*: Utilizar el *operador do*, para realizar la intervención $Z := z$ en el modelo causal estructural M, obteniendo el modelo $M' = M[Z := z]$.
- *Predicción*: Calcular el objetivo contrafactual $Y_{Z \leftarrow z}(E)$ usando U' como semilla aleatoria en M'.

10.2 EQUIDAD CONTRAFACTUAL

Todos los criterios de equidad definidos en el Capítulo 6, tienen limitaciones en su aplicación a problemas reales. Mientras que la equidad por desconocimiento es un criterio insuficiente debido a la gran cantidad de características correlacionadas con los atributos sensibles, la equidad individual tiene el problema de ser directamente dependiende de una distancia fiable entre individuos. Por otro lado, los criterios de equidad de grupo, son observacionales y no pueden utilizarse para encontrar la causa de la disparidad entre grupos.

Como solución a estos problemas, aparecen las *medidas causales* donde, en este trabajo, profundizaremos en la *equidad contrafactual* (Russell et al. [2017]) que puede ser considerada como una subclase de las mismas. Este concepto considera que, para un individuo, una decisión es justa si coincide en el mundo real y en un mundo "contrafactual" en el que el individuo perteneciese a un grupo demográfico diferente. Esta suposición construye un método para comprobar el tratamiento dispar, que surge al sustituir únicamente el atributo sensible y además, proporciona una explicación del impacto del sesgo a través de un grafo causal.

Definición 72 (Equidad contrafactual)**.** Dado $A \in \mathcal{A}$ un atributo sensible multivaluado, $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario, $X \in \mathcal{X} \setminus \mathcal{A}$ una variable observada cualquiera y (U, V, F) un modelo causal donde $V \equiv A \cup X$. Se dice que g satisface la *equidad contrafactual* si, y solo si,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a),$$

para todo y y $a' \neq a$.

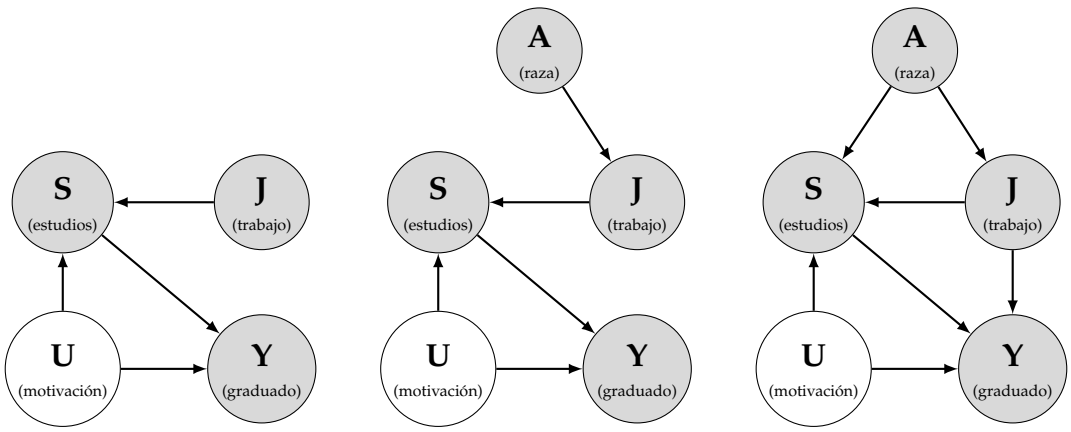


Figura 26: Ejemplo de grafo causal en un problema real.

La Figura 26 muestra varios ejemplos sobre un caso relacionado con la admisión a la universidad (Russell et al. [2017]). Si sustituimos el atributo sensible (raza) por su valor contrafactual, todas las características correlacionadas con él también se verían influidas (estudios y trabajo), propagándose hacia abajo en el grafo causal a través de las ecuaciones estructurales. Cualquier atributo que no descienda del atributo sensible permanecerá igual.

10.2.1 *Implicaciones de la definición de equidad*

En este apartado, presentaremos algunas implicaciones de la definición de equidad contrafactual, y algunos resultados que proporcionan un método directo para satisfacer esta noción de justicia dado un modelo (Kusner et al. [2018]).

Lema 4. Sea G el grafo causal del modelo dado por (U, V, F), $A \in \mathcal{A}$ un atributo sensible multivaluado y $g\colon \mathcal{X} \to \mathcal{Y}$ un clasificador arbitrario. Entonces g satisface la equidad contrafactual, si es una función que no depende de los nodos descendientes de A.

Demostración. Sea W una variable no descendiente de A en G. Entonces $W_{A\leftarrow a}(U)$ y $W_{A\leftarrow a'}(U)$ tienen la misma distribución para los tres pasos de la Definición 71. Por lo tanto, la distribución de cualquier función g de los nodos no descendientes de A, es invariante con respecto a los valores contrafactuales de A. □

Defectos y limitaciones

El concepto de equidad contrafactual, teóricamente hablando, puede parecer una buena idea para eliminar todos los defectos del resto de criterios estudiados. En la práctica, la dependencia de conceptos como la inferencia causal o el estudio de contrafactuales lo hacen una de las nociones de equidad más complejas de implementar.

Otro problema surge cuando queremos acordar cómo debería ser el grafo causal, o decidir qué características vamos a utilizar incluso disponiendo de dicho grafo. Esto se debe a que podríamos perder precisión si rechazamos variables de estudio que pudiesen ser problemáticas a la hora de realizar el estudio causal.

Parte IV

ANÁLISIS EXPERIMENTAL

Elaboración de un ejemplo práctico sobre la equidad contrafactual y su discusión frente a otras nociones de equidad estudiadas.

11

DESCRIPCIÓN Y DISEÑO

En este capítulo se describirá un problema real de justicia en el ámbito de la educación, y se propondrán varios diseños de modelos causales que puedan ser tratados por un algoritmo de equidad contrafactual, con el objetivo de eliminar el tratamiento dispar en el problema.

11.1 ALGORITMO DE APRENDIZAJE JUSTO

El algoritmo propuesto por Kusner et al. [2018] parte de la necesidad de relacionar $\hat{Y}$ con Y. Para ello restringiremos $\hat{Y}$ para que funcione como una función parametrizada de los nodos no descendientes de A apoyándonos en el Lema 4.

Calculemos la predicción $\hat{Y}$ a partir de un clasificador parametrizado por $\boldsymbol{\theta}$ al que denominaremos como $g_{\boldsymbol{\theta}}(U, X_{\not\succ A})$, donde $X_{\not\succ A} \subseteq X$ denota el conjunto de no descendientes de A. Dada una función de pérdida $l(\cdot,\cdot)$ (*squared* o *logistic loss*) y un conjunto de datos $\mathcal{D} = \{(A^{(i)}, X^{(i)}, Y^{(i)}) \ : \ i = 1, \ldots, n\}$, definimos

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[l(y^{(i')}, g_{\boldsymbol{\theta}}(U^{(i')}, x^{(i')}_{\not\succ A})) \mid x^{(i)}, a^{(i)}],$$

como la pérdida empírica al minimizar sobre $\boldsymbol{\theta}$. Cada esperanza se calcula respecto a la variable $U^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$, donde $P_{\mathcal{M}}(U \mid x, a)$ es la distribución condicional de las variables de ruido dada por el modelo causal $\mathcal{M}$. En tiempo de predicción, crearemos una nueva variable $\tilde{Y} = \mathbb{E}[\hat{Y}(U^*, x^*_{\not\succ A}) \mid x^*, a^*]$ para un nuevo elemento del conjunto de datos (a^*, x^*).

Para aproximar la esperanza, utilizaremos el *método de Monte Carlo basado en cadenas de Markov* o MCMC (Andrieu et al. [2003]). Los métodos MCMC son una clase de algoritmos de simulación para el muestreo y estimación de distribuciones de probabilidad a posteriori. Al construir un cadena de Markov usando una distribución deseada como distribución de equilibrio de la cadena, se puede hacer un muestreo de la distribución, registrando los diferentes estados del grafo. Cuántas más iteraciones realicemos con el método de Monte Carlo, la distribución de la cadena se acercará más a la distribución real deseada.

Algoritmo 5: FairLearning($\mathcal{D}, \mathcal{M}$)

Entrada: $\mathcal{D}$, conjunto de datos y $\mathcal{M}$, modelo causal.
Salida: parámetros aprendidos $\hat{\theta}$.
para $i \in \mathcal{D}$ **hacer**
  muestrear m ejemplos MCMC $U_1^{(i)}, \ldots, U_m^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.
fin
Creamos $\mathcal{D}'$ donde cada punto $(a^{(i)}, x^{(i)}, y^{(i)})$ en $\mathcal{D}$ es sustituido por los correspondientes m puntos $\{(a^{(i)}, x^{(i)}, y^{(i)}, u_j^{(i)})\}$;
$\hat{\theta} \leftarrow \text{argmin}_{\theta} \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_{\theta}(U^{(i')}, x_{\not\succ A}^{(i')}))$;
devolver $g_{\hat{\theta}}$;

11.2 DISEÑO DEL MODELO CAUSAL DE ENTRADA

El modelo $\mathcal{M}$ debe proporcionarse al Algoritmo 5; recordemos que los modelos causales requieren de fuertes suposiciones cuando se hacen afirmaciones contrafactuales (Barocas et al. [2019]). Existen infinitas ecuaciones estructurales compatibles con la misma distribución observable; por lo que, teóricamente estos modelos deberían ser propensos a modificaciones si, por ejemplo se incorporan nuevos datos anteriormente no observados que contradijesen el modelo actual.

En nuestro trabajo, no será necesario especificar un modelo totalmente determinista y relajaremos las ecuaciones estructurales a partir de la definición de distribución condicional. En particular, el concepto de equidad contrafactual es válido bajo los siguientes tres niveles, que vienen especificados en el trabajo de Kusner et al. [2018] como:

- **Nivel 1**: predecir $\hat{Y}$ usando sólo las variables observables no descendientes de A. Esta consideración no requiere de ninguna suposición causal, pero en la mayoría de los problemas la mayor parte de las variables observables serán descendientes de atributos protegidos, y por tanto tendremos menos información manejable.
- **Nivel 2**: suponer variables de ruido que actúan como causas no deterministas de las variables observables, basadas en el conocimiento explícito del dominio y en algoritmos de aprendizaje. La información sobre X pasará a $\hat{Y}$ a través de $P(U \mid x, a)$.
- **Nivel 3:** construir un modelo totalmente determinista con variables de ruido. Por ejemplo, la distribución $P(V_i \mid pa_i)$ puede tratarse como un modelo de error aditivo, $V_i = f_i(pa_i) + \epsilon_i$ (Peters et al. [2014]). El término de error ϵ_i sirve como una entrada a $\hat{Y}$ calculada a partir de las variables observadas. Esto maximizará la información extraída por el clasificador.

11.3 APLICACIÓN EN UN PROBLEMA REAL

Ilustraremos la aplicación de justicia contrafactual sobre un problema del mundo real que requiere equidad. El objetivo de este experimento es cuantificar el comportamiento del Algoritmo 5 con tamaños de muestra finitos sobre una suposición real compatible con un modelo sintético.

11.3.1 *Descripción del problema*

El Consejo de Admisión de las Facultades de Derecho realizó una encuesta en 163 facultades de Derecho de Estados Unidos (Wightman [1998]). Contiene información sobre 21.790 estudiantes de Derecho, tales como las puntuaciones de su examen de acceso (LSAT), su media del expediente (GPA) antes de entrar en la facultad, y su nota media del primer año (FYA) en la carrera de Derecho.

A partir de estos datos, una escuela podría querer predecir si un solicitante tendrá un FYA alto. También podría ser interesante asegurarse de que estas predicciones no están sesgadas por la raza y el sexo del individuo. Sin embargo, es bastante probable que los resultados del LSAT, GPA y FYA estén sesgados por factores sociales. Nuestro trabajo consistirá en aplicar las herramientas aprendidas para equidad contrafactual en diversos escenarios y comparar su actuación con otras nociones de equidad estudiadas.

11.3.2 *Escenarios de predicción*

Utilizaremos el mismo escenario de experimentación que el propuesto en Kusner et al. [2018]. Dividiremos el conjunto de datos en un 80-20 (entrenamiento-test) para evaluar los modelos, preservando el equilibrio de las etiquetas. Para predecir los resultados utilizaremos un predictor basado en regresión lineal, y mediremos la exactitud alcanzada por cada modelo a con la métrica RMSE y el coeficiente de determinación.

Según hemos descrito en la Sección 11.2, existen tres aproximaciones a partir de las cuales podemos construir un predictor de FYA que satisfaga la equidad contrafactual:

- **Nivel 1**: usaremos cualquier característica que no sea descendiente de la raza y el sexo para la predicción. Como suponemos que el LSAT, el GPA y el FYA están sesgados por la raza y el sexo, no podremos utilizar ninguna de las características observadas para construir un clasificador justo contrafactual.
- **Nivel 2** (Fair K): supondremos que una variable de ruido: los conocimientos del estudiante (K), afecta a las puntuaciones de GPA, LSAT y FYA. El gráfico

causal correspondiente a este modelo se muestra en la Figura 27. Emplearemos las siguientes distribuciones:

$$\begin{aligned} \text{GPA} &\sim \mathcal{N}(b_G + K\mathbf{w}_G^K + [A_R, A_S]\mathbf{w}_G^A, \sigma_G), \\ \text{LSAT} &\sim \textit{Poisson}(\exp(b_L + K\mathbf{w}_L^K + [A_R, A_S]\mathbf{w}_L^A)), \\ \text{FYA} &\sim \mathcal{N}(K\mathbf{w}_F^K + [A_R, A_S]\mathbf{w}_F^A, 1), \\ \text{K} &\sim \mathcal{N}(0, 1). \end{aligned}$$

Realizamos la inferencia sobre este modelo utilizando un conjunto de entrenamiento observado para estimar la distribución posterior de K.

- **Nivel 3** (Fair *Add*): modelaremos las puntuaciones de GPA, LSAT y FYA como variables continuas con términos de error aditivos independientes de la raza y el sexo. Estimamos los términos de error ϵ_G, ϵ_L ajustando primero dos modelos que utilizan la raza y el sexo para predecir individualmente el GPA y LSAT. A continuación, calculamos los residuos de cada modelo (aplicando por ejemplo, $\epsilon_G = \text{GPA} - \hat{Y}_{\text{GPA}}(A_R, A_S)$). Finalmente, utilizaremos las estimaciones residuales de ϵ_G, ϵ_L para predecir FYA. Este modelo se muestra en la Figura 27. En este caso las distribuciones vienen dadas por:

$$\begin{aligned} \text{GPA} &= b_G + [A_R, A_S]\mathbf{w}_G^A + \epsilon_G, \quad \epsilon_G \sim P(\epsilon_G) \\ \text{LSAT} &= b_L + [A_R, A_S]\mathbf{w}_L^A + \epsilon_L, \quad \epsilon_L \sim P(\epsilon_L), \\ \text{FYA} &= b_F + [A_R, A_S]\mathbf{w}_F^A + \epsilon_F, \quad \epsilon_F \sim P(\epsilon_F). \end{aligned}$$

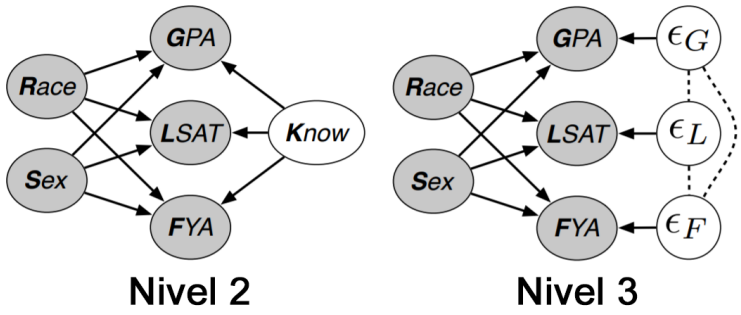


Figura 27: Grafos causales de los escenarios para los niveles 2 y 3, respectivamente.

Compararemos el escenario propuesto por la equidad contrafactual con dos líneas de base injustas:

- **Modelo completo**: utilizará todas las características disponibles para el individuo, incluidos los atributos sensibles.
- **Modelo por desconocimiento**: aplicaremos la noción de equidad por desconocimiento discutida en la Sección 6.2.

12

IMPLEMENTACIÓN Y RESULTADOS

En este capítulo explicaremos los procedimientos empleados en la implementación de los diseños y algoritmos definidos en el capítulo anterior. Además realizaremos un estudio de la equidad del problema haciendo uso del software Aequitas.

12.1 OBTENCIÓN Y TRATAMIENTO DE LOS DATOS

El conjunto de datos ha sido extraído del repositorio Ethik, que contiene diferentes *datasets* relacionados con el estudio de equidad en aprendizaje automático, y que está disponible en este enlace[1].

