Universidad de la República
Facultad de Ingeniería

CEDUCA - Ciencia de Datos y Educación en Uruguay: Evaluación de la Equidad desde un Enfoque Interseccional

Memoria de proyecto presentada a la Facultad de Ingeniería de la Universidad de la República por

Carmen Beatriz Salinas De la Vega, María Victoria Tournier Nieto

en cumplimiento parcial de los requerimientos
para la obtención del título de
Ingeniera en Sistemas de Comunicación.

Tutoras

Lorena Etcheverry	Universidad de la República
Aiala Rosá	Universidad de la República
Noelia Beltramelli	Universidad de la República

Tribunal

Mercedes Marzoa	Universidad de la República
Libertad Tansini	Universidad de la República
Martín Rocamora	Universidad de la República

Montevideo
lunes 2 diciembre, 2024

CEDUCA - Ciencia de Datos y Educación en Uruguay: Evaluación de la Equidad desde un Enfoque Interseccional, Carmen Beatriz Salinas De la Vega, María Victoria Tournier Nieto.

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 84 páginas.
Compilada el lunes 2 diciembre, 2024.
http://iie.fing.edu.uy/

Agradecimientos

Queremos expresar nuestro más sincero agradecimiento a todas las personas que hicieron posible la realización de este proyecto y que fueron fundamentales en su desarrollo. En especial reconocer y agradecer a nuestras tutoras Lorena Etcheverry, Aiala Rosá y Noelia Beltramelli, así como a María Goñi, por su constante disposición y guía a lo largo de todo el proceso. También queremos destacar a aquellas personas que se mostraron dispuestas a responder nuestras preguntas y a reunirse con nosotras para compartir su conocimiento: Juan Valle Lisboa, Álvaro Cabana, Camila Zugarramurdi, Daniel Alessandrini,Libertad Tansini, Santiago Cardozo y Cristian Cechinel. Finalmente, extendemos nuestro más sentido agradecimiento a nuestras familias, amigos y todas las personas que nos acompañaron y apoyaron a lo largo de nuestra carrera universitaria. Sin su comprensión y apoyo no habría sido posible.

Esta página ha sido intencionalmente dejada en blanco.

Resumen

La motivación principal de este proyecto es explorar el creciente uso de la ciencia de datos en la educación, estudiando sus aplicaciones desde una perspectiva de equidad y feminista, que considera no sólo el género, sino también la raza, clase y otras dimensiones de la identidad, con el fin de garantizar un manejo más inclusivo y equitativo de los datos. El objetivo general es aprender y recopilar información sobre cómo se están implementando estos sistemas, qué herramientas existen para realizar análisis de equidad en algoritmos, cómo aplicar técnicas de mitigación de sesgos y qué implicaciones tienen. Además, este proyecto responde al creciente interés institucional y regional por buscar soluciones innovadoras en el ámbito educativo a través de la ciencia de datos. Esta tendencia refleja un espíritu de colaboración y progreso en la región, orientado a mejorar los sistemas educativos.

En este contexto, se realiza una revisión bibliográfica que aborda conceptos como el feminismo de datos y la equidad en la ciencia de datos, incluyendo la interseccionalidad y el desarrollo de un proceso de evaluación de equidad en sistemas algorítmicos. Dicho proceso considera tanto los sesgos en los algoritmos como las causas que los originan. Se lleva a cabo un relevamiento de proyectos en Uruguay relacionados con ciencia de datos y educación. Este relevamiento pretende contribuir a visibilizar iniciativas y a mejorar el acceso al conocimiento en este campo en el país. Finalmente, se aplica un proceso de evaluación de equidad utilizando los conceptos y las tecnologías estudiadas en un caso real de un sistema de ciencia de datos aplicado en el ámbito educativo.

Esta página ha sido intencionalmente dejada en blanco.

Tabla de contenidos

Capítulo 1

Introducción

La ciencia de datos ha demostrado ser una herramienta poderosa para mejorar procesos en diversos ámbitos y la educación no es una excepción. Sin embargo, se hace cada vez más evidente la necesidad de garantizar que estos no perpetúen desigualdades ni refuercen estereotipos preexistentes. La creciente preocupación por la equidad en los algoritmos que procesan y analizan datos educativos nos lleva a reflexionar sobre cómo se diseñan e implementan estos sistemas.