Es recomendable que antes de aplicar cualquier modelo o algoritmo sobre un conjunto de datos, se realice un tratamiento de los mismos (`preprocess_data`). En nuestro caso, haciendo uso de la biblioteca Pandas, hemos realizado las siguientes modificaciones sobre el conjunto de datos original:

- Categorizar cada valor del atributo raza, obteniendo una columna para que cada tipo de raza tome el valor 1 o 0 para cada individuo, señalando la pertenencia o no del mismo a la raza indicada por la columna específica (`get_dummies`).
- Sustituir el atributo sexo por dos columnas ('Male'-'Female'), que indiquen con 1 o 0 la característica del individuo concreto.
- Discretizar el valor de 'LSAT' (convertir cada valor a tipo entero).
- A partir del análisis del Apartado A.2.2: agrupar los individuos de raza mexicana y puertorriqueña en la raza hispánica, y los de raza amerindia en otras.

UGPA	LSAT	ZFYA	Female	Male	Asian	Black	Hispanic	Other	White	region_first	sander_index	first_pf
3.1	42	0.36	1	0	0	0	0	0	1	NE	0.82023	1
3.3	39	-0.99	0	1	0	0	0	0	1	NE	0.80178	0
2.9	44	-0.77	0	1	1	0	0	0	0	GL	0.61369	1

Tabla 5: Conjunto *law_data* tras el preprocesamiento de los datos.

1 https://github.com/XAI-ANITI/ethik/tree/master/ethik/data

Definiremos un array con las categorías de los atributos protegidos, conteniendo los siguientes valores: [”Female”, ”Male”, ”Asian”, ”Black”, ”Hispanic”, ”Other”, ”White”]. Estos serán los valores considerados como sensibles en la definición del problema.

Una vez realizado el tratamiento de los datos, y teniendo el conjunto de datos como se presenta en la Tabla 5, almacenaremos los datos útiles en un diccionario (definido a partir de la función `create_pystan_dic`) que nos servirá como estructura de datos contenedora del grafo causal del experimento. Esta elección la realizamos, ya que el método StanModel, que construye el modelo causal, trabaja con diccionarios en Python.

Almacenaremos en el diccionario los siguientes valores:

- **N** - Número total de ejemplos para ese diccionario.
- **C** - Número de categorías diferentes de atributos sensibles.
- **A** - Array con el contenido de cada una de las filas de los atributos protegidos.
- **GPA** - Array con los valores de UGPA.
- **LSAT** - Array con los valores de LSAT.
- **FYA** - Array con los valores de ZFYA.

Realizaremos, haciendo uso del método `train_test_split`, una partición del 80-20 para los conjuntos de entrenamiento y test. Crearemos un diccionario para los datos de entrenamiento y otro para los de test (el cuál no contendrá los valores de FYA).

12.2 IMPLEMENTACIÓN DE LOS MODELOS

El lenguaje de programación seleccionado para la implementación del proyecto ha sido Python en su versión 3.8.5. La elección de Python se debe a que es un lenguaje muy popular en el ámbito de la ciencia de datos, con una gran cantidad de bibliotecas útiles para la visualización de datos y altamente compatible con otros lenguajes de programación que nos pueden ser útiles, tanto para el tratamiento de datos, como para la construcción de modelos causales.

En este apartado, se presentará el diseño de la implementación realizada en Python de forma esquematizada, para más información del funcionamiento de cada modelo, consultar el Capítulo 11. Si por el contrario, quiere consultarse el manual de ejecución del experimento, ir a la Sección 12.5.

12.2.1 *Modelos injustos*

Como no podemos implementar el nivel 1 presentado en el Apartado 11.3.2, ya que el LSAT, GPA y FYA están sesgados por la raza y el sexo. En su lugar (y con el fin de comparar el rendimiento con los modelos de nivel 2 y 3) construiremos los siguientes dos modelos de base injusta.

Para construir los conjuntos con los que entrenaremos en cada modelo, haremos uso de la función `hstack` de la biblioteca NumPy. Añadiremos según el modelo, los valores específicos de las columnas para los diccionarios pertinentes. Nuestro objetivo será predecir la nota media del primer año para cada individuo (FYA). Utilizaremos un regresor lineal (`LinearRegression` de la biblioteca Scikit-learn) para predecir el valor de FYA para cada individuo sobre el conjunto de datos de prueba. El regresor, habrá sido previamente entrenado con `linear_regresor.fit(x_train, y_labels)` usando los datos del conjunto de entrenamiento. Para predecir los valores haremos uso del método `linear_regresor.predict(test)`.

Modelo completo

Funciones creadas para la implementación del *modelo completo*:

- `mod_full(dic_train, dic_test)` :
  Utilizaremos los datos de entrenamiento (`dic_train`) relativos a los atributos protegidos (`dic_train['A']`) y los valores de GPA (`dic_train['GPA']`) y LSAT (`dic_train['LSAT']`). Construiremos el modelo completo, y usaremos el regresor lineal para entrenarlo. Una vez hecho esto, podremos calcular las predicciones sobre el conjunto de datos de prueba (`dic_test`).

Modelo por desconocimento

Funciones creadas para la construcción del *modelo por desconocimiento*:

- `mod_unaware(dic_train, dic_test)` :
  Usaremos los valores de LSAT (`dic_train['LSAT']`) y GPA (`dic_train['GPA']`) del conjunto de entrenamiento (`dic_train`) para entrenar el modelo. Finalmente, comprobaremos su actuación sobre el conjunto de datos de prueba (`dic_test`). En este caso, no utilizamos los atributos protegidos en el entrenamiento.

12.2.2 *Modelos justos*

En este apartado, definiremos las funciones utilizadas para implementar los escenarios de predicción para los niveles 2 y 3 del Apartado 11.3.2.

Modelo de variable latente (Fair K)

Para la creación del *modelo de variable latente*, definiremos las siguientes funciones:

- `MCMC(dic_post, path_model, path_stan)` :
  Obtenemos las muestras para un modelo Stan (definido en `path_stan`) con los datos contenidos en `dic_post`. Construiremos el modelo usando la función `StanModel` de la biblioteca PyStan, lo entrenaremos haciendo uso del método

`sampling` y finalmente extraeremos las muestras con `extract`. Usaremos 2000 iteraciones y 1 cadena de Markov para replicar el ejemplo de Kusner et al. [2018]. Para evitar ejecutar el modelo cada vez, guardaremos los modelos entrenados (en `path_model`) usando las funciones `load` y `dump` del módulo `pickle`.

- `get_mean_params(samples, dic_test)` :
  Creamos un diccionario con la media de los parámetros obtenidos en el entrenamiento previo del modelo, almacenados en `samples`. Mantendremos la estructura del diccionario `dic_test` para los parámetros que no varíen, como son $[N, C, A, GPA, LSAT]$, y haremos la media para el resto de parámetros que no dependan de FYA, a saber $[\mathbf{w}_G^K, \mathbf{w}_G^A, \mathbf{w}_L^K, \mathbf{w}_L^A, \sigma_G, b_G, b_L]$.

- `fair_learning(dic_train, dic_test)` :
  Utilizamos el método `MCMC` sobre el modelo `law_school_train.stan` para obtener las muestras para cada punto del conjunto de datos `dic_train`. Guardamos la media de la variable K para `dic_train`. Usamos la distribución de K aprendida y hacemos las medias del resto de variables para `dic_test`. Volvemos a inferir sobre el modelo, esta vez usando `law_school_only_k.stan` (sin información sobre FYA) para `dic_test`. Finalmente, calculamos la media de la variable K para el conjunto de prueba y la devolveremos junto con la calculada para el conjunto de entrenamiento.

- `mod_fair_k(dic_train)` :
  Usaremos los arrays con las medias de la variable K inferida en el método anterior, para el conjunto de entrenamiento (`train_k`) y el conjunto de prueba (`test_k`). Entrenaremos el modelo haciendo uso de la información contenida en `train_k` y `dic_train['FYA']`. Finalmente usaremos el regresor lineal para calcular las predicciones sobre el conjunto de prueba.

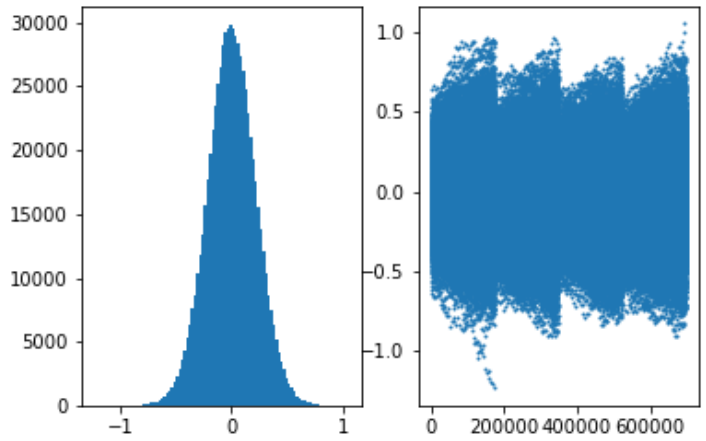


Figura 28: Distribución de la traza para una muestra de `train_k`.

Modelo de error aditivo (Fair Add)

Crearemos las siguientes funciones para construir el *modelo de error aditivo*:

- `calculate_eps(dic_train,dic_test,var)` :
  Estima el error ϵ, entrenando el modelo utilizando el conjunto total de datos (`dic_train+dic_test`) para una variable observada `var` usando además los atributos protegidos. Calculamos los residuos de cada modelo como

  $$\epsilon_{\texttt{var}} = \texttt{var} - \hat{Y}_{\texttt{var}}(\mathrm{A}),$$

  donde `var` puede tomar el valor de LSAT o GPA.

- `mod_fair_add(dic_train,dic_test)` :
  Estimamos el error para GPA (ϵ_{GPA}) y LSAT (ϵ_{LSAT}) para el conjunto de entrenamiento (`dic_train`) utilizando el método previo. Usamos los valores de ϵ_{GPA} y ϵ_{LSAT} para entrenar el modelo. Finalmente, usaremos el regresor lineal para calcular las predicciones sobre el conjunto de datos de prueba (`dic_test`).

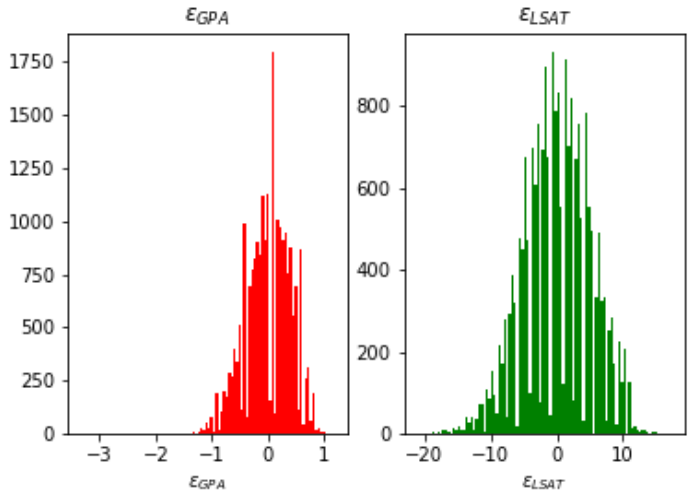


Figura 29: Distribución de los residuos calculados para cada variable (GPA y LSAT).

12.2.3 *Exactitud de los modelos*

Para estimar el rendimiento y evaluar el ajuste del modelo, usaremos el error cuadrático medio y el coeficiente de determinación de la predicción (ver su definición en el Apartado 5.4).

En la práctica calcularemos el RMSE haciendo la raíz cuadrada (`sqrt` de la biblioteca NumPy) a la función `mean_squared_error` sobre los conjuntos de las predicciones de FYA y sus valores reales. Por otro lado, calcularemos el valor de R^2, haciendo uso del método `score` del regresor lineal, sobre los valores y etiquetas reales del conjunto de prueba.

12.2.4 *Contrafactuales*

Para calcular las distribuciones contrafactuales definiremos las siguientes funciones:

- `cambiar_individuos(dic, protected_attr, referencia, objetivo)` :
  Realiza una copia (con el método `deepcopy` del módulo copy) del diccionario `dic`, cambiando el atributo `objetivo` perteneciente a los atributos protegidos (`protected_attr`) por el valor del `referencia` para todos los individuos que lo cumplan.

- `preds_individuos(orig_data, protected_attr, atributo, preds)` :
  Devuelve la restricción del array de predicciones (`preds`) para el grupo de individuos del conjunto de datos `orig_data` que presenten el valor del `atributo` perteneciente a los atributos protegidos (`protected_attr`).

- `preds_array_individuos(orig_data, protected_attr, atributos, preds)` :
  Mismo funcionamiento que el anterior, pero para un array de `atributos`.

12.3 CONTRASTE DE LOS RESULTADOS

12.3.1 *Actuación de los modelos*

En un modelo de regresión, el objetivo es predecir el valor numérico de una cantidad desconocida a partir de unas características dadas. Llamaremos error de predicción a la diferencia entre la predicción y el valor real. Usaremos como medida de evaluación el error cuadrático medio (RMSE), que indica cómo de cerca están los valores observados de los valores predichos por el modelo. Si comparamos el RMSE entre dos modelos, un valor más bajo indicará un mejor ajuste de los datos.

También usaremos el coeficiente de determinación (R^2), que refleja la bondad del ajuste de un modelo a la variable que pretende explicar. El resultado del coeficiente de determinación oscila entre 0 y 1. Cuánto más cerca de 1 se sitúe su valor, mejor será el ajuste del modelo. Por el contrario, cuánto más cerca de cero, menos ajustado estará el modelo y, por tanto, será menos fiable.

	Completo	**Desconocimiento**	**Fair** K	**Fair Add**
RMSE	0.8720	0.8916	0.9312	0.9191
R^2	0.1440	0.1051	0.0239	0.0490

Tabla 6: Valores de RMSE y R^2 para los cuatro modelos.

Comparamos el RMSE alcanzado por la regresión lineal sobre el conjunto de prueba para cada modelo en la Tabla 6. El modelo completo obtiene el RMSE más bajo, ya que utiliza mayor número de atributos para reconstruir con mayor precisión el valor de FYA. El modelo por desconocimiento usa las variables injustas GPA y LSAT, pero como no utiliza las variables raza y sexo, un menor número de atributos, hace que no pueda igualar el RMSE del modelo completo. Los modelos Fair *K* y Fair *Add*, al satisfacer la equidad contrafactual, compensan algo de exactitud. El modelo Fair *K* utiliza supuestos más débiles, y por tanto, el RMSE es más elevado. Por otro lado, usando los supuestos del modelo Fair *Add*, producimos un modelo justo contrafactualmente, que sacrifica suposiciones ligeramente más fuertes por un RMSE más bajo.

Respecto al valor de R^2, en general, los modelos no obtienen un valor muy bueno, teniendo en cuenta que este valor se encuentra en el intervalo [0,1], siendo 1 el ajuste perfecto. Esto se debe a que probablemente un regresor lineal no sea el mejor modelo para predecir las etiquetas de nuestro problema, no obstante lo hemos usado para simplificar el problema, ya que nuestro trabajo se centra en el estudio de la equidad. En un futuro próximo, podríamos probar otros modelos de regresión no lineales para obtener unos mejores resultados en la predicción de las etiquetas.

12.3.2 *Estudio de la equidad*

A continuación, utilizaremos los escenarios propuestos para realizar un estudio de la equidad sobre los diferentes aspectos estudiados en la teoría.

Realizaremos un cálculo de los contrafactuales basándonos en el escenario del nivel 2, y visualizaremos gráficamente el cambio en la distribución de los individuos cuando modificamos sus atributos. Por otro lado, presentaremos un estudio en Aequitas sobre las nociones de equidad, que cumplen los datos bajo las predicciones de los cuatro modelos estudiados.

12.3.2.1 *Distribuciones contrafactuales*

Comprobaremos empíricamente si los modelos propuestos son justos desde el punto de vista contrafactual. Para ello, supondremos que el modelo real del mundo viene dado por la Figura 30, que corresponde con el modelo de variable latente (Fair *K*). Podemos ajustar los parámetros de este modelo utilizando los datos observados, y evaluar la equidad contrafactual mediante un muestreo del mismo.

Generaremos muestras para los atributos de raza y sexo a partir de las variables observadas y las variables contrafactuales. Ajustaremos los modelos tanto a la muestra original como a la contrafactual, y pintaremos una gráfica (usando la función `displot`) mostrando los cambios en la distribución de la FYA predicha para ambos modelos de referencia.

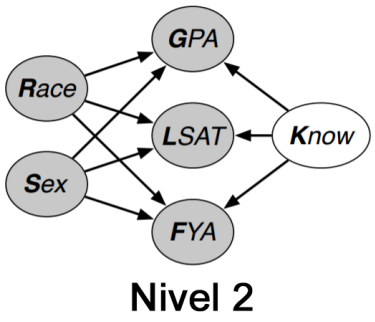


Figura 30: Grafo causal del escenario propuesto para el nivel 2.

En la Tabla 7, visualizamos un ejemplo del cálculo del contrafactual sobre un individuo seleccionado aleatoriamente en la población. En nuestro caso, hemos escogido un individuo de la población con los atributos de *sexo*=mujer y *raza*=asiática, con unos valores determinados de GPA, LSAT y FYA (para simplificar, tomarán valores binarios (*fail*/*ok*)). Para que el modelo cumpla la equidad contrafactual, la decisión predicha, debe ser igual a la predicción sobre el mismo individuo con un cambio en uno de sus atributos protegidos. Normalmente el cambio se realiza de un valor objetivo por su valor de referencia.

1. Seleccionar un individuo:				
A_S	A_R	GPA	LSAT	FYA
female	asian	FAIL	OK	FAIL

2. Cambiar el valor del atributo protegido:				
a′	A_R	GPA	LSAT	FYA
male	asian	FAIL	OK	FAIL

3. Calcular la variable K no observada del modelo causal:					
a′	A_R	GPA	LSAT	FYA	K
male	asian	FAIL	OK	FAIL	OK

4. Recalcular las variables observadas del modelo causal:					
a′	A_R	$\text{GPA}_{A_S \leftarrow a'}$	$\text{LSAT}_{A_S \leftarrow a'}$	$\text{FYA}_{A_S \leftarrow a'}$	K
male	asian	OK	OK	OK	OK

Tabla 7: Proceso de cálculo de contrafactual sobre un individuo para la variable K.

En el ejemplo de la Tabla 7, realizamos el cálculo del contrafactual para el atributo sexo, por lo que cambiamos el valor *female* por *male* (históricamente considerado más aventajado). Una vez cambiado el valor del atributo para el individuo, calcularemos el valor de la variable no observada K, que en nuestro caso, se calcula siguiendo el modelo Fair K. Finalmente, eliminamos los valores restantes y predecimos estos valores (en este problema, solo nos interesa el valor de FYA) a partir del nuevo valor de K. Como podemos observar; el valor de FYA ha cambiado, por lo que para este individuo, el modelo utilizado en la predicción, no sería justo contrafactualmente para el atributo protegido sexo.

Replicaremos el proceso de la Tabla 7 para todos los individuos del *dataset*. Los cambios en los atributos se harán usando el método `cambiar_individuos` implementado. La Figura 31 muestra el cambio de las distribuciones contrafactuales usando el modelo completo, donde cada columna corresponde al cambio contrafactual para el atributo específico. La Figura 32 muestra lo mismo, pero para el modelo por desconocimiento. En cada gráfica, la distribución azul muestra la densidad de los datos originales y la distribución roja la de los datos contrafactuales. Si un modelo es contrafactualmente justo, estas distribuciones serán lo más parecidas posibles. Cabe destacar que la gráfica en cada columna representa la distribución condicionada a los dos grupos de individuos indicados, para ello usamos la función `preds_array_individuos`.

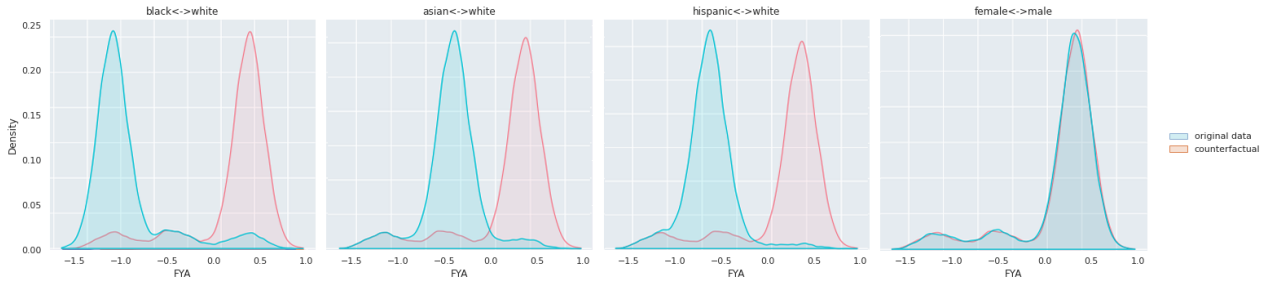


Figura 31: Distribución contrafactual para el modelo completo.

En la Figura 31, vemos que el modelo es injusto para los tres valores de los atributos más comunes de raza (black, asian, hispanic), mientras que es justo contrafactualmente para el atributo sexo.

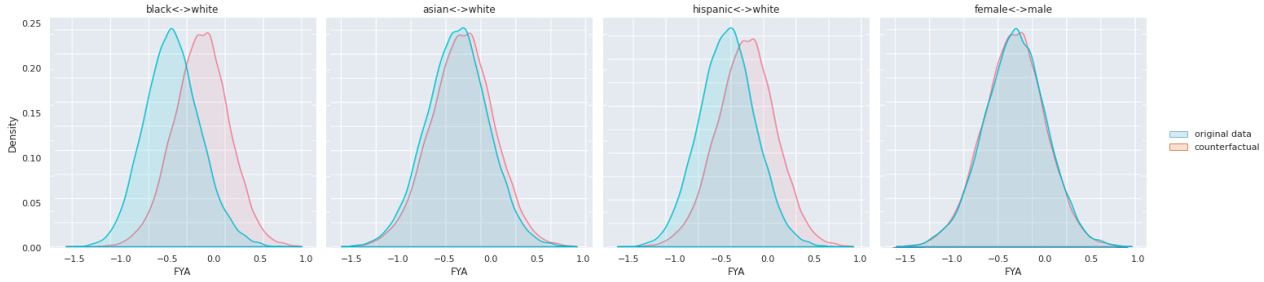


Figura 32: Distribución contrafactual para el modelo por desconocimiento.

La Figura 32 obtiene resultados a primera vista más justos para los atributos de raza. En nuestro caso, la raza asiática se podría considerar justa contrafactualmente para este modelo. Al igual que para el modelo completo, el atributo sexo no parece estar sesgado en este problema. Haciendo la media de los pesos para todos los individuos masculinos y femeninos obtenemos un valor de (0.91, 1.03) para GPA y (0.98,0.99) para LSAT. Concluimos que estos modelos son justos con respecto al sexo, ya que existe una relación causal muy débil entre el atributo sexo y los atributos GPA y LSAT.

12.3.2.2 *Análisis extendido en Aequitas*

Definiremos las siguientes funciones, que nos ayudarán a convertir el array de predicciones calculadas por cada modelo, a una estructura de datos tratable por Aequitas:

- `get_aequitas_data(train_orig, preds_train, test_orig, preds_test)`:
  Construye un *dataset* utilizando los valores originales de *raza, sexo* y *first_pf* (contenidos en `train_orig` y `test_orig`) y guarda como *score* las nuevas predicciones `preds_train` y `preds_test` para el modelo en cuestión. Finalmente, concatena ambos subconjuntos y define la estructura compatible con Aequitas definida en el Apéndice A.

- `save_graficas(data, path, aq_palette, attr)`:
  Genera gráficas de distribución del atributo `attr` para un conjunto de datos (`data`). Utiliza el color determinado en `aq_palette` y guarda la imagen en la ruta especificada por `path`.

- `tabla_metrica_grupo(data)`:
  Utiliza la clase *Group()* de Aequitas, para construir la tabla de las métricas de grupo para el conjunto de datos (`data`).

- `tabla_metrica_sesgo(data, attr_ref)`:
  Utiliza la clase *Bias()* de Aequitas, para construir la tabla de las métricas de sesgo para el conjunto de datos (`data`). El parámetro `attr_ref` especificará los valores del grupo de referencia.

- `tabla_medidas_equidad(data, attr_ref, tau=0.8)`:
  Utiliza la clase *Fairness()* de Aequitas, para construir la tabla de las medidas de equidad para el conjunto de datos (`data`). El parámetro `attr_ref` indicará los valores del grupo de referencia y el valor de `tau`, por defecto 0.8, especificará el umbral tomado para la regla p %.

En este apartado, haremos una selección de las métricas de grupo que nos serán útiles para comparar las diferentes nociones de equidad estudiadas, a saber:

- **Paridad estadística** - Aequitas utiliza la tasa de positivos predichos (PPR).
- **Probabilidades igualadas** - Cumplir simultáneamente las métricas asociadas a las tasas de falsos y verdaderos positivos (FPR y TPR, respectivamente).
- **Tasa de paridad predictiva** - Cumplir simultáneamente los valores positivo y negativo predictivo (PPV y NPV, respectivamente). En Aequitas el PPV se registra por su otro nombre: precision.

- **Paridad tipo II** - Al encontrarnos con un problema de tipo asistencial, podría ser relevante el estudio de esta medida de equidad. Esta medida se cumple cuando se aceptan como justas simultáneamente las métricas relativas a la tasa de falsa omisión (FOR) y falsos negativos (FNR).

El objetivo de este apartado, es comprobar si la actuación de los modelos justos, favorecen el cumplimiento de otras métricas de grupo, las cuales hemos estudiado en teoría. El análisis se reducirá a hacer un estudio de las tablas de métricas de grupo, sesgo y finalmente estudiar para qué casos se cumple una medida de equidad u otra. Un análisis sobre los conjuntos de datos originales para *law_data* y *COMPAS*, profundizando en otras métricas y las herramientas gráficas que ofrece Aequitas, podrá encontrarse en el Apéndice A.

Presentaremos para cada modelo: la distribución de sus puntuaciones predichas, su tabla de métricas de grupo y sesgo, y las medidas de equidad que cumplen. Finalmente, haremos un contraste general de los resultados para los cuatro modelos.

Modelo completo

Comparando las distribuciones de las predicciones (*score*) con las etiquetas reales (*label_value*), vemos que en la Figura 34, la actuación del modelo completo, a primera vista, tiene unos resultados injustos en términos de equidad respecto al atributo raza. Mientras que las puntuaciones de los individuos de raza blanca, son muy similares a los de la etiqueta real, las puntuaciones favorables hacia individuos de otra raza son mínimas o inexistentes.

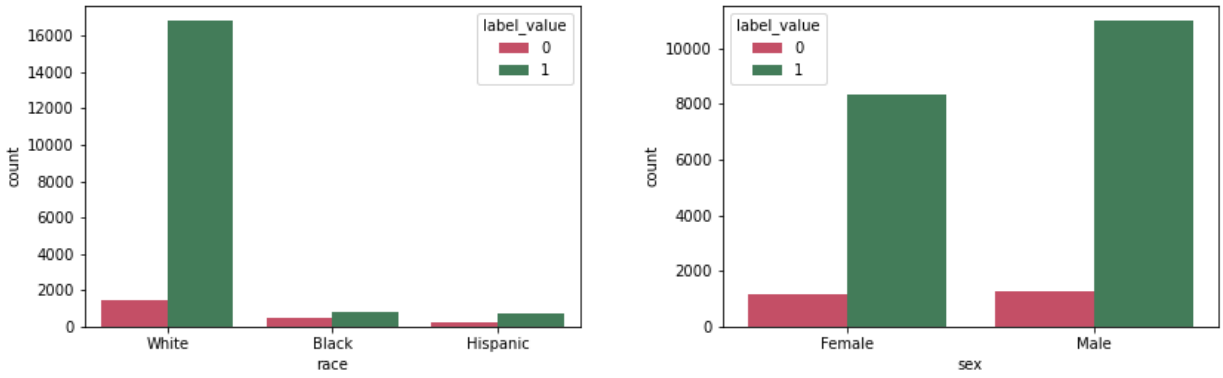


Figura 33: Distribución de las etiquetas reales para los atributos protegidos.

El modelo completo era el que mejores resultados de RMSE y R^2 ofrecía, esto se debe a que al entrenar con todos los atributos, favorece la predicción correcta para los individuos de la clase mayoritaria (raza blanca), mientras que los individuos de grupos desfavorecidos obtienen resultados nefastos en la regresión realizada.

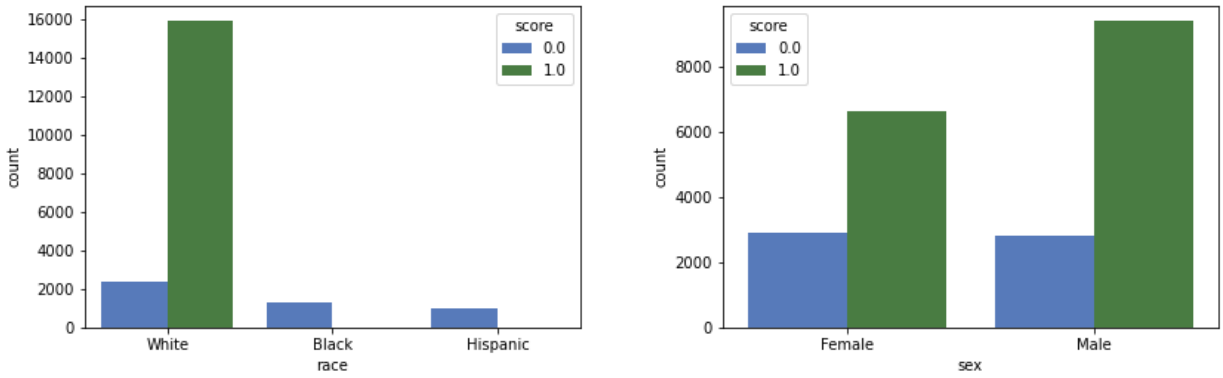


Figura 34: Distribución de la puntuación para el modelo completo.

Definiremos las tablas para las métricas de grupo (Tabla 8) y sesgo (Tabla 9) calculadas con las predicciones de este modelo. Además, también aportaremos una tabla, indicando el cumplimiento o no de las nociones de equidad (Tabla 10) sobre las que nos hemos centrado.

Nombre del Atributo	Valor	PPR	FPR	TPR	PPV	NPV	FOR	FNR
race	Asian	0.0	0.01	0.11	0.97	0.20	0.80	0.89
race	Black	0.0	0.0	0.0	-	0.38	0.62	1.0
race	Hispanic	0.0	0.0	0.03	0.95	0.26	0.74	0.97
race	Other	0.0	0.01	0.24	0.99	0.25	0.75	0.76
race	White	0.99	0.66	0.89	0.94	0.21	0.79	0.11
sex	Female	0.41	0.34	0.75	0.94	0.27	0.73	0.25
sex	Male	0.59	0.45	0.81	0.94	0.24	0.76	0.19

Tabla 8: Métricas de grupo para el modelo completo.

Para las métricas de sesgo, trataremos como atributos de referencia la raza blanca y el sexo masculino. Además, utilizaremos el criterio de la regla del 80 %, por lo que los valores de la tabla de métricas de sesgo que no estén definidas en el intervalo [0.8,1.25] serán consideradas como injustas y coloreadas en rojo, mientras que los valores contenidos en el intervalo vendrán indicados en verde. Para el resto de modelos, mantendremos el mismo valor de umbral (80 %), y mantendremos la acepción del coloreado para la tabla de métricas de sesgo.

Nombre del Atributo	Valor	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	Asian	0.0	0.02	0.12	1.03	0.95	1.01	8.09
race	Black	0.0	0.0	0.0	-	1.81	0.78	9.09
race	Hispanic	0.0	0.0	0.03	1.01	1.24	0.94	8.82
race	Other	0.0	0.02	0.27	1.05	1.19	0.95	6.91
race	White	1.0	1.0	1.0	1.0	1.0	1.0	1.0
sex	Female	0.69	0.76	0.93	1.0	1.13	0.96	1.32
sex	Male	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 9: Métricas de sesgo para el modelo completo.

Nombre del Atributo	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
sex	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE

Tabla 10: Medidas de equidad para el modelo completo.

Modelo por desconocimiento

Para este modelo, la predicción de la puntuación para los individuos de las clases minoritarias sufre cierta mejoría respecto al modelo completo analizado anteriormente.

En la Figura 36 podemos observar cómo ya existen algunos individuos de otras razas que obtienen una puntuación favorable. Esto se debe a que no tenemos en cuenta los atributos protegidos en la predicción. No obstante, probablemente sigan existiendo sesgos entre los diferentes grupos demográficos existentes. El modelo por desconocimiento también favorece la clasificación negativa (*score*=0) de individuos por sexo, en comparación con el modelo anterior.

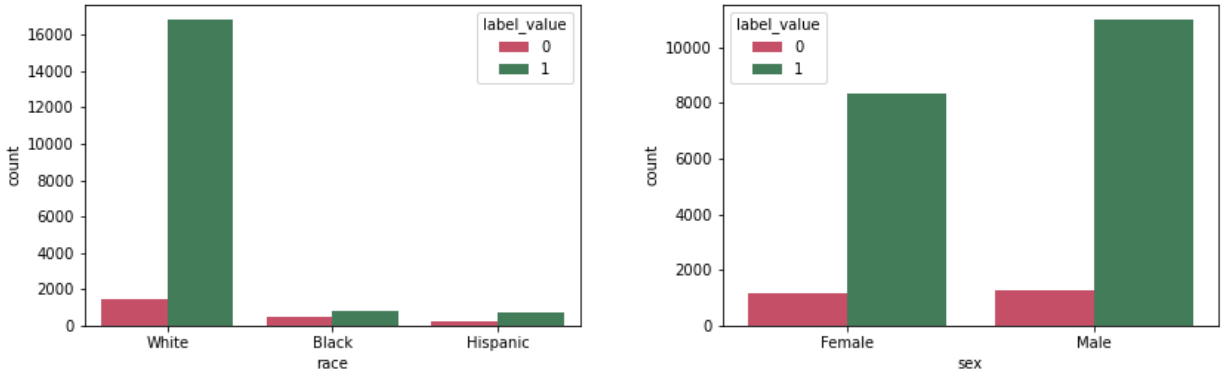


Figura 35: Distribución de las etiquetas reales para los atributos protegidos.