El proyecto *CEDUCA – Ciencia de Datos y Educación en Uruguay: Evaluación de la Equidad desde un Enfoque Interseccional*[1] surge de la necesidad de examinar críticamente la equidad en los sistemas algorítmicos aplicados al ámbito educativo. Este estudio además de plantearse como una exploración técnica de la ciencia de datos en educación, se plantea como una oportunidad para integrar enfoques feministas e interseccionales que permitan entender cómo las desigualdades pueden influir en los procesos de recolección, análisis y uso de los datos con impacto social. Al hacerlo, buscamos también promover la creación de soluciones más inclusivas y mejorar las prácticas que las rodean, contribuyendo así a un sistema educativo más equitativo.

Por un lado, el proyecto responde a la demanda creciente de herramientas basadas en datos que puedan ayudar a mejorar el rendimiento académico y facilitar la toma de decisiones en el sector educativo. Por otro, resalta el desafío de aplicar estos avances de manera justa, considerando las diversas realidades de los estudiantes. La ciencia de datos, si no se utiliza con una perspectiva crítica, corre el riesgo de exacerbar las desigualdades que debería ayudar a mitigar. Por tanto, este proyecto no solo se enfoca en el desarrollo y análisis técnico de herramientas, sino en el diseño de procesos de evaluación de equidad que sirvan para mitigar el sesgo algorítmico en los sistemas educativos.

En el contexto de Uruguay, CEDUCA pretende sentar las bases para futuras investigaciones en este campo. Al centralizar información sobre los proyectos de ciencia de datos en educación en el país, el proyecto busca también visibilizar el trabajo de diferentes actores e impulsar un diálogo sobre cómo mejorar la equidad en el uso de datos en la educación. Este estudio cobra particular importancia en un momento en que las instituciones educativas de la región están interesadas en implementar tecnologías que puedan ayudar a reducir la brecha de acceso y rendimiento académico.

1.1. Objetivos

El objetivo general de este proyecto es abordar la equidad en la ciencia de datos en ámbitos educativos y promover perspectivas más igualitarias. Para afrontar este desafío, se analizan diversas formas de estudiar la

[1] El repositorio que contiene los análisis realizados en este proyecto se encuentra en `https://gitlab.fing.edu.uy/maria.victoria.tournier/ceduca`.

equidad en la ciencia de datos, incorporando conceptos como el feminismo de datos. Luego, se busca aplicar estas técnicas de análisis en proyectos que empleen algoritmos de clasificación en el ámbito educativo.

Dentro de los objetivos particulares para cumplir el objetivo general, el primero es llevar a cabo un relevamiento de proyectos educativos en Uruguay. El propósito es investigar qué se está haciendo en el país y qué información está disponible públicamente. Consideramos fundamental centralizar esta información para que esté al alcance de todos, con el objetivo de aumentar la conciencia pública y respaldar futuras investigaciones en este ámbito. Para realizar este relevamiento, se intercambió directamente con los investigadores y responsables de los proyectos.

Otro objetivo particular es familiarizarse con el estudio de sesgos algorítmicos mediante la aplicación de distintas herramientas utilizadas en el área y de técnicas de mitigación de sesgos. Para cumplir con este objetivo, se realiza un estudio de caso que permita aplicar dichas técnicas y herramientas.

1.2. Alcance

El alcance de este proyecto implica realizar una revisión de la literatura sobre la equidad en ciencia de datos y examinar estudios previos relacionados con proyectos de ciencia de datos en entornos educativos, con un enfoque particular en el rendimiento académico. Además, se lleva a cabo un relevamiento de sistemas que aplican ciencia de datos en el ámbito educativo en Uruguay, estableciendo contacto y realizando entrevistas con los involucrados en dichos proyectos para recopilar información relevante. El proyecto incluye un estudio de caso en el que se analizan datos y modelos proporcionados por un sistema real, investigando posibles sesgos y prejuicios que puedan surgir en su uso. Para ello, se acuerda previamente una definición de equidad y se determinan las poblaciones a considerar en el análisis, garantizando la privacidad y seguridad de los datos de los estudiantes, así como la transparencia en las decisiones tomadas por el algoritmo. A partir de este análisis de equidad, se proponen posibles mejoras a la solución original aplicando técnicas de mitigación de sesgos algorítmicos. Se excluye del alcance del proyecto la adquisición de los datos, el desarrollo de la herramienta predictiva original y el análisis social.

1.3. Organización del Documento

Luego de presentar el **Capítulo 1** introductorio, el documento se organiza cómo sigue:

- **Capítulo 2**: Se presentan los principales conceptos a trabajar desde la ciencia de datos y cómo se aplica a la educación pasando por los principales principios de Feminismo de Datos y finalmente se especifican las métricas utilizadas para la equidad en ciencia de datos.
- **Capítulo 3**: Presenta el relevamiento realizado de aplicaciones de ciencia de datos en la educación uruguaya y expone, con una perspectiva de feminismo de datos, un análisis sobre los proyectos relevados.
- **Capítulo 4**: En base a las herramientas estudiadas para análisis de sesgos algorítmicos y mitigación de los mismos se presenta un estudio de caso donde se aplican técnicas y se evalúan desempeños de las mismas.
- **Capítulo 5**: Se presentan las conclusiones del proyecto y trabajo a futuro.

Al final del documento se presentan los **Anexos A y B**, que surgen de realizar un análisis en más profundidad de dos de los sistemas relevados.

Capítulo 2

Ciencia de Datos y Equidad

Este capítulo aborda varias perspectivas sobre la equidad en la ciencia de datos, explorando conceptos clave y estrategias para identificar y mitigar los sesgos. Antes de profundizar en los aspectos relacionados con la equidad, se explican conceptos de la ciencia de datos que son esenciales para comprender el contexto.

2.1. Conceptos de Ciencia de Datos

La ciencia de datos es un campo interdisciplinario que se enfoca en extraer información valiosa a partir de datos. En el documento *50 Years of Data Science* [25], Donoho define la ciencia de datos como la ciencia de aprender a partir de los datos. Se explica que no solo incluye la teoría y las técnicas estadísticas, sino que también abarca la preparación de datos, la visualización, la gestión de grandes volúmenes de datos heterogéneos y la presentación de resultados. Los resultados deben ser replicables, transparentes, verificables y pueden generar nuevos conocimientos o teorías a partir de los mismos.

Dentro del campo de la ciencia de datos, el aprendizaje automático es actualmente uno de sus pilares. Este enfoque se utiliza cuando no existe una solución analítica exacta para el problema. Se basa en encontrar aproximaciones o reglas que describen patrones para hacer predicciones o clasificaciones basadas en datos. En la Figura 2.1 se presenta un diagrama sobre un problema genérico de aprendizaje automático donde se pueden diferenciar las etapas del proceso. El proceso incluye la selección de un modelo o algoritmo a utilizar, su entrenamiento con datos, y su posterior aplicación para hacer predicciones o clasificaciones sobre nuevos datos. Una vez entrenado, el modelo se evalúa utilizando un conjunto de prueba, que consiste en datos que no se han visto antes, para medir su capacidad de generalización y precisión en la predicción de nuevas observaciones.

2.1.1. Modelos de Predicción

Mediante el proceso de entrenamiento, que consiste en una serie de ajustes y optimizaciones iterativas, los modelos de predicción modifican sus parámetros internos para identificar patrones en los datos de entrenamiento y, finalmente, generar predicciones sobre datos no vistos.

Uno de los enfoques más comunes es el aprendizaje automático supervisado, en el cual los datos de entrenamiento incluyen ejemplos con las salidas correctas conocidas para cada entrada. Los modelos se especializan en resolver problemas específicos, como regresión, que se utiliza para predecir valores continuos, como el precio de una vivienda, o clasificación, que se emplea para predecir o clasificar valores discretos. Los algoritmos de clasificación pueden ser binarios, cuando predicen entre dos posibles categorías, como identificar si un correo electrónico es “spam” o “no spam”, o de clasificación multi-clase, cuando se predicen más de dos categorías, como en la clasificación de distintas especies de animales en imágenes.

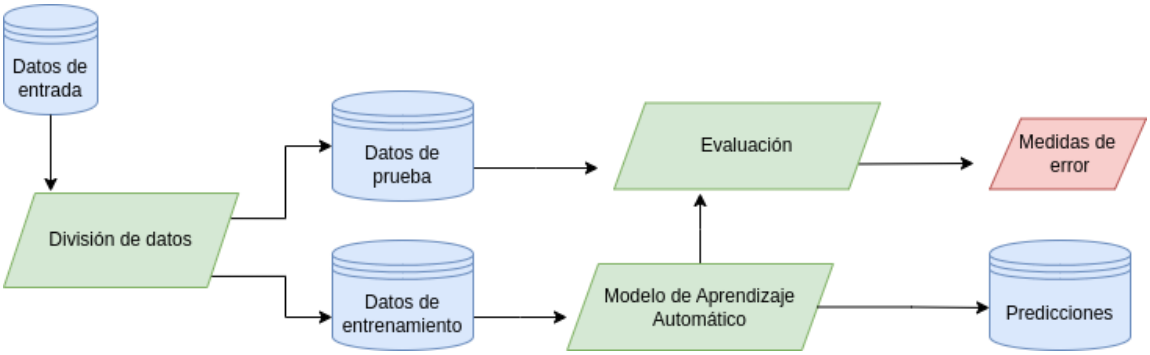


Figura 2.1: Diagrama general de un problema de Aprendizaje Automático.