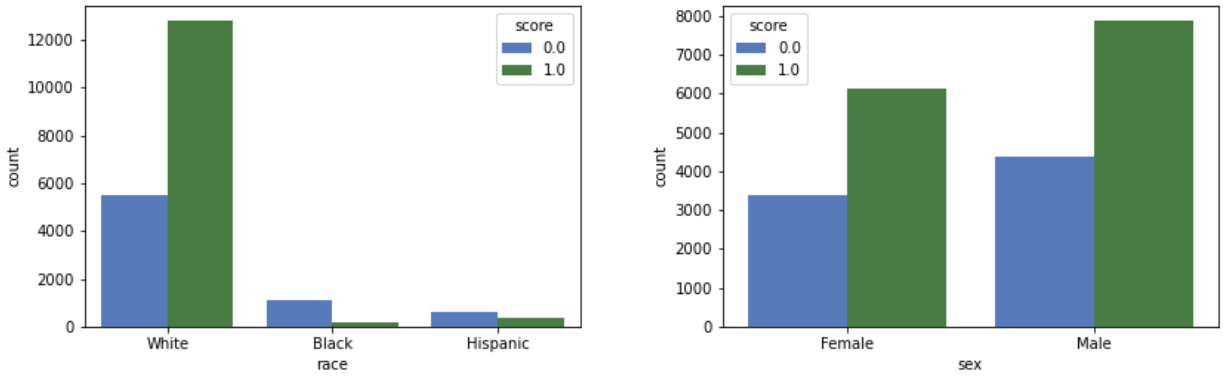


Figura 36: Distribución de la puntuación para el modelo por desconocimiento.

A continuación definimos las tablas para las métricas de grupo (Tabla 11) y sesgo (Tabla 12) calculadas con las predicciones de este modelo. También presentamos la tabla con las medidas de equidad (Tabla 13) para este modelo concreto.

Nombre del Atributo	Valor	PPR	FPR	TPR	PPV	NPV	FOR	FNR
race	Asian	0.03	0.35	0.63	0.89	0.29	0.71	0.37
race	Black	0.01	0.04	0.19	0.89	0.42	0.58	0.81
race	Hispanic	0.03	0.16	0.45	0.89	0.34	0.66	0.55
race	Other	0.01	0.23	0.58	0.91	0.32	0.68	0.42
race	White	0.91	0.43	0.72	0.95	0.15	0.85	0.28
sex	Female	0.44	0.28	0.70	0.95	0.25	0.75	0.30
sex	Male	0.56	0.34	0.68	0.95	0.19	0.81	0.32

Tabla 11: Métricas de grupo para el modelo por desconocimiento.

Nombre del Atributo	Valor	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	Asian	0.03	0.81	0.88	0.94	1.93	0.84	1.32
race	Black	0.01	0.09	0.26	0.94	2.8	0.68	2.89
race	Hispanic	0.03	0.37	0.63	0.94	2.27	0.78	1.96
race	Other	0.01	0.53	0.81	0.96	2.13	0.8	1.5
race	White	1.0	1.0	1.0	1.0	1.0	1.0	1.0
sex	Female	0.79	0.82	1.03	1.0	1.32	0.93	0.94
sex	Male	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 12: Métricas de sesgo para el modelo por desconocimiento.

Nombre del Atributo	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
sex	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE

Tabla 13: Medidas de equidad para el modelo por desconocimiento.

Modelo de variable latente (Fair K)

Para este modelo podemos observar en la Figura 38 que, a diferencia de sus predecesores injustos, obtenemos mejores resultados para los grupos desfavorecidos. Analizando la gráfica, notamos cierta mejoría en la clasificación positiva (*score*=1) para los grupos minoritarios. Además, obtenemos un mayor balance para las clases positivas y negativas en general (destacando el grupo de raza blanca y la gráfica para el atributo sexo).

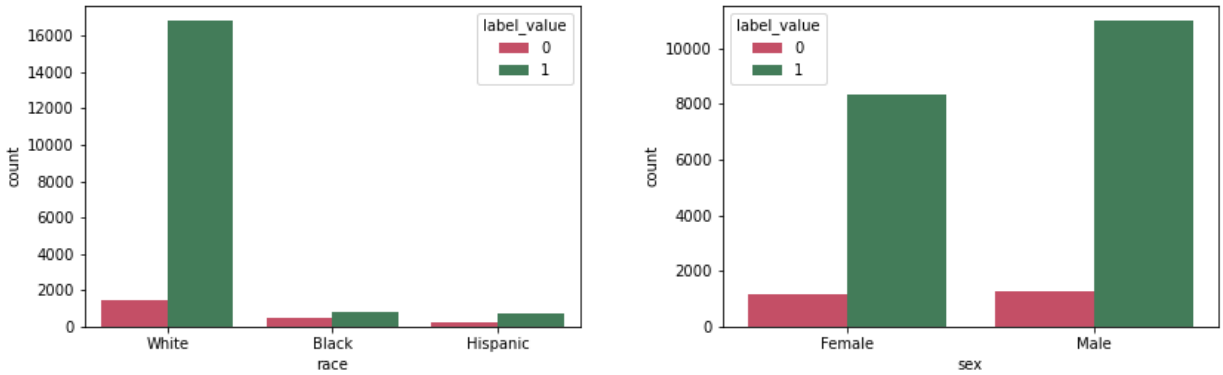


Figura 37: Distribución de las etiquetas reales para los atributos protegidos.

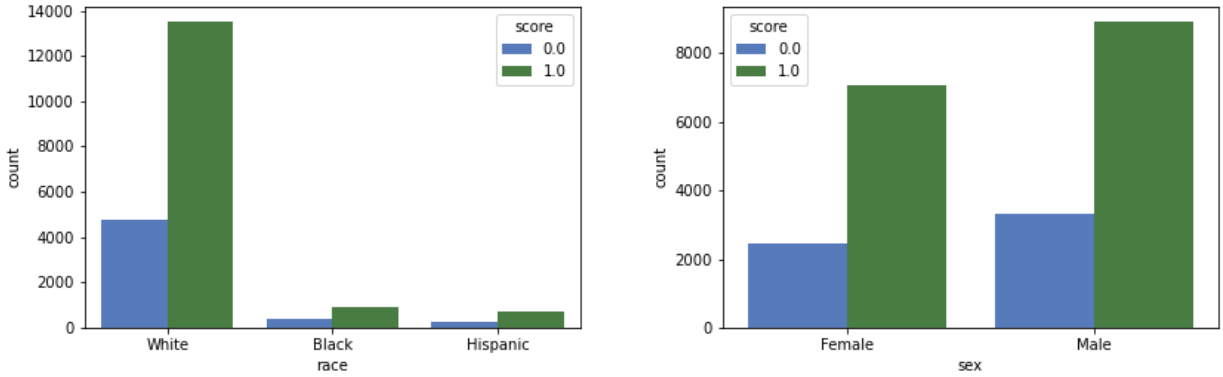


Figura 38: Distribución de la puntuación para el modelo Fair *K*.

Existe un aumento de los individuos etiquetados negativamente que no viene reflejado en la Figura 37, por lo que podríamos contar con individuos que aun cumpliendo las condiciones de buena nota (*label_value*=1) obtuviese una predicción negativa (*score*=0). Esta diferencia, entre predicciones y valor real, es la que influye en los malos resultados de este modelo para los valores de RMSE y R^2; sin embargo, esto se debe a que el modelo busca un equilibrio entre equidad-exactitud, por lo que sacrificamos la correcta clasificación de una parte de los individuos de la clase mayoritaria para evitar el sesgo hacia las clases minoritarias.

A continuación, definiremos las tablas para las métricas de grupo (Tabla 14) y sesgo (Tabla 15), y para las medidas de equidad (Tabla 16) para el modelo Fair *K*.

Nombre del Atributo	Valor	PPR	FPR	TPR	PPV	NPV	FOR	FNR
race	Asian	0.04	0.54	0.74	0.86	0.29	0.71	0.26
race	Black	0.06	0.62	0.74	0.66	0.48	0.52	0.26
race	Hispanic	0.04	0.64	0.76	0.78	0.33	0.67	0.24
race	Other	0.02	0.46	0.78	0.87	0.39	0.61	0.22
race	White	0.84	0.56	0.76	0.94	0.14	0.86	0.24
sex	Female	0.44	0.57	0.77	0.91	0.21	0.79	0.23
sex	Male	0.56	0.58	0.75	0.92	0.16	0.84	0.25

Tabla 14: Métricas de grupo para el modelo Fair *K*.

Nombre del Atributo	Valor	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	Asian	0.05	0.96	0.97	0.91	2.07	0.83	1.08
race	Black	0.07	1.11	0.97	0.70	3.43	0.60	1.08
race	Hispanic	0.05	1.14	1.0	0.83	2.36	0.78	1.0
race	Other	0.02	0.82	1.03	0.93	2.79	0.71	0.92
race	White	1.0	1.0	1.0	1.0	1.0	1.0	1.0
sex	Female	0.79	0.98	1.03	0.99	1.31	0.94	0.92
sex	Male	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 15: Métricas de sesgo para el modelo Fair *K*.

Nombre del Atributo	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE
sex	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE

Tabla 16: Medidas de equidad para el modelo Fair *K*.

Modelo de error aditivo (Fair Add)

Para este modelo obtenemos unos resultados muy similares a los del modelo Fair *K*. A simple vista, la única diferencia notable entre estos dos modelos la podemos observar comparando los gráficos de la Figura 40 y los de la Figura 38. Comentaremos si existen diferencias más notables relativas al cumplimiento de los criterios de equidad más adelante.

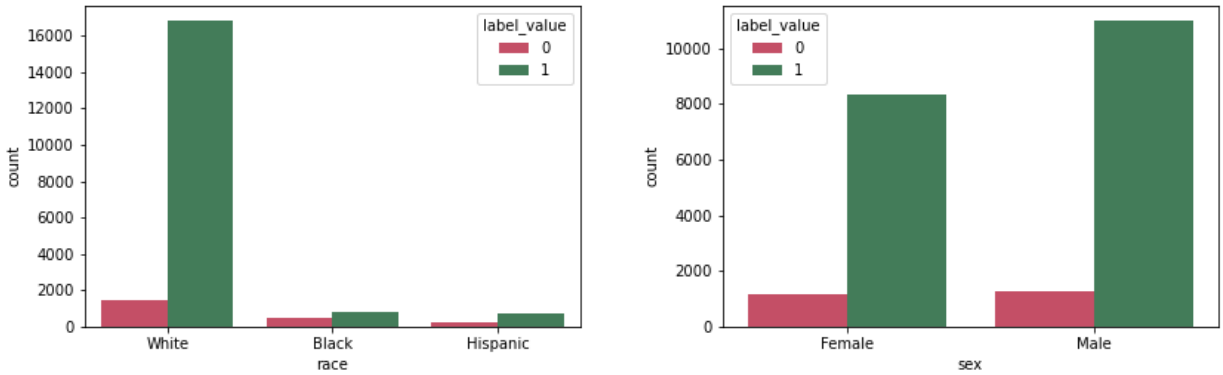


Figura 39: Distribución de las etiquetas reales para los atributos protegidos.

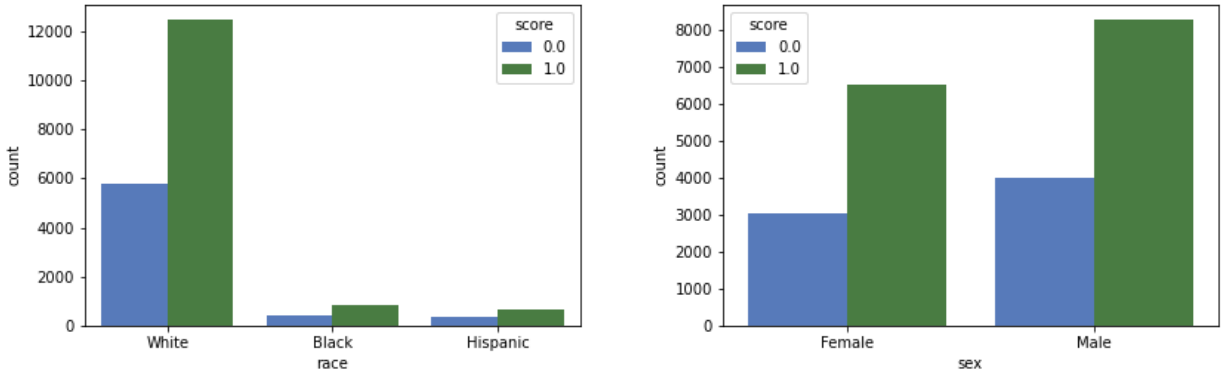


Figura 40: Distribución de la puntuación para el modelo Fair *Add*.

Seguidamente, presentaremos las tablas para las métricas de grupo (Tabla 17) y sesgo (Tabla 18) para el modelo Fair *Add*. Como hemos hecho para el resto de modelos, también incluiremos el análisis para las medidas de equidad (Tabla 19).

Nombre del Atributo	Valor	PPR	FPR	TPR	PPV	NPV	FOR	FNR
race	Asian	0.04	0.38	0.68	0.89	0.30	0.70	0.32
race	Black	0.06	0.50	0.77	0.71	0.58	0.42	0.23
race	Hispanic	0.04	0.44	0.73	0.83	0.42	0.58	0.27
race	Other	0.02	0.33	0.74	0.90	0.40	0.60	0.26
race	White	0.84	0.40	0.71	0.95	0.15	0.85	0.29
sex	Female	0.44	0.40	0.72	0.93	0.23	0.77	0.28
sex	Male	0.56	0.45	0.70	0.93	0.17	0.83	0.30

Tabla 17: Métricas de grupo para el modelo Fair *Add*.

Nombre del Atributo	Valor	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	Asian	0.05	0.95	0.96	0.94	2.0	0.82	1.1
race	Black	0.07	1.25	1.08	0.75	3.87	0.49	0.79
race	Hispanic	0.05	1.1	1.03	0.87	2.8	0.68	0.93
race	Other	0.02	0.83	1.04	0.95	2.67	0.71	0.9
race	White	1.0	1.0	1.0	1.0	1.0	1.0	1.0
sex	Female	0.79	0.89	1.03	1.0	1.35	0.93	0.93
sex	Male	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 18: Métricas de sesgo para el modelo Fair *Add*.

Nombre del Atributo	PPR Disparity	FPR Disparity	TPR Disparity	PPV Disparity	NPV Disparity	FOR Disparity	FNR Disparity
race	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
sex	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE

Tabla 19: Medidas de equidad para el modelo Fair *Add*.

Comparativa de las nociones de equidad

Para concluir esta sección, presentaremos en la Tabla 20 un resumen de los datos obtenidos para los cuatro modelos estudiados. Agruparemos las métricas de equidad en cuatro nociones de justicia, las cuales tres de ellas corresponden con las estudiadas en la Sección 6.4, y la última la presentamos en el Apéndice A.

A la vista de los resultados, los modelos justos, a diferencia de los injustos, cumplen con el criterio de las probabilidades igualadas. Esto se debe, a que el cumplimiento de la equidad contrafactual impone de cierta forma un equilibrio entre las clases minoritarias y la mayoritaria.

De esta forma, obtenemos individuos de raza blanca etiquetados incorrectamente, pero evitamos el sesgo en el resto de grupos demográficos. Es lógico, por tanto, que el criterio satisfecho sea el de probabilidades igualadas.

Modelo	Nombre del Atributo	Paridad Estadística	Probabilidades Igualadas	Tasa de Paridad Predictiva	Paridad Tipo II
Completo	*race*	FALSE	FALSE	FALSE	FALSE
	sex	FALSE	FALSE	TRUE	FALSE
Desconocimiento	*race*	FALSE	FALSE	FALSE	FALSE
	sex	FALSE	TRUE	FALSE	TRUE
Fair K	*race*	FALSE	TRUE	FALSE	FALSE
	sex	FALSE	TRUE	FALSE	TRUE
Fair Add	*race*	FALSE	TRUE	FALSE	FALSE
	sex	FALSE	TRUE	FALSE	TRUE

Tabla 20: Tabla comparativa para todos los modelos de equidad.

Por otro lado, podemos observar cómo en ninguno de los casos se satisfacen simultáneamente los tres criterios de equidad de grupo. Esto corrobora el teorema de la imposibilidad (Capítulo 9), y hace evidente la idea de incompatibilidad de las tres nociones de grupo presentadas.

Podemos comprobar que la paridad de tipo II se satisface para el atributo sexo en los modelos justos, pero también en el modelo por desconocimiento, el cual también cumple para este atributo el criterio de probabilidades igualadas. Esto constata las gráficas presentadas en el Apartado 12.3.2.1; las cuales, para el atributo sexo eran justas contrafactualmente. El modelo completo sólo cumple la tasa de paridad predictiva para el atributo sexo; esto se debe a que al predecir con la información completa del problema (por ello obtiene la mejor tasa de RMSE), tiene buena actuación para el etiquetado de los individuos de la clase positiva (Precisión), y una proporción equilibrada para los de la negativa respecto al grupo de referencia (NPV Disparity). Por lo que, al cumplir las medidas de paridad para PPV y NPV simultáneamente, cumplirá la tasa de paridad predictiva.