Además, algunos algoritmos de regresión, como la regresión lineal o logística, pueden ser utilizados para clasificación al estimar la probabilidad de que una instancia pertenezca a una clase particular. Si la probabilidad estimada supera un umbral predefinido, el modelo predice que la instancia pertenece a esa clase.

A continuación, se describen los modelos más utilizados en los trabajos estudiados, basados en el libro de Aurélien Géron [33].

2.1.1.1. Regresión Lineal

El modelo de regresión lineal se utiliza para predecir un valor continuo a partir de un conjunto de características. La predicción de un modelo de regresión lineal está dada por la ecuación:

$$\hat{y} = h_\theta(\mathbf{x}) = \theta \cdot \mathbf{x}$$

donde h_θ es la función hipótesis, que usa los parámetros del modelo θ, θ es el vector de parámetros del modelo, que contiene el término independiente θ_0 y los pesos de las características θ_1 a θ_n, $\mathbf{x}$ es el vector de características de la instancia, que contiene desde x_0 hasta x_n, con x_0 siempre igual a 1 y $\theta \cdot \mathbf{x}$ es el producto punto entre los vectores θ y $\mathbf{x}$, que es igual a $\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$.

2.1.1.2. Regresión Logística

La regresión logística es un modelo de clasificación que estima la probabilidad de que una instancia pertenezca a una clase. La probabilidad estimada se calcula mediante la ecuación:

$$\hat{p} = h_\theta(\mathbf{x}) = \sigma(\theta^\top \mathbf{x})$$

donde $\sigma(\cdot)$ es la función sigmoide, que devuelve un valor entre 0 y 1. Esta función está definida como:

$$\sigma(t) = \frac{1}{1 + \exp(-t)}.$$

2.1.1.3. Árboles de Decisión

Un árbol de decisión es un algoritmo no paramétrico que se utiliza tanto para tareas de clasificación como de regresión. La estructura del árbol está compuesta por nodos de decisión y nodos hoja, donde los primeros representan divisiones basadas en características, y los segundos indican predicciones finales.

El objetivo del árbol de decisión es crear un modelo que prediga el valor de una variable objetivo aprendiendo reglas de decisión simples inferidas de las características de los datos. Es fácil de interpretar y visualizar, lo que lo convierte en un modelo muy popular.

2.1.1.4. Ensamblado de Modelos

El *ensamblado de Modelos* es una técnica que combina las predicciones de múltiples modelos para obtener un modelo más robusto y preciso que los modelos individuales. Dentro de esta técnica, existen varios enfoques destacados:

- **Random Forest**: El modelo de *Random Forest* es un conjunto de árboles de decisión. Cada árbol se entrena de manera independiente sobre una muestra diferente de los datos, y las predicciones finales se obtienen mediante un promedio de las predicciones de los árboles individuales.
- **AdaBoost (Adaptive Boosting)**: AdaBoost es un algoritmo de *boosting* que entrena modelos de manera secuencial, corrigiendo los errores de los clasificadores anteriores en cada paso. En cada iteración, AdaBoost asigna un peso mayor a las instancias mal clasificadas por el clasificador anterior, enfocándose en corregir los errores. Al final, el modelo realiza una votación ponderada de todos los clasificadores para generar una predicción.
- **Voting Classifier**: El *Voting Classifier* es una técnica de ensamblado en la que se combinan las predicciones de varios modelos diferentes (como regresión logística, árboles de decisión, etc.). Existen dos tipos de votación:
  - **Hard voting**: La clase más votada por todos los modelos es la clase final predicha.
  - **Soft voting**: En lugar de elegir la clase directamente, se promedian las probabilidades de las clases predichas, y la clase con mayor probabilidad promedio es la predicción final.

2.1.2. Métricas de Desempeño

Para evaluar el rendimiento de los algoritmos, una de las herramientas que se suele utilizar es la matriz de confusión. Permite visualizar las predicciones de un modelo en comparación con los valores reales. La matriz de confusión de un sistema de clasificación binario se organiza en cuatro cuadrantes, como se muestra en la Tabla 2.1.

	Predicción Positiva	**Predicción Negativa**
Positivo Real	Verdadero Positivo (VP)	Falso Negativo (FN)
Negativo Real	Falso Positivo (FP)	Verdadero Negativo (VN)

Tabla 2.1: Matriz de confusión.

Los verdaderos positivos (VP) hacen referencia a cuando el sistema predice correctamente la clase positiva. Por otro lado los falsos positivos (FP) son aquellas instancias en las cuales se predice que es positiva cuando en realidad no lo es. En el caso de los verdaderos negativos (VN) se da cuando el sistema predice correctamente una instancia como negativa, mientras que los falsos negativos (FN) se refiere a cuando el sistema predice que la instancia es negativa cuando no lo es.

En este contexto, el porcentaje de acierto, la precisión, la sensibilidad y la especificidad son métricas importantes para evaluar el rendimiento de un modelo. Son detalladamente explicadas por Joel Grus en su libro *Data Science from Scratch* [36].

El **porcentaje de acierto** es una métrica que mide la proporción de predicciones correctas sobre el total de predicciones realizadas. Indica qué tan frecuentemente el modelo acierta en sus predicciones. Se calcula como:

$$\text{Porcentaje de acierto} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}.$$

La **precisión** mide la proporción de instancias predichas como positivas que realmente lo son. Se calcula de la siguiente manera:

$$\text{Precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}}.$$

Por otro lado, la **sensibilidad** mide la capacidad del modelo para identificar correctamente todas las instancias positivas. Su fórmula es:

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

Finalmente, la **especificidad** mide la capacidad del modelo para identificar correctamente todas las instancias negativas. Se calcula como:

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}}.$$

Las métricas presentadas se utilizan a lo largo del documento para evaluar el desempeño de los sistemas estudiados.

2.1.3. Ciencia de Datos en la Educación

Cuando se habla de ciencia de datos en la educación, se hace referencia a una rama que se enfoca en el análisis y uso de datos relacionados con el ámbito educativo. Su objetivo principal es mejorar los resultados de aprendizaje, optimizar las prácticas docentes, influir en las políticas educativas y fomentar la innovación en el ámbito educativo. Dentro de este campo, se utilizan términos como *minería de datos educativos*, *analítica de aprendizaje* y *ciencia de datos educativos*, que aunque tienen diferencias sutiles, comparten el mismo propósito general de aplicar técnicas de ciencia de datos para resolver problemas educativos y mejorar el sistema educativo en su conjunto.

El manejo de datos en el ámbito educativo implica varios desafíos que deben ser considerados, además de la necesidad de garantizar su protección y privacidad debido a su alta sensibilidad. Los datos relacionados con la educación presentan ciertas particularidades, destacadas por Piety, Hickey y Bishop (2014) [50]:

- **Creación humana**: Gran parte de los datos educativos requiere intervención humana, lo que aumenta la posibilidad de errores y manipulaciones.
- **Imprecisión en la medición**: Presentan problemas de precisión, especialmente cuando se trata de evaluaciones del aprendizaje estudiantil o de las capacidades sistémicas. Pueden ser inexactas y estar influenciadas por factores como el contexto socioeconómico del estudiante, las técnicas de instrucción y las condiciones en las que se realiza la prueba.
- **Desafíos de comparabilidad**: Las diferencias entre distintos centros educativos pueden complicar la interpretación de los datos.
- **Fragmentación**: Los datos educativos están altamente fragmentados. Muchas organizaciones distintas poseen partes de la información educativa, lo que dificulta el acceso y la estandarización.

Ejemplos de proyectos de ciencia de datos en educación incluyen la implementación de sistemas de recomendación que sugieren recursos educativos personalizados basados en el desempeño y las necesidades individuales de los estudiantes, el análisis de datos de interacción en plataformas de aprendizaje en línea y la predicción del desempeño estudiantil. En ese contexto, varios países de la región comenzaron a desarrollar *Sistemas de Alerta Temprana (SAT)* para prevenir un evento no deseado en la trayectoria educativa de un estudiante, como la desvinculación o repetición [7].

Estos sistemas se enfocan en la prevención oportuna, lo que implica identificar a los estudiantes que presentan un riesgo elevado con suficiente antelación para permitir la implementación de medidas de apoyo que fortalezcan dichas trayectorias. Esta estimación no garantiza con certeza que el evento ocurra, sino que ofrece una probabilidad basada en los indicadores observados. Un documento elaborado por Perusia y Cardini (2021) [49] presenta un listado de SAT implementados en América Latina, así como los principales indicadores utilizados para identificar a los estudiantes en riesgo.

Después de identificar a los estudiantes en riesgo, se implementan protocolos de intervención a nivel individual o grupal. Es crucial realizar una evaluación del funcionamiento del SAT. Este proceso de retroalimentación permite ajustar y mejorar continuamente el sistema, analizando la efectividad de las intervenciones y refinando los criterios para identificar riesgos.