Cabe destacar que, si tuviésemos que elegir entre un modelo justo para aplicarlo en un problema, el modelo de error aditivo (Fair *Add*) sería la mejor opción. Además de que es mucho más rápido en ejecución que el modelo Fair *K* (ya que éste necesita realizar el muestreo con MCMC), cumple las mismas nociones de justicia y además, como observamos en la Tabla 6, tiene mejores tasas de predicción sobre los datos originales. El único inconveniente, es que parte de suposiciones más fuertes, que necesitán del conocimiento del conjunto total de los datos, el cual podría no ser conocido en determinados casos.

12.4 CONDICIONES DE LA EXPERIMENTACIÓN

En esta sección se hará un repaso de los diferentes entornos en los que se ha trabajado para la elaboración de este experimento, así como las bibliotecas que se utilizan para la programación del mismo.

12.4.1 *Entorno de ejecución*

Los experimentos realizados en este capítulo se han ejecutado en un equipo con las siguientes características:

- Arquitectura x86_64.
- AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz.
- 8 núcleos con 2 hilos de procesamiento por núcleo.
- 16 GB RAM DDR4.
- Sistema Operativo: Ubuntu 20.04.2 LTS.

12.4.2 *Entorno de programación*

Spyder es un potente entorno de desarrollo integrado (IDE) de código abierto para el lenguaje Python. Dispone de funciones avanzadas de edición, control interactivo de pruebas y depuración. Gracias a su intérprete interactivo de Python y su compatibilidad con bibliotecas como NumPy, SciPy o matplotlib, se ha convertido en uno de los entornos de programación para Python más populares.

Spyder incluye soporte de herramientas interactivas para la inspección de datos e incorpora controles de calidad específicos de Python. Además puede instalarse a través del sistema de gestión de paquetes de código abierto Anaconda, considerándose un IDE multiplataforma.

12.4.3 *Bibliotecas y herramientas auxiliares*

Algunas de las bibliotecas más importantes utilizadas junto al número de sus versiones han sido:

- Pandas 1.2.4
- NumPy 1.19.2
- Scikit-learn 0.24.2
- matplotlib 3.4.3
- seaborn 0.11.2
- Aequitas 0.42.0
- pathlib2 2.3.6
- Módulos usados incluidos en Python: pickle, copy, os.

Además se ha utilizado el módulo PyStan 2.19.1.1 que funciona como una interfaz de Python para el lenguaje de programación probabilística Stan. El cual permite generar muestras de datos a partir de métodos que trabajan con cadenas de Markov (como MCMC) y crear modelos causales sobre los que operar gracias a la inferencia estadística Bayesiana.

12.5 TUTORIAL DE EJECUCIÓN DEL EXPERIMENTO

Incluimos un tutorial de ejecución de los experimentos en el formato Jupyter Notebook. Un archivo de Jupyter Notebook es un documento JSON, que sigue un esquema versionado y que contiene una lista ordenada de celdas de entrada/salida que pueden contener código, gráficos y texto, generalmente terminado con la extensión ".ipynb".

El tutorial de ejecución y el resto del código implementado, se puede encontrar en la carpeta `experimentos` del repositorio público del proyecto, al cual puede acceder desde el siguiente enlace[2].

2 https://github.com/danibolanos/TFG-Ensuring_Fairness_in_ML/blob/main/experimentos/tutorial.ipynb

Parte V

CONCLUSIONES Y VÍAS FUTURAS

Conclusiones extraídas a lo largo del desarrollo del proyecto y debate de posibles rutas de trabajo futuras para el mismo.

13

CONCLUSIÓN

La revisión previa de diferentes ejemplos como los citados en Fuster et al. [2018], Miller [2015] o Angwin et al. [2016] nos da indicios de la importancia de detectar y mitigar el sesgo dentro de una población que pudiera ser usada como conjunto de entrenamiento, o modelo de predicción de una aplicación programada para resolver problemas del mundo real.

En este documento, se ha realizado un análisis de los conceptos de equidad más comúnes en la literatura de justicia en aprendizaje automático, con el objetivo de valorar las ventajas e inconvenientes de cada uno. También se ha realizado un estudio de las herramientas software actuales enfocadas en la detección de sesgo y sus limitaciones en la práctica, tanto en falta de conceptos implementados como en sus facilidades de uso para cualquier usuario.

En principio, las nociones de equidad por desconocimiento o equidad individual, podrían parecer suficientes para evitar el sesgo de un conjunto de datos, pero como hemos visto, la correlación invisible entre los atributos protegidos y otras características, o la fuerte dependencia de una métrica de distancia definida entre individuos, los hacen dos conceptos que a pesar de ofrecer una aproximación simple al concepto de equidad, tienen muchas desventajas a la hora de aplicarlos en la práctica.

Los conceptos de equidad de grupo surgen como una definición más formal del concepto de equidad y, a diferencia de las medidas discutidas previamente, su definición a partir de probabilidades condicionadas podría confundirse con una aplicación ideal y precisa de las nociones de equidad provenientes de la literatura de las ciencias sociales, algo bastante lejos de la realidad. La incompatibilidad entre sus tres medidas principales, demostrada a partir del teorema de imposibilidad, que impide imponer restricciones fuertes de equidad sobre un conjunto de datos. La relajación de varios de los conceptos impuestos con el objetivo de facilitar la aplicación práctica de estas métricas; así como la característica observacional de las medidas, que impide la explicación de la causa del sesgo en el problema. Son algunos de los inconvenientes con los que tiene que lidiar el concepto de equidad de grupo.

Con la aparición del concepto de medidas causales, que surge como solución a los problemas consecuentes de los anteriores criterios de equidad descritos, definimos la equidad contrafactual como una subclase de las mismas. No obstante, la equidad contrafactual a pesar de tener ventajas, como dar una explicación de la causa del sesgo, también tiene algunos problemas evidentes. El concepto de equidad contrafactual implica un conocimiento previo tanto de la inferencia causal, como de varios de sus herramientas básicas de trabajo, tales como los modelos causales y los grafos. La definición de los contrafactuales también implica un conocimiento matemático amplio necesario antes de trabajar en este contexto. Además, la creación del grafo causal del problema debe ser realizada por un experto en sociología o demografía, por lo que podría ocasionar inconvenientes durante la programación de la aplicación. Sin embargo, como vemos en la práctica, este concepto en general consigue eliminar el sesgo existente en los datos, da una explicación de la causa del sesgo, y ofrece herramientas visuales para la correcta interpretación de los datos por parte de un usuario sin amplios conocimientos en programación.

Durante la fase experimental, hemos replicado un ejemplo de equidad contrafactual sobre un conjunto de datos, donde hemos visto que el concepto de equidad contrafactual, a pesar de que obtiene peores resultados en exactitud, respecto a otras nociones básicas como la equidad por desconocimiento, consigue ser mejor en términos de mitigación de sesgo. Además, podemos ver que si auditamos un conjunto, utilizando el software de Aequitas para las diferentes medidas de equidad de grupo antes y después de que el conjunto de datos sea entrenado usando un modelo causal, nos encontramos con resultados sorprendentes, en cuanto a la existencia del sesgo en determinados grupos demográficos.

Finalmente, la utilización de herramientas software para la detección de sesgo como Aequitas sobre varios conjuntos de datos, nos presenta la necesidad de utilizar algoritmos de optimización, como el replicado en el experimento, para trabajar con un conjunto de datos que pudiese integrar tratamientos injustos sobre grupos específicos de la población. Además, señala la evidente falta de herramientas de interacción y visualización de resultados que pudieran ser utilizados, tanto por científicos de datos, como por otros expertos relacionados con el ámbito de las ciencias sociales. La existencia de estas utilidades facilitarían una interfaz de trabajo más amigable, con el objetivo de mejorar la comunicación entre los expertos de las ramas sociales y tecnológicas.

Concluimos asegurando que los objetivos presentados al inicio del trabajo se han cumplido según las expectativas previstas. Sin embargo, como se podrá leer en el capítulo siguiente, han surgido muchas vías de desarrollo relacionadas con la implementación de nuevas herramientas de interacción y visualización en el ámbito del proyecto. Las cuales hecho en falta haber implementado, o al menos introducido en este trabajo.

14

TRABAJOS FUTUROS

En el campo del aprendizaje automático, los sistemas de detección de sesgo, así como el desarrollo de nuevos conceptos de equidad, está evolucionando rápidamente. Es un hecho que en el mundo actual es totalmente necesario el tratamiento de comportamientos injustos que pudiesen existir en los conjuntos de datos usados para predecir resultados y tomar decisiones sobre los mismos

A lo largo de este proyecto hemos hecho una revisión de todos los conceptos de equidad que existen en la actualidad, analizando las ventajas e inconvenientes de cada uno de ellos. Uno de los principales problemas, es la incompatibilidad entre diversas nociones de la equidad, que como ya estudiamos en el Capítulo 9, obligan a relajar las restricciones en determinados problemas, si queremos imponer varios conceptos de equidad de grupo simultáneamente.

Una de las líneas de trabajo que podríamos desarrollar en un futuro, es la iniciada en Saravanakumar [2021] y del cuál hemos hablado en la Sección 9.2. Añadir una variable de corrección al grafo causal que elimine la incompatibilidad entre la paridad demográfica y el criterio de probabilidades igualadas, puede ser interesante a la hora de aplicarlo a determinados problemas con el objetivo de que el modelo use la información del atributo protegido, mientras evita un abuso del mismo de manera discriminatoria.

A raíz del estudio realizado en el Apéndice A sobre diferentes paquetes de software para garantizar justicia en *machine learning*, hemos podido comprobar que actualmente existen diversas herramientas para satisfacer la equidad en este ámbito. Como comentamos previamente, este tema es de evolución acelerada, y hay muchas instituciones públicas y privadas e investigadores interesados en el mismo, por lo que la construcción de herramientas progresa muy rápidamente. No obstante, hemos detectado la ausencia de utilidades para facilitar el trabajo a los científicos de datos, y expertos de otras ramas sin conocimientos realmente avanzados en programación. A continuación, proponemos algunas posibles vías de desarrollo centradas en los conceptos de equidad individual y equidad contrafactual:

Equidad individual

Uno de los principales inconvenientes de esta noción es la fuerte dependencia a una métrica de distancia, definida entre los individuos de un conjunto de datos. El problema de esta métrica es que depende directamente del problema y en principio podría estar sesgada subjetivamente a la definición de la misma. Por ello proponemos una revisión desde la perspectiva matemática de diferentes clases de funciones que pudiesen generar unos pesos adecuados para cada individuo, y que no dependiesen en gran medida del problema en concreto, manteniendo un equilibrio razonable entre equidad y precisión de los resultados.

Una vía de trabajo más práctica, sería construir un software que generase matrices de distancia con los pesos de los individuos, a partir de un conjunto de datos y unas especificaciones, determinadas por el experto en cuestión.

Equidad contrafactual

La inferencia causal es un rama que no se restringe solamente a las matemáticas, si no que tiene un papel importante en otras ciencias tales como la sociología o medicina. En el ámbito de la equidad contrafactual, sería entonces de vital importancia implementar herramientas que pudiesen ser utilizadas de forma interactiva por expertos de estas otras ramas.

Actualmente existen implementaciones de paquetes como Ethik, que incorporan diversos tratamientos y estudios gráficos para la noción de equidad contrafactual. Sin embargo, no muchos trabajos se centran en el apartado de la construcción del grafo causal, ya que esta tarea es normalmente desarrollada en el ámbito de la sociología política o demografía.

Proponemos el desarrollo de una aplicación que, haciendo uso de una biblioteca como D3.js, implemente una herramienta interactiva para construir los grafos causales y establecer las relaciones entre las diferentes variables, dado un *dataset* con el problema. El software implementaría como *FrontEnd*, tanto la construcción interactiva de los grafos causales como la visualización de resultados, en un entorno amigable para el usuario. En el *BackEnd* se implementaría la construcción del modelo causal, haciendo uso, por ejemplo, del algoritmo utilizado en el experimento de la Parte IV. De esta forma, se facilitaría la construcción de un grafo causal, el cual podría ser depurado por el propio experto. Obtendríamos una mejor comunicación entre las dos partes que colaboran (ciencias sociales e ingeniería), durante el desarrollo de modelos u algoritmos causales, y evitaríamos problemas en el diseño del grafo causal que pudiesen aparecer durante la implementación algorítmica.

BIBLIOGRAFÍA

Yaser Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from Data: A Short Course*. AMLBook, 2012. ISBN 1600490069.

Adverse Impact Analysis / Four-Fifths Rule. Adverse impact analysis / four-fifths rule. https://www.prevuehr.com/resources/insights/adverse-impact-analysis-four-fifths-rule/, Center for Data Science and Public Policy, 2009.

Christophe Andrieu, Nando De Freitas, Arnau Doucet, and Michael I. Jordan. An introduction to mcmc for machine learning, 2003.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness in Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *International Journal of Epidemiology 43, no. 2*, 2014.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy, 2021.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 0387310738.

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention, 2017.

Michael Collins. Convergence proof for the perceptron algorithm. *Columbia University*, 2012. http://www.cs.columbia.edu/~mcollins/courses/6998-2012/notes/perc.converge.pdf.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

Frederick H. Cramer. *Random Variables and Probability Distributions*. Cambridge University Press, 2004. ISBN 0521604869.

Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky. *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large*. Springer, 2012. ISBN 3642304869.

Amir Dembo. *Probability Theory: STAT310/MATH230*. CreateSpace Independent Publishing Platform, 2014. ISBN 1502955652.

Nicholas Diakopoulos. Algorithmic-accountability: the investigation of black boxes. *Tow Center for Digital Journalism*, 2014.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011.

Roland Fryer, Glenn Loury, and Tolga Yuret. An economic analysis of color-blind affirmative action. *Journal of Law, Economics and Organization*, 2008.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance*, 2018.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning, 2018.

Chris Godsil and Gordon Royle. *Algebraic Graph Theory*. Springer, 2001. ISBN 0387952209.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

L.L. Humphrey, B.K.S. Chan, and H.C. Sox. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of Internal Medicine 137, no. 4*, 2002.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018.

Claire C. Miller. Can an algorithm hire better than a human? *New York Times*, 2015.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, Mar 2021. ISSN 2326-831X. doi: 10.1146/annurev-statistics-042720-125902. URL http://dx.doi.org/10.1146/annurev-statistics-042720-125902.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK, 2000. ISBN 0521773628.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models, 2014.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019.

Frank Rosenblatt. The perceptron: A perceiving and recognizing automaton (project para). *Cornell Aeronautical Laboratory*, 1957.

Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness, 2017.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit, 2019.

Kailash Karthik Saravanakumar. The impossibility theorem of machine fairness – a causal perspective, 2021.

Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. Disciplined convex-concave programming, 2016.

Masashi Sugiyama, Anton Schwaighofer, Neil D. Lawrence, and Joaquin Quiñonero-Candela. Dataset shift in machine learning. *The MIT Press*, 2017.

Title VII of the Civil Rights Act: Equal Employment Opportunities. Title vii of the civil rights act of 1964: Equal employment opportunities. https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964, United States, 2 de julio de 1964.

Sahil Verma and Julia Rubin. Fairness definitions explained. *Indian Institute of Technology Kanpur and University of British Columbia*, 2018.

Hao Wang, Berk Ustun, and Flavio P. Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions, 2019.

Linda F. Wightman. Lsac national longitudinal bar passage study, 1998.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017a. doi: 10.1145/3038912.3052660. URL http://dx.doi.org/10.1145/3038912.3052660.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2017b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification, 2017c.

Richard Zemel, Yu L. Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations, 2013.

Ciyou Zhu, Richard H. Byrd, Jorge Nocedal, and Peihuang Lu. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization, 1997.

APÉNDICES

A

HERRAMIENTAS SOFTWARE PARA GARANTIZAR JUSTICIA

En este capítulo hablaremos de algunas herramientas y paquetes de Python que existen actualmente para garantizar la equidad en aprendizaje automático. Hablaremos de algunos de los más influyentes, entre los que destacaremos Aequitas.

A.1 PAQUETES DE PYTHON PARA EQUIDAD EN AA

ETHIK

Ethik es un paquete de Python para realizar estudios sobre equidad en aprendizaje automático y obtener explicaciones sobre la misma. Ethik utiliza el enfoque de la equidad contrafactual para dar solución a escenarios hipotéticos de injusticia en una población concreta. El paquete contiene una recopilación de conjuntos de datos sobre problemas de equidad a partir de la cual hemos podido extraer el *dataset* denominado *law_data.csv*, que contiene el problema del Apartado 11.3.1, usando la función `load_law_school()`.

Podemos encontrar la documentación y herramientas que proporciona el paquete en el siguiente enlace[1].

FAIRLEARN

Fairlearn es un paquete de Python que permite a los científicos de datos y desarrolladores de aplicaciones de *machine learning* evaluar la equidad de su sistema, y mitigar cualquier problema de sesgo en una población para un conjunto de datos observado. Fairlearn contiene algoritmos de mitigación, así como métricas para la evaluación del modelo.