2.2. Feminismo de Datos

El feminismo de datos se centra en analizar cómo las prácticas estándar en la ciencia de datos sirven para reforzar las desigualdades existentes y, en segundo lugar, en utilizar la ciencia de datos para desafiar y cambiar la distribución del poder. De acuerdo con D'Ignazio y Klein [24], el feminismo de datos implica una forma de pensar sobre los datos, tanto en sus usos como en sus límites, que se basa en la experiencia directa, el compromiso con la acción y el pensamiento feminista. Además, el feminismo de datos no se limita únicamente al género, sino que también aborda aspectos como raza, clase, sexualidad, edad, religión, geografía y más, reconociendo que estos factores influyen de manera conjunta en la experiencia y oportunidades de cada persona en el mundo. El feminismo de datos se desarrolla en base a los siguientes siete principios:

1. **Examinar el poder**: Busca analizar cómo operan las estructuras de poder. Este proceso implica identificar y comprender las estructuras de opresión, que a menudo pasan desapercibidas, especialmente para aquellos en posiciones privilegiadas. En el ámbito de los datos, esto implica reconocer quiénes se benefician (y quiénes no) de los trabajos de ciencia de datos, quiénes están llevando a cabo ese trabajo y cuáles son las metas prioritarias, entre otros aspectos relevantes. Las preguntas que se sugieren hacer para examinar el poder son:

    - **¿Quién realiza el trabajo?**: Todos los individuos están influenciados por su experiencia personal que introduce un sesgo en su perspectiva. Por este motivo es fundamental analizar el contexto del cual provienen quienes llevan a cabo los trabajos. Aquellos pertenecientes a grupos privilegiados pueden experimentar el fenómeno conocido como "peligro del privilegio" (*privilege hazard*), que ocurre cuando la persona no es consciente de su propio sesgo. Esta falta de reconocimiento se debe a la dificultad de percibir estos sesgos desde una posición de privilegio.
    - **¿Quién se beneficia del trabajo?¿Y quién es invisibilizado?**: Sobre esta pregunta se busca estudiar el conjunto de datos. Es esencial analizar si el conjunto de datos es representativo de la realidad o qué realidad se está intentando representar, ya que esto puede generar un desfavorecimiento de ciertas poblaciones.
    - **¿Qué beneficios se obtienen?**: Esta pregunta busca indagar el propósito del trabajo realizado, con el objetivo de analizar si existe algún interés sesgado por parte de quienes lo implementan. El análisis puede dividirse en tres categorías: estudios realizados por universidades, estudios realizados para el gobierno y estudios realizados por corporaciones o empresas. Cada una de estas instituciones persigue fines específicos al llevar a cabo el trabajo, por lo que es crucial examinar si obtienen algún beneficio derivado del mismo.

2. **Desafiar el poder**: Implica tomar acción, movilizar la ciencia de datos para contrarrestar las estructuras de poder. Consiste en cuatro puntos de partida:

    - **Recolectar**: Confrontar la falta de datos mediante la compilación de *contradatos*. La falta de datos se da cuando información relevante para el bienestar de grupos de personas no es recolectada o suele estar incompleta, difícil de acceder y subrepresentada. En un estudio del año 2022 [23] se profundiza en el concepto de compilar contradatos, discutiendo el trabajo realizado por diez organizaciones activistas y de la sociedad civil en seis países, quienes combaten la falta de datos sobre los feminicidios. Uno de los proyectos estudiados es Feminicidio Uruguay [39], una base de datos y mapa interactivo sobre los feminicidios en Uruguay, liderado por la activista y comunicadora social Helena Suárez Val.
    - **Analizar**: Analizar y auditar algoritmos para dar pruebas de sesgos y daños generados por sistemas de ciencias de datos, así como responsabilizar a las instituciones. Diakopoulos [22] utiliza el concepto de *responsabilidad algorítmica*, argumentando cómo los algoritmos están tomando decisiones cada vez más importantes en diversos aspectos de la vida y sin embargo son a menudo *cajas negras*, siendo difícil entender cómo funcionan y cómo afectan a las personas,

lo que plantea preocupaciones sobre su imparcialidad, sesgos y errores. Además describe una metodología para investigar y exponer estos algoritmos. El proceso de auditar algoritmos y sistemas de ciencia de datos se describe en la Sección 2.3.

- **Imaginar**: Imaginar el objetivo final no como justicia, sino como co-liberación. Esto implica que, aunque abordar los sesgos en un algoritmo es valioso, no es suficiente. No se trata solo de realizar auditorías retroactivas, sino de diseñar con la meta de la co-liberación. La co-liberación requiere un compromiso y una creencia en el beneficio mutuo tanto para los miembros de grupos dominantes como para los de grupos minoritarios.
- **Enseñar**: Enseñar e involucrar a las generaciones futuras de científicos de datos, transformando el sistema educativo.