Podemos encontrar la documentación y una guía de uso del paquete en el siguiente enlace[2].

1 https://xai-aniti.github.io/ethik/
2 https://fairlearn.org/v0.5.0/user_guide/index.html

A.2 AEQUITAS

Aequitas (Saleiro et al. [2019]) es una herramienta de auditoría desarrollada por el *Center for Data Science and Public Policy* de la Universidad de Chicago. Es una herramienta de código abierto que consta de diversas utilidades de soporte para la auditoría de sesgos; creado para ser utilizado por analistas de todo tipo relacionados con el ámbito del aprendizaje automático, y cuyo principal objetivo es auditar los modelos de *machine learning*, con el fin de encontrar posibles discriminaciones en ellos y evitarlas en un futuro.

Aequitas nos permite detectar dos tipos de sesgos:

- Acciones sesgadas que no ocurren de forma representativa en la población.
- Resultados sesgados, a causa de errores de clasificación de nuestro sistema, con respecto a ciertos grupos de la población.

Para utilizar la herramienta se necesita aportar los siguientes datos:

- Datos sobre los atributos específicos (raza, sexo, etc.) que queramos auditar.
- El conjunto de personas de la población mencionada que el sistema de evaluación de riesgos seleccionó para una intervención.

Estructura de los datos de entrada y resultados

Podemos dividir en tres apartados (conformados por columnas en Aequitas) los datos que debemos aportar para el correcto funcionamiento de la herramienta.

- **score**: representa la conclusión a la que llega un modelo, puede ser binaria (0 o 1) o continua (decimal entre 0 y 1). Esta decisión representa si el sujeto es apto o no, por ejemplo, si se le concede un crédito bancario.
- **label_value**: representa los datos reales; es decir, si la predicción realizada por el modelo fue correcta. Por ejemplo, el sujeto fue capaz de devolver el crédito en su totalidad. Es por esto por lo que el modelo solo puede ser auditado después de su aplicación y no antes. Se representa como un valor binario, 1 significa que la predicción fue correcta, 0 que no lo fue.
- **attributes**: categorías de los atributos definidos por el usuario y utilizados para decidir la equidad del modelo. Algunos ejemplos de atributos son la raza, sexo, educación, edad o ingresos.

El funcionamiento de Aequitas viene determinado por los conceptos y métricas definidas previamente en la Sección 6.4. Podemos acceder a una lista de los conceptos y métricas que utiliza Aequitas desde el siguiente enlace[3].

La herramienta devuelve una interpretación gráfica de los resultados junto a información relevante acerca de tres tipos de métricas:

3 https://dssg.github.io/aequitas/metrics.html

- **Métricas de grupo -** Los grupos podrán definirse utilizando la clase *Group()*.
- **Métricas de disparidad -** Las disparidades podrán ser calculadas haciendo uso de la clase *Bias()*, una vez calculadas las métricas de grupo.
- **Medidas de equidad -** Estudiaremos la equidad y gráficos de distribución de las métricas de grupo y disparidad a partir de la clase *Fairness()*.

A continuación mostraremos dos ejemplos de uso de Aequitas haciendo uso de las diferentes alternativas que ofrece para auditar un *dataset*. Por un lado, usaremos la herramienta WEB para analizar la información acerca de las medidas de equidad ofrecidas, y por otro la API de Python que incluye un mayor número de métricas y gráficas para interpretar los resultados.

A.2.1 *Ejemplo: Puntuación del riesgo de reincidencia delictiva*

En este apartado, analizaremos cómo se definen las diferentes métricas con las que trabaja Aequitas a partir de los conceptos previamente estudiados. Mostraremos un ejemplo sobre el conjunto de datos *COMPAS*, cuyo objetivo era identificar a las personas con riesgo de reincidencia, para respaldar las decisiones de liberación previa al juicio.

score	label_value	race	sex	age_cat
1	0	Hispanic	Male	Less than 25
0	1	African-American	Female	25-45
0	0	Caucasian	Male	Greater than 45

Tabla 21: Ejemplo del *dataset COMPAS* aportado a Aequitas.

Los datos están basados en las estadísticas recogidas en el condado de Broward y puestas a disposición del público por la organización ProPublica (Angwin et al. [2016]). Este conjunto de datos contiene una puntuación del riesgo de reincidencia para 7.214 individuos, sus resultados reales de reincidencia a los dos años, y una serie de atributos recogidos entre 2013 y 2014. Para tratar la puntuación de riesgo con Aequitas, convertiremos el *score* original (1-10) a uno binario donde 0 indica riesgo bajo y 1 medio o alto.

Análisis de los datos

En los gráficos siguientes podemos notar una gran diferencia en la distribución de las puntuaciones por raza (Figura 41), con una mayoría de personas de raza caucásica pronosticadas como riesgo bajo (*score*=0) y una mayoría de población afroamericana pronosticada como riesgo medio o alto (*score*=1). En cuanto al atributo edad (Figura 42), se predicen como riesgo bajo la mayoría de las personas de edad superior a 25 años y como riesgo medio-alto a la mayoría de los menores de 25 años. Finalmente,

aunque existen muchos más hombres que mujeres en la base de datos (los hombres tienen un índice de criminalidad mayor que las mujeres); parece que, en principio, el sexo (Figura 43) no es determinante a la hora de puntuar la reincidencia.

Además, incluimos gráficos de cómo se distribuyen las etiquetas reales según los atributos protegidos. Las etiquetas reales contienen la información sobre si el sujeto reincidió (*label_value*=1) o no (*label_value*=0); es decir, si el individuo sufrió una nueva detención en un plazo de dos años. Con esta información podremos comprobar directamente la precisión de las predicciones.

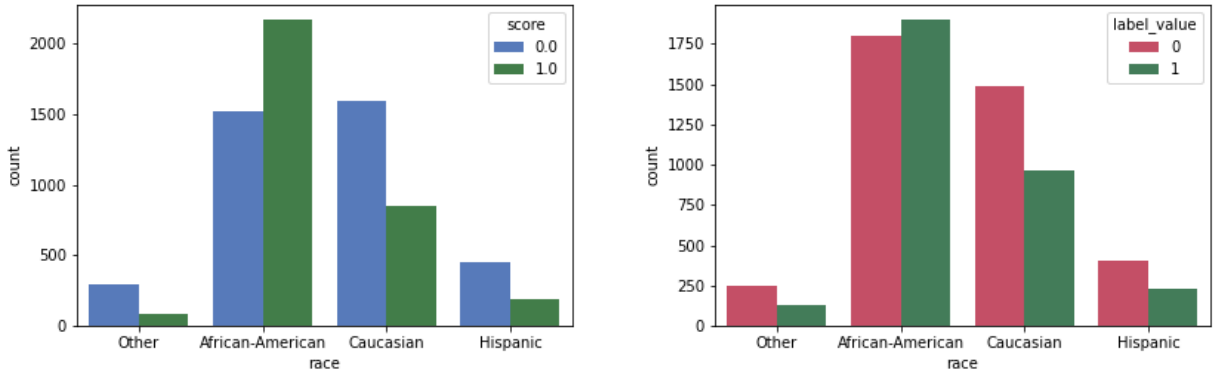


Figura 41: Gráfico de barras de la puntuación y etiqueta real para el atributo *race*.

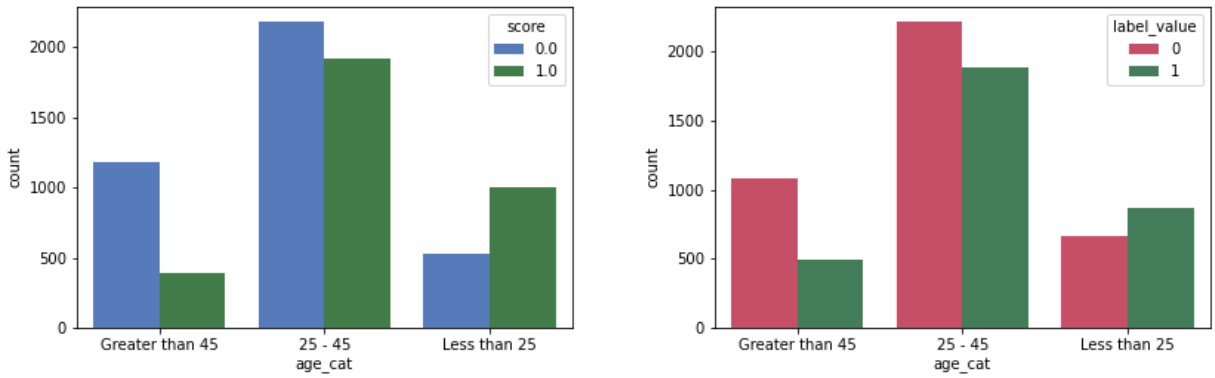


Figura 42: Gráfico de barras de la puntuación y etiqueta real para el atributo *age_cat*.

En la Figura 41 podemos observar que donde se detecta una mayor diferencia entre las gráficas de puntuación de riesgo y la reincidencia real, es en el grupo de los individuos de raza afroamericana. Este hecho induce a preguntarnos si estos patrones podrían reflejar o no un sesgo en la población, que fuese perjudicial a la hora de utilizar *COMPAS* como un conjunto de datos de entrenamiento para un modelo de aprendizaje automático.

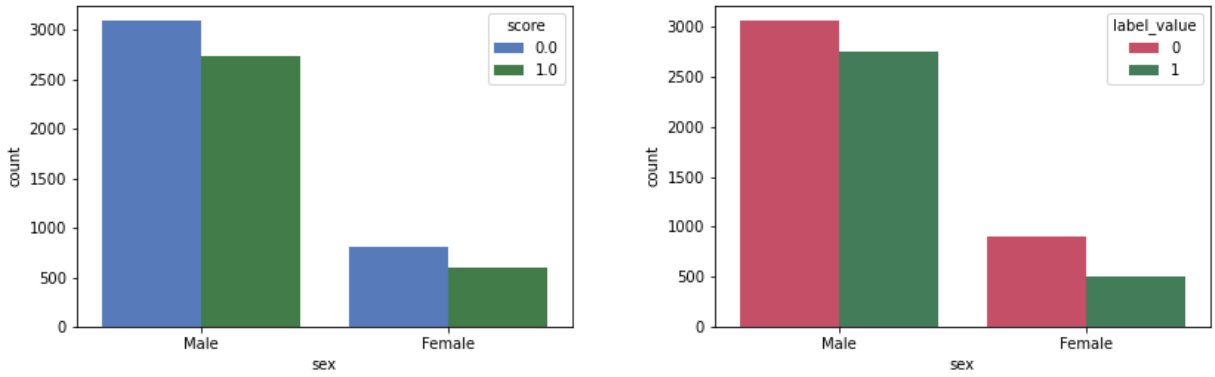


Figura 43: Gráfico de barras de la puntuación y etiqueta real para el atributo *sex*.

Utilizaremos la herramienta de Aequitas para realizar un estudio de los posibles patrones de sesgo que pudiesen existir entre los diferentes grupos de la población que representa el conjunto de datos *COMPAS*.

Métricas de grupo

Haciendo uso de la clase *Group()*, calcularemos una tabla con los valores de las métricas de grupo para un atributo concreto (p.ej. raza).

race

Attribute Value	Group Size Ratio	PPR	PPREV	FDR	FPR	FOR	FNR
African-American	0.51	0.66	0.59	0.37	0.45	0.35	0.28
Asian	0	0.0	0.25	0.25	0.09	0.12	0.33
Caucasian	0.34	0.26	0.35	0.41	0.23	0.29	0.48
Hispanic	0.09	0.06	0.3	0.46	0.21	0.29	0.56
Native American	0	0.0	0.67	0.25	0.38	0.17	0.1
Other	0.05	0.02	0.21	0.46	0.15	0.3	0.68

Tabla 22: Tabla con las principales métricas de grupo para el atributo *race*.

Métricas de disparidad

Tomaremos como atributo de referencia la raza *Caucasian*. En nuestro caso la raza *Caucasian* no es la raza mayoritaria en el conjunto de datos, pero la tomaremos como

referencia debido a que históricamente en el ámbito en el que nos movemos es la que sufre menos sesgos.

A continuación usaremos la clase *Bias()* para medir la disparidad entre cada grupo objetivo y el grupo de referencia seleccionado. Aequitas calcula la disparidad a partir de la fórmula que ya definimos en el Apartado 6.4.6.

$$\text{Métrica de disparidad}_{G(a_o)} = \frac{\text{Métrica}_{a_o}}{\text{Métrica}_{a_r}},$$

donde Métrica hace referencia a una métrica de grupo de la Sección 6.4. Es evidente que la disparidad de cualquier métrica sobre el grupo de referencia será 1.

$$\text{Métrica de disparidad}_{G(a_r)} = \frac{\text{Métrica}_{a_r}}{\text{Métrica}_{a_r}} = 1.$$

Si queremos calcular por ejemplo la disparidad del ratio de falsos negativos (FNR) sobre el grupo de raza *Asian*, se calculará de la siguiente forma:

$$\text{FNR}_{\text{G(Asian)}} = \frac{\text{FNR}_{\text{Asian}}}{\text{FNR}_{\text{Caucasian}}} = \frac{0{,}33}{0{,}48} = 0{,}7.$$

Para ver otro ejemplo de cómo realiza el cálculo Aequitas, mediremos la disparidad para la tasa de falso descubrimiento (FDR) sobre el grupo de raza *Hispanic*:

$$\text{FDR}_{\text{G(Hispanic)}} = \frac{\text{FDR}_{\text{Hispanic}}}{\text{FDR}_{\text{Caucasian}}} = \frac{0{,}46}{0{,}41} = 1{,}12.$$

Completando la tabla con las métricas de disparidad para todas las métricas de grupo obtendremos el siguiente resultado:

race

Attribute Value	PPR Disparity	PPREV Disparity	FDR Disparity	FPR Disparity	FOR Disparity	FNR Disparity
African-American	2.55	1.69	0.91	1.91	1.21	0.59
Asian	0.01	0.72	0.61	0.37	0.43	0.7
Caucasian	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	0.22	0.86	1.12	0.92	1.0	1.17
Native American	0.01	1.92	0.61	1.6	0.58	0.21
Other	0.09	0.6	1.12	0.63	1.05	1.42

Tabla 23: Tabla con las métricas de disparidad para el atributo *race* con umbral del 80 %.

Medidas de equidad

La equidad siempre se define en relación con un grupo de referencia. Podemos ver que el cálculo de la equidad, depende de la métrica de disparidad. En la evaluación del criterio de equidad definiremos un $\tau \in (0,1]$ como umbral de equidad. Diremos que un grupo cumple con la paridad métrica si

$$\tau \leq \text{Métrica de disparidad}_G \leq \frac{1}{\tau}.$$

En nuestro ejemplo, hemos tomado $\tau = 0{,}8$ por lo que cualquier métrica de disparidad se considerará justa si se está contenida en el intervalo $[0{,}8, 1{,}25]$.

En la Tabla 23 hemos destacado en verde los valores contenidos en el intervalo, y en rojo los que se encuentran fuera del mismo. En los resultados finales, si para todos los atributos existe una métrica de disparidad que contiene todos sus valores en color verde, el modelo se evaluará como justo para esa métrica. De lo contrario, lo considerará injusto, y procederá a enumerar los grupos afectados según los criterios de equidad dados.

Resumen de los resultados	
Equal Parity - Todos los grupos protegidos tienen igual representación en el conjunto seleccionado.	Fallo
Proportional Parity - Todos los grupos protegidos son seleccionados de forma proporcional a su porcentaje de la población	Fallo
FPR Parity - Todos los grupos protegidos tienen la misma tasa de falsos positivos que el grupo de referencia.	Fallo
FDR Parity - Todos los grupos protegidos tienen igual proporción de falsos positivos entre el conjunto seleccionado (comparado con el de referencia)	Fallo
FNR Parity - Todos los grupos protegidos tienen la misma tasa de falsos negativos que el grupo de referencia.	Fallo
FOR Parity - Todos los grupos protegidos tienen igual proporción de falsos negativos entre el conjunto no seleccionado (comparado con el de referencia)	Fallo

Tabla 24: Tabla de medidas de equidad con $\tau = 0{,}8$.

En nuestro ejemplo estamos usando la regla del 80 % recomendada por la EEOC (Adverse Impact Analysis / Four-Fifths Rule), donde $\tau = \frac{p}{100}$ y por tanto $\tau = 0{,}8$. En este caso, todas las métricas de disparidad aparecen como injustas, como podemos observar en la Tabla 24. En la versión WEB podemos obtener más detalles sobre cada métrica de disparidad, como podemos ver en la Figura 44 para la FDR Parity.

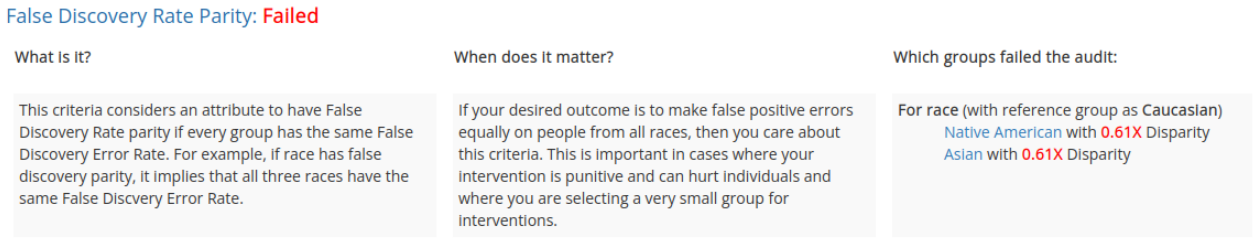


Figura 44: Resultado injusto para la paridad métrica FDR con $\tau = 0{,}8$.

Si cambiamos la regla p % de un 80 % a, por ejemplo, un 60 %, el intervalo del umbral para $\tau = 0{,}6$ sería $[0{,}6, 1{,}67]$, obteniendo los siguientes resultados.

race

Attribute Value	PPR Disparity	PPREV Disparity	FDR Disparity	FPR Disparity	FOR Disparity	FNR Disparity
African-American	2.55	1.69	0.91	1.91	1.21	0.59
Asian	0.01	0.72	0.61	0.37	0.43	0.7
Caucasian	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	0.22	0.86	1.12	0.92	1.0	1.17
Native American	0.01	1.92	0.61	1.6	0.58	0.21
Other	0.09	0.6	1.12	0.63	1.05	1.42

Tabla 25: Tabla con las métricas de disparidad para el atributo *race* con umbral del 60 %.

En la Tabla 25 podemos observar que la paridad de la tasa de falso descubrimiento (FDR) tiene todos los valores de su columna en verde. Aequitas indicará entonces que, basándonos en el umbral establecido ($\tau = 0{,}6$), todos los grupos cumplen la definición de la métrica y por tanto considera justo el concepto de FDR Parity.

Estudio de la equidad

Finalmente utilizaremos la clase *Fairness()* para hacer un estudio más exhaustivo sobre los criterios de paridad estudiados en la Tabla 23.

La Figura 45 muestra los resultados detallados de algunas de las métricas de grupo más populares: las barras verdes representan los grupos para los que el modelo no presenta un sesgo dentro de esa métrica, mientras que las barras rojas indican un sesgo desfavorable en comparación con el grupo de referencia. Seguimos utilizando el umbral de equidad $\tau = 0{,}8$. Los resultados muestran que para cada métrica existe al menos un grupo con algún tipo de sesgo respecto al grupo de referencia. *COMPAS* considera mayoritariamente como de alto riesgo a los menores de 25 años, a los hombres y a los afroamericanos. Según las métricas PPR, que muestran las entidades con *score*=1; vemos que, en comparación con el tamaño de cada grupo, los más jóvenes, los nativos-americanos y los afroamericanos son seleccionados de forma desproporcionada.

Para visualizar las disparidades entre los distintos grupos, la herramienta también produce resultados para las métricas de disparidad. A continuación tenemos que determinar qué medida de sesgo es relevante para nuestro entorno. Dado que en el marco del *COMPAS*, las predicciones se utilizan para tomar decisiones de liberación previa al juicio, las intervenciones serán punitivas (proporcionar esta intervención a los individuos que son falsos positivos les perjudicará), así que deberemos tener en cuenta las tasas de falsos descubrimientos (FDR) y las tasas de falsos positivos (FPR).

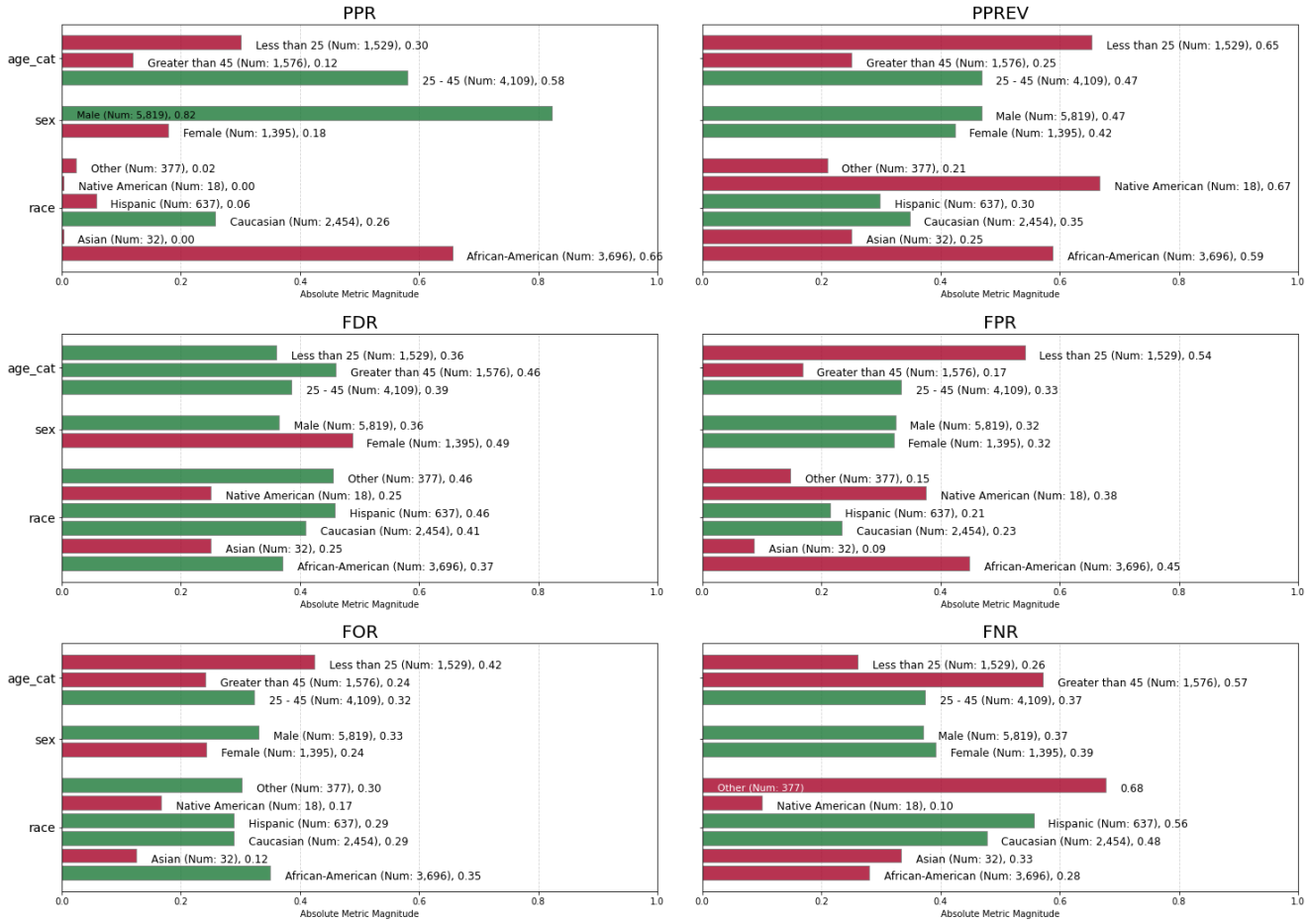


Figura 45: Resultado de las métricas de grupo para COMPAS con $\tau = 0{,}8$.

Observando la Figura 46 podemos ver que el conjunto de datos *COMPAS* es justo para la métrica FDR para la raza, pero la FPR para los afroamericanos es casi el doble que para los caucásicos, por lo que existe un problema evidente de sesgo. También existen problemas de sesgo para los atributos de sexo y edad. En cuanto al sexo, los resultados del FPR son justos, pero el FDR para las mujeres es 1.34 veces mayor que el de los hombres. Por otro lado, si consideramos la distribución de errores falsos positivos teniendo en cuenta la edad, observamos lo contrario: el modelo es justo para el FDR, pero el FPR para los menores de 25 años es 1.62 veces mayor que el de 25-45.

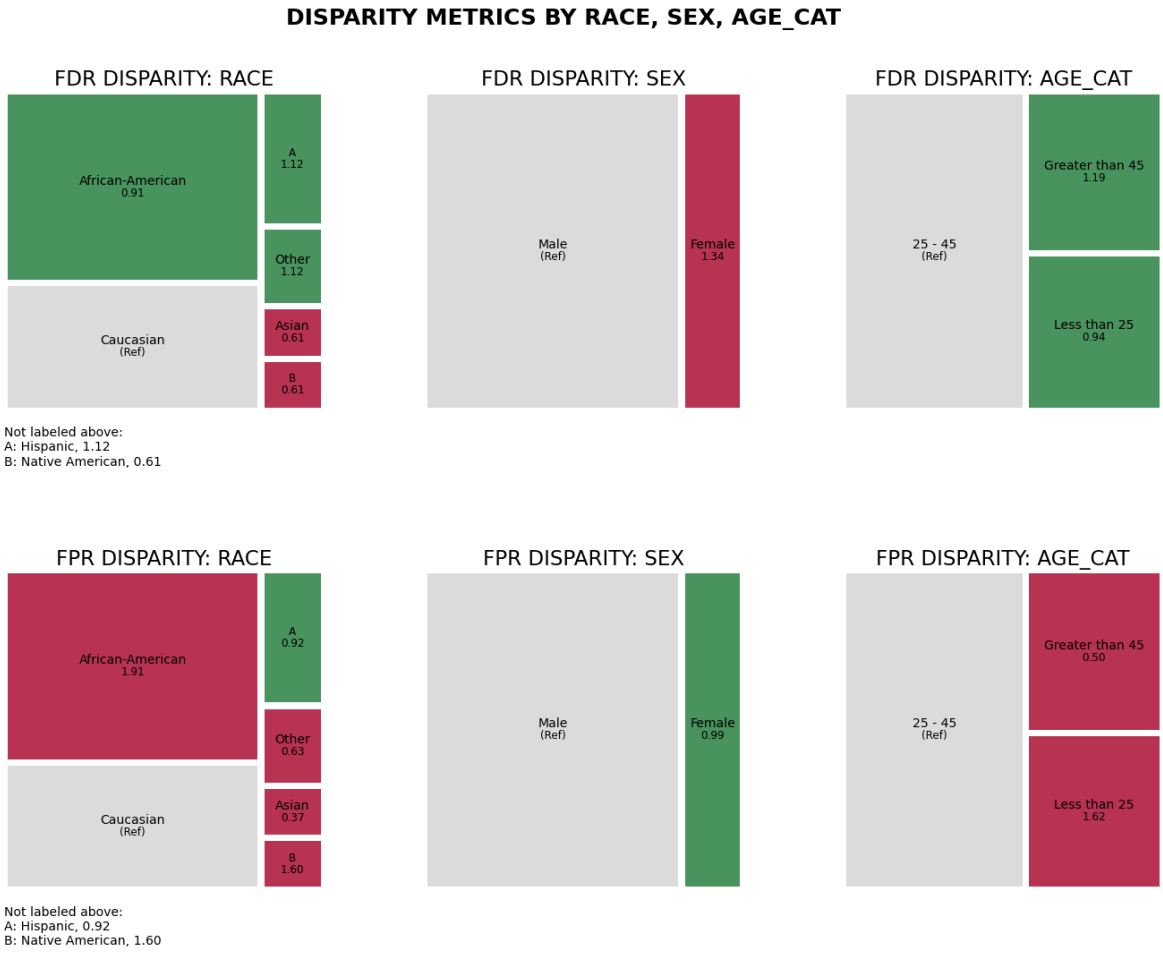


Figura 46: Resultado de las métricas de disparidad para *COMPAS* con $\tau = 0{,}8$.

A.2.2 *Ejemplo: Predicción de notas en la facultad de derecho*

Analizaremos el ejemplo sobre el conjunto de datos *law_data* que presentamos en el Apartado 11.3.1 de nuestra fase experimental. Para tratar la puntuación de riesgo con Aequitas, convertiremos el *score* original (FYA), que toma valores en el intervalo $(-2{,}5, 2{,}5)$, a uno binario: donde 0 indicará una nota baja en el primer año de carrera y 1 una nota alta.

Además eliminaremos cualquier atributo que pudiera entorpecer al estudio de los datos en Aequitas, quedándonos únicamente con las columnas que contienen los valores para los dos atributos protegidos del problema (sexo y raza), la columna con los nuevos valores de *score* y la columna de *label_value*, cuyo valor binario indica la decisión real que fue tomada en el problema.

score	label_value	race	sex
1	1	White	Female
1	1	Asian	Male
0	1	Black	Male

Tabla 26: Ejemplo del *dataset law_data* aportado a Aequitas.

En la Tabla 27, realizamos un estudio previo de la fracción sobre el total de individuos por raza de cada una de las categorías existentes.

Attribute Value	Group Size Ratio	Attribute Value	Group Size Ratio
Amerindian	0	Mexican	0.02
Asian	0.04	Other	0.01
Black	0.06	Puertorican	0.01
Hispanic	0.02	White	0.84

Tabla 27: Distribución de los individuos por *raza*.

Como queremos realizar un análisis lo más general posible, agruparemos los individuos de razas con una representación demasiado pequeña en otra categoría ya existente, en la que puedan figurar. Para ello, añadiremos a la categoría raza hispánica a los individuos de raza mexicana y puertorriqueña, y a la categoría otras a los sujetos de raza amerindia.

Análisis de los datos

En los gráficos siguientes podemos notar una gran diferencia en la distribución de las puntuaciones por raza (Figura 47), con una mayoría de personas de raza blanca pronosticadas con notas altas en el primer año (*score*=1), y una mayoría de población negra e hispánica pronosticada con puntuación baja (*score*=0). En cuanto al atributo sexo (Figura 42), parece que, en principio, no es determinante a la hora de puntuar, aunque existen más individuos de sexo masculino que femeninos puntuados como aptos.

Mostramos mediante varios gráficos cómo se distribuyen las etiquetas reales según los atributos protegidos. Las etiquetas reales contienen la información real sobre si el individuo obtuvo una calificación alta en su primer año (*label_value*=1) o no (*label_value*=0). Con esta información podremos comprobar directamente la precisión de las predicciones.

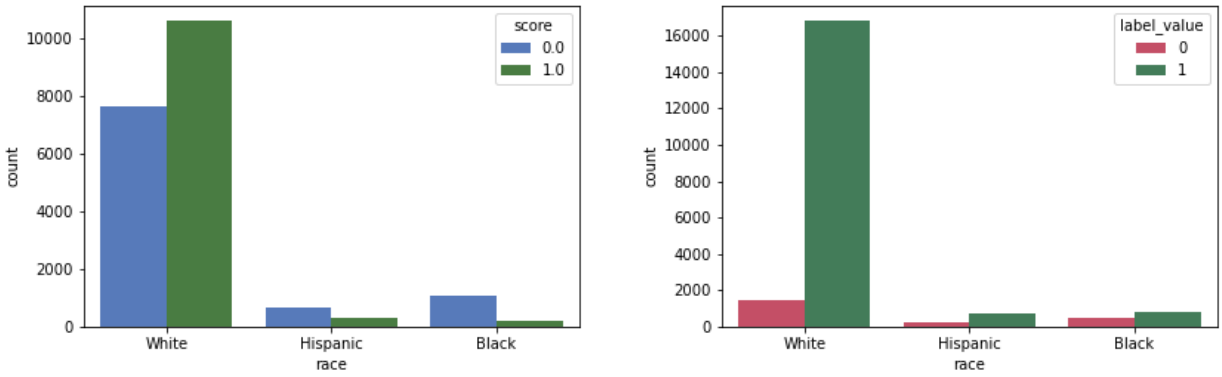


Figura 47: Gráfico de barras de la puntuación y etiqueta real para el atributo *race*.

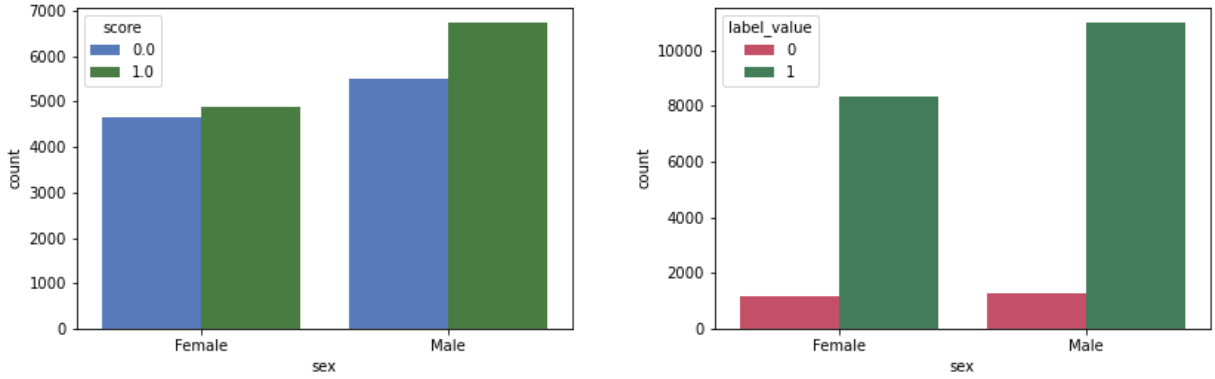


Figura 48: Gráfico de barras de la puntuación y etiqueta real para el atributo *sex*.