Dos ejemplos de proyectos que aplican este principio son el Observatorio para la Igualdad de Género [18] en Uruguay y el Sistema de Indicadores con perspectiva de género de la provincia de Buenos Aires [19].

3. **Valorar la emoción y corporalidad**: Habla sobre reconocer el poder de las emociones humanas en la visualización de los datos. Cuestiona las visualizaciones de datos, si verdaderamente el minimalismo visual es más neutral. La visualización de datos puede parecer objetiva, neutral, pero muchas veces no lo es y está sujeta al contexto histórico y social de quien presenta la visualización. Además argumenta que incluir la emoción a la hora de comunicarnos con datos ayuda a aprender y generar un pensamiento crítico sobre lo que se presenta.

4. **Replantear binarios y jerarquías**: Problematiza los sistemas de representación y clasificación de datos. Cuando un conjunto de datos busca caracterizar a una persona, los atributos individuales se simplifican en variables que no logran capturar la complejidad del ser humano, resultando en una representación reducida y, en muchos casos, inexacta. Se argumenta que la forma en que las personas son representadas en estos sistemas influye directamente en su nivel de visibilidad (o invisibilidad) dentro de ellos. Replantear los binarios y jerarquías requiere pensar en la forma en la cual se clasifican las personas dentro de un conjunto de datos.

   Un ejemplo claro es la clasificación de género y sexo de las personas, que refleja construcciones sociales sujetas a variaciones contextuales e históricas. Muchas veces en la ciencia de datos se utilizan los términos género y sexo indistintamente como si fuesen sinónimos. La elección de uno, otro, ambos o ninguno de estos atributos debe depender del problema que se esté abordando. Judith Butler [10], pionera en este ámbito, argumentó que el género no es binario ni mutuamente excluyente. Se suele reducir la elección de género o sexo a femenino o masculino, dejando afuera la posibilidad de identificarse con otro atributo. Las experiencias de las minorías a menudo son desplazadas a los márgenes del análisis o excluidas por completo. En particular, en Uruguay, una encuesta realizada en el año 2022 [20] reveló que las personas que se identifican como no binarias encuentran problemático expresar su identidad de género en ámbitos como la vía pública, el trabajo, los centros educativos o incluso dentro de sus propias familias . Esta reflexión implica cuestionar la infraestructura subyacente a estas clasificaciones, así como considerar quién realiza este proceso de clasificación y bajo qué intereses.

   Por otra parte, es crucial considerar que al contabilizar y clasificar, se hacen visibles ciertos atributos asociados a un grupo de personas, lo cual puede presentar desventajas o beneficios para dichos grupos. Por ello, es imperativo asumir la responsabilidad inherente al acto de contabilizar y clasificar, dado que esto puede reforzar desigualdades existentes o, en contrapartida, recuperar datos históricamente poco representados. Este proceso, conlleva un ejercicio de poder que debe ser manejado con una conciencia crítica y un compromiso ético.

5. **Abrazar el pluralismo**: Abrazar el pluralismo implica valorar y respetar la diversidad de perspectivas y voces a lo largo de todo el proceso de trabajo con datos. Este enfoque reconoce la importancia de integrar la diversidad de información, reconociendo y analizando datos atípicos, e interpretando

su existencia de manera crítica. Se debe ser responsables en el proceso de filtrado de datos para no caer en un conjunto de datos poco representativos.

El riesgo de caer en la denominada "violencia espistémica" [48] es considerable; este concepto se refiere a cómo los conocimientos y perspectivas del individuo que realiza el trabajo pueden condicionar la interpretación del mundo. Por ello, es fundamental especificar en qué contexto son realizados los proyectos, identificando los datos que son ignorados y explicando las razones detrás de estas decisiones. Abrazar el pluralismo es ir en contra de esta violencia, ser más reflexivos sobre el conjunto de datos que se tiene. Es también importante detallar las decisiones que generaron desacuerdo entre los miembros del equipo, así como las hipótesis iniciales y las que se confirmaron al final del proyecto. Fomenta la colaboración entre comunidades y expertos en datos, lo que contribuye a obtener una visión más transparente respecto a la que se obtendría de un trabajo individual.