En la Figura 47 podemos observar que donde se detecta una mayor diferencia entre las gráficas de puntuación y etiquetas reales es en el grupo de los individuos de raza negra e hispánica. Además, la distribución de etiquetas reales para el atributo sexo de la Figura 48 deja ver que hay muchos individuos que se puntúan con *score*=0, y que finalmente acaban obteniendo buenas notas. Para realizar un estudio de los datos utilizaremos la herramienta de Aequitas, analizando los posibles patrones de sesgo que pudiesen existir entre los diferentes grupos de la población del conjunto de datos *law_data*.

Métricas de grupo

Usando la clase *Group()* calcularemos los valores de las métricas de grupo para los atributos de raza y sexo. Observamos en la Tabla 28 que la mayoría de los individuos son de raza blanca y el resto de categorías está medianamente equilibrada en cuanto a grupos de población. Por otro lado, la Tabla 29 muestra que la distribución del atributo sexo contiene una proporción mínimamente mayor de hombres que de mujeres.

race

Attribute Value	Group Size Ratio	PPR	PPREV	FDR	FPR	FOR	FNR
Asian	0.04	0.02	0.34	0.06	0.1	0.75	0.61
Black	0.06	0.02	0.18	0.15	0.07	0.57	0.76
Hispanic	0.05	0.03	0.32	0.08	0.1	0.67	0.61
Other	0.02	0.01	0.41	0.02	0.05	0.67	0.5
White	0.84	0.91	0.58	0.03	0.2	0.85	0.39

Tabla 28: Tabla con las principales métricas de grupo para el atributo *race*.

sex

Attribute Value	Group Size Ratio	PPR	PPREV	FDR	FPR	FOR	FNR
Female	0.44	0.42	0.51	0.03	0.14	0.78	0.44
Male	0.56	0.58	0.55	0.03	0.16	0.81	0.41

Tabla 29: Tabla con las principales métricas de grupo para el atributo *sex*.

Métricas de disparidad

Tomaremos como atributos de referencia la raza *White* y el sexo *Male*, por ser tanto los mayoritarios, como los valores que históricamente sufren menos impacto dispar.

A continuación usaremos la clase *Bias()* para medir la disparidad entre cada grupo objetivo y los grupos de referencia seleccionados.

Veamos un ejemplo de cómo realiza el cálculo Aequitas en este caso concreto. Mediremos la disparidad para la tasa de falsa omisión (FOR) sobre el grupo de raza *Black* y el grupo de sexo *Female*:

$$\text{FOR}_{\text{G(Black)}} = \frac{\text{FOR}_{\text{Black}}}{\text{FOR}_{\text{White}}} = \frac{0{,}57}{0{,}85} = 0{,}67.$$

$$\text{FOR}_{\text{G(Female)}} = \frac{\text{FOR}_{\text{Female}}}{\text{FOR}_{\text{Male}}} = \frac{0{,}78}{0{,}81} = 0{,}97.$$

Completando la tabla con las métricas de disparidad para todas las métricas de grupo obtendremos los siguientes resultados:

race

Attribute Value	PPR Disparity	PPREV Disparity	FDR Disparity	FPR Disparity	FOR Disparity	FNR Disparity
Asian	0.03	0.58	2.07	0.52	0.88	1.58
Black	0.02	0.3	5.59	0.35	0.67	1.97
Hispanic	0.03	0.55	2.84	0.49	0.79	1.58
Other	0.02	0.71	0.92	0.26	0.8	1.29
White	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 30: Tabla con las métricas de disparidad para el atributo *race* con umbral del 80 %.

sex

Attribute Value	PPR Disparity	PPREV Disparity	FDR Disparity	FPR Disparity	FOR Disparity	FNR Disparity
Female	0.72	0.93	1.16	0.89	0.97	1.08
Male	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 31: Tabla con las métricas de disparidad para el atributo *sex* con umbral del 80 %.

Medidas de equidad

Para cumplir con la regla del 80 % tomamos $\tau = 0{,}8$, por lo que cualquier métrica de disparidad se considerará justa si está contenida en el intervalo $[0{,}8, 1{,}25]$.

En las Tablas 30 y 31 destacamos en verde los valores contenidos en el intervalo, y en rojo los que se encuentran fuera del mismo. Si para todos los atributos existe una métrica de disparidad que contiene todos sus valores en color verde, el modelo se evaluará como justo. De lo contrario, lo considerará injusto. Podemos ver que para este conjunto de datos, y el umbral $\tau = 0{,}8$, no se cumple ninguna medida de equidad.

Resumen de los resultados	
Equal Parity - Todos los grupos protegidos tienen igual representación en el conjunto seleccionado.	Fallo
Proportional Parity - Todos los grupos protegidos son seleccionados de forma proporcional a su porcentaje de la población	Fallo
FPR Parity - Todos los grupos protegidos tienen la misma tasa de falsos positivos que el grupo de referencia.	Fallo
FDR Parity - Todos los grupos protegidos tienen igual proporción de falsos positivos entre el conjunto seleccionado (comparado con el de referencia)	Fallo
FNR Parity - Todos los grupos protegidos tienen la misma tasa de falsos negativos que el grupo de referencia.	Fallo
FOR Parity - Todos los grupos protegidos tienen igual proporción de falsos negativos entre el conjunto no seleccionado (comparado con el de referencia)	Fallo

Tabla 32: Tabla de medidas de equidad con $\tau = 0{,}8$.

Estudio de la equidad

Finalmente utilizaremos la clase *Fairness()* para hacer un estudio más exhaustivo sobre los criterios de paridad estudiados en las Tablas 30 y 31.

La Figura 49 muestra los resultados detallados de algunas de las métricas de grupo más populares. Los resultados muestran que para cada métrica existe al menos un grupo con algún tipo de sesgo respecto al grupo de referencia. En el ámbito del problema, dado que las predicciones se utilizan para tomar decisiones de asistenciales (proporcionar esta asistencia a los individuos que son falsos positivos no les perjudicará), así que deberemos tener en cuenta las tasas de falsas omisiones (FOR) y las tasas de falsos negativos (FNR).

Figura 49: Resultado de las métricas de grupo para *law_data* con $\tau = 0{,}8$.

Observando la Figura 50 podemos ver que el conjunto de datos *law_data* no es justo para ninguna de las métricas para el atributo raza, siendo la FNR para los individuos de raza negra casi el doble que para los de raza blanca y 1.5 veces mayor para el resto de razas. La tasa de falsos negativos (FNR) nos dice que los individuos de cualquier raza tienen casi entre un 1.5-2 veces más de probabilidad de ser etiquetados con notas bajas, cuando realmente obtendrían una mejor calificación. Esto hace evidente la existencia de un problema de sesgo para esta métrica. En cuanto al sexo, los resultados son justos para ambas métricas. Esto nos puede dar la idea de que en este problema, no existe un sesgo determinante para el sexo, conclusión que ya obtuvimos en la Sección 12.3.

Figura 50: Resultado de las métricas de disparidad para *law_data* con $\tau = 0{,}8$.

B

ESTIMACIÓN DEL COSTE Y PLANIFICACIÓN

En este apéndice se realizará una estimación, tanto del coste como de la planificación del trabajo durante el período de desarrollo, con el objetivo de simular la valoración y presupuesto de un proyecto real en el ámbito laboral.

ESTIMACIÓN DEL PRESUPUESTO DEL PROYECTO

Haremos un presupuesto del proyecto, en el que incluiremos las horas dedicadas a cada tema estudiado y realizaremos una estimación a precio 7 euros/hora. El análisis del presupuesto se puede observar en la Tabla 33.

Para facilitar la estimación de los costes, hemos dividido el trabajo en tres partes:

- **Parte teórica**: recoge las tareas relacionadas con el análisis y el estudio de los conceptos de carácter teórico contenidos en el trabajo. Incluiremos la formalización de las nociones básicas para el proyecto y el desarrollo de las demostraciones matemáticas.
- **Parte práctica**: reúne las prácticas relacionadas con la programación, análisis y validación de los experimentos. Además tendremos en cuenta el equipo utilizado y los tiempos dedicados a instalación de bibliotecas y software empleados.
- **Parte general**: agrupa las labores de elaboración de la memoria y reuniones con los tutores (presencial u online).

El cómputo total es de 8.047 euros. Teniendo en cuenta que el período de trabajo útil ha sido de aproximadamente 5 meses, el sueldo medio mensual equivale a 1.609 euros brutos. Para un trabajador sin experiencia, este valor es bastante fiel a la realidad.

PLANIFICACIÓN DEL TRABAJO

El diseño de la planificación se ha realizado con el software GanttProject, que es un programa de código abierto utilizado para administrar proyectos, pudiendo usar entre otras muchas herramientas, un diagrama de Gantt. Para el desarrollo de la planificación, hemos usado la misma división en partes que la presentada en la sección anterior.

En la planificación original nos habría gustado presentar el trabajo para la convocatoria de septiembre, pero debido a algunos problemas derivados de la carga docente a lo largo del curso 2020-2021 y a la concreción de los experimentos asociados al proyecto, se ha acabado retrasando hasta el mes de noviembre.

Además es destacable, si comparamos las Figuras 51 y 52, que el tiempo dedicado tanto al estudio como la formalización de las diferentes medidas de equidad, ha sido mayor al previsto inicialmente.

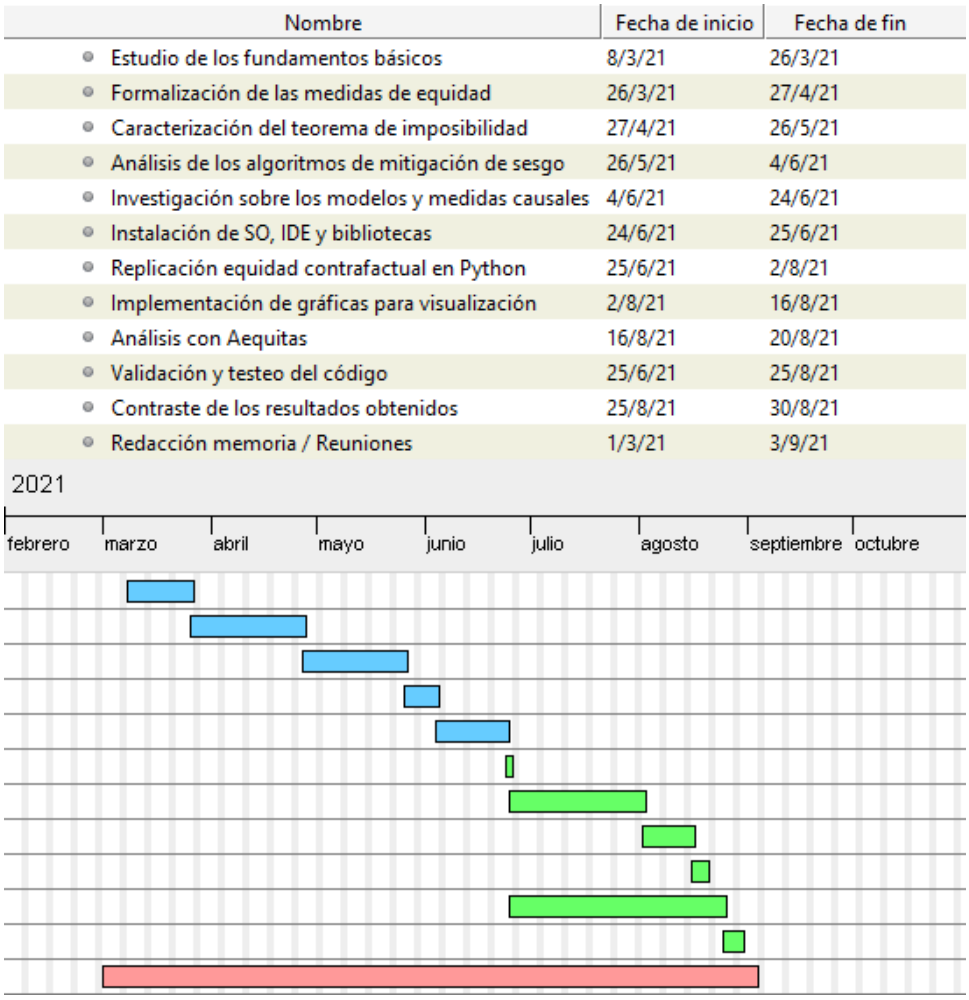


Figura 51: Planificación original del desarrollo del trabajo.

En ambos casos, la parte general es una tarea que se desarrolla a lo largo del período de trabajo, ya que la redacción de la memoria y las tutorías se realizan de forma simultánea al análisis teórico y práctico del proyecto.

En la planificación final, hemos establecido el mes de abril como inicio del proyecto. Además el periodo de exámenes de junio también ralentizó en gran medida el avance del proyecto propuesto. Todos estos factores se pueden intuir observando la Figura 52, donde indicamos en color azul la parte teórica, en verde la parte práctica y en rojo la parte general.

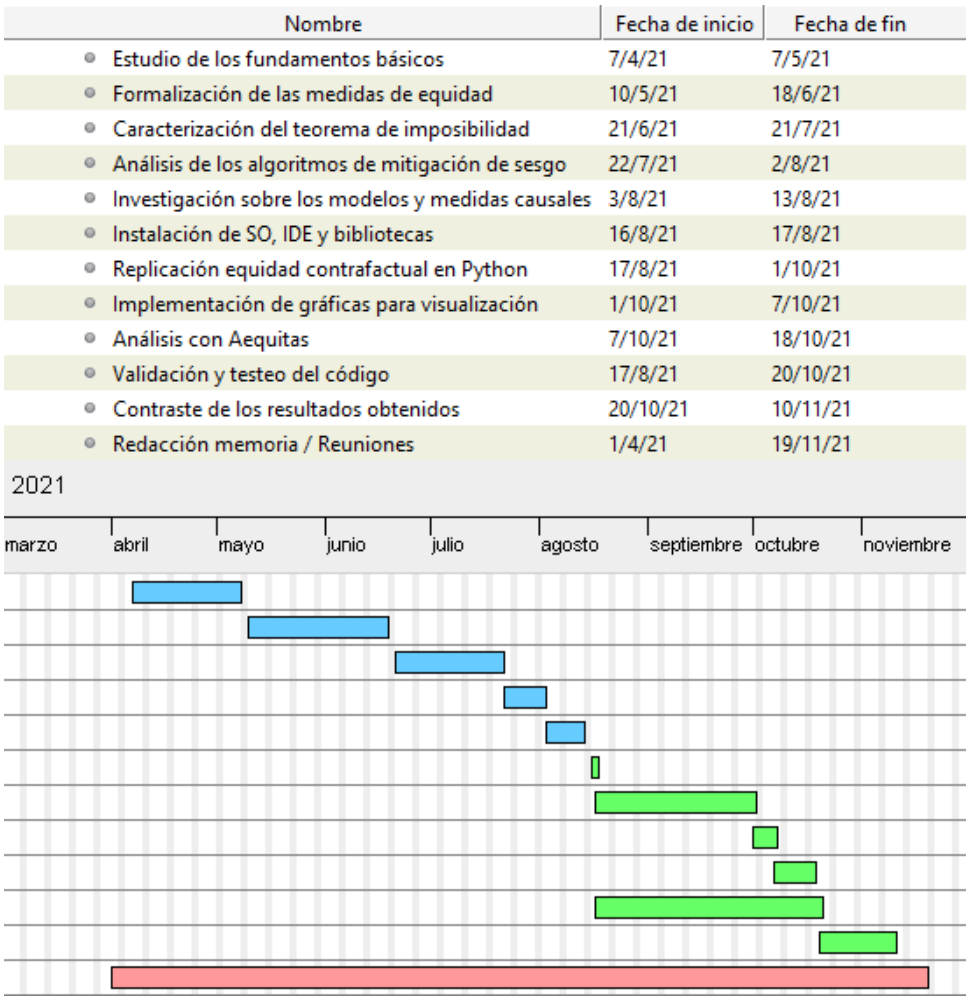


Figura 52: Planificación final del desarrollo del trabajo.

Concepto	Tiempo *(horas)*	Coste *(euros/hora)*	Coste total *(euros)*
Parte general			
Redacción de la memoria	300	7	2.100
Reuniones con los tutores	14	-	-
Parte teórica			
Estudio de los fundamentos básicos	80	7	560
Formalización de las medidas de equidad	150	7	1.050
Caracterización del teorema de imposibilidad	90	7	630
Análisis de los algoritmos de mitigación de sesgo	32	7	224
Investigación sobre los modelos y medidas causales	45	7	315
Parte práctica			
Equipo de trabajo: ASUS TUF	-	-	949
Instalación de SO, IDE y bibliotecas	3	7	21
Replicación de un modelo de equidad contrafactual en Python	120	7	840
Implementación de gráficas para visualización	16	7	112
Análisis con Aequitas	28	7	196
Validación y testeo del código	110	7	770
Contraste de los resultados obtenidos	40	7	280
Cómputo total del proyecto	1.028	-	8.047

Tabla 33: Estimación del coste del proyecto.