6. **Considerar el contexto**: Afirma que a la hora de enfrentarse a una fuente de conocimiento hay que hacerse preguntas sobre las condiciones históricas, sociales, culturales, institucionales que la rodean. En lugar de ver los conjuntos de datos como datos *crudos* que pueden simplemente incorporarse a un análisis estadístico o una visualización de datos, un enfoque feminista enfatiza la necesidad de conectar los datos con su contexto de origen, comprendiendo tanto sus limitaciones como su validez. Existen métodos para proporcionar y analizar el contexto. La herramienta *datasheets for datasets* [31], es una solución propuesta para documentar la motivación detrás de la creación, composición y procesamiento de un conjunto de datos, respondiendo preguntas clave. Para los creadores del conjunto de datos, el objetivo es fomentar una reflexión cuidadosa sobre el proceso de creación, distribución y mantenimiento. Para los usuarios, el principal propósito es asegurar que dispongan de la información necesaria para tomar decisiones informadas sobre su uso.
7. **Hacer visible el trabajo**: La visibilidad del trabajo es esencial para garantizar que el trabajo subvalorado e invisible reciba el reconocimiento que merece. También es fundamental para entender el verdadero costo y las consecuencias ambientales y humanas del trabajo con datos. En la ciencia de datos, muchas tareas permanecen invisibles. A menudo citamos la fuente del conjunto de datos y mencionamos a las personas que diseñaron e implementaron el código, pero rara vez se profundiza en quién creó, recopiló y procesó los datos originalmente. Dentro de la jerarquía de trabajos en ciencia de datos, tareas como el ingreso y procesamiento de datos, a menudo llevadas a cabo por personas que realizan un trabajo subpago y subvalorado, suelen quedar invisibles.

El feminismo de datos ofrece una perspectiva necesaria para la ciencia de datos. Centra la atención en cómo las prácticas estándar pueden perpetuar desigualdades y cómo los datos pueden usarse para desafiar las estructuras de poder.

2.3. Equidad en Ciencia de Datos

Uno de los conceptos más importantes del presente estudio es la definición de *equidad*. No existe un consenso sobre el concepto en el campo de la ciencia de datos y la inteligencia artificial. Según Ghosh, Genuit y Reagan (2021) [34], es *la ausencia de prejuicio o preferencia para un individuo o grupo en función de sus características*. Según Skirpan y Gorelick (2017) [55], *un sistema solo puede ser justo con una justificación contextual de la elección de la definición de equidad y ofreciendo un medio para las partes afectadas para activamente aceptar o no aceptar la equidad del sistema*. A la hora de preguntarse si un sistema en específico cumple con las definiciones de *equidad*, corresponde cuestionarse si el sistema sigue un enfoque equitativo para resolver el problema en particular. En el artículo *Model Cards for Model Reporting* [47], se propone una metodología para que los sistemas sean acompañados de documentación que detalle sus características de rendimiento e información sobre su uso previsto y autores.

Un sistema de ciencia de datos o inteligencia artificial puede generar resultados injustos por diversas razones. Esto puede deberse a que los datos empleados a menudo reflejan sesgos sociales preexistentes, a

la falta de representatividad de ciertos grupos o a la forma en que se aplica el sistema. Como se menciona en el principio *Considerar el contexto* de la Sección 2.2, la correlación sin contexto es insuficiente y puede perpetuar los mismos sesgos presentes en las circunstancias sociales, políticas e históricas que dieron lugar a los datos. En muchas ocasiones, la causa exacta puede ser difícil de determinar debido a la complejidad y la interacción de múltiples factores, y en algunos casos, pueden ser varias causas actuando simultáneamente. Por esta razón, la definición y evaluación de la equidad en estos sistemas frecuentemente se centra en medir los potenciales daños más que en identificar las causas subyacentes. Kate Crawford [14] identifica dos tipos de daños asociados:

- **Daño de asignación**: Ocurre cuando el sistema favorece o desfavorece la asignación de oportunidades, recursos, información a ciertos individuos o grupos. Por ejemplo, en procesos de contratación o en la aprobación de préstamos, un sistema podría sesgar sus decisiones basándose en características como el género, la raza o la orientación sexual.
- **Daño de representación**: Este tipo de daño surge cuando un sistema refleja o refuerza estereotipos, o cuando no logra reconocer o representar adecuadamente a ciertos grupos. Por ejemplo, en el procesamiento del lenguaje natural, un sistema podría tener dificultades para reconocer o respetar el género en la generación de textos, o en sistemas de reconocimiento facial, podría tener dificultades para identificar con precisión a individuos con tonos de piel específicos.

Como mencionan Jacobs y Wallach [41], dentro del concepto de *equidad*, se puede hacer una primera clasificación entre lo individual y lo grupal:

- **Equidad individual**: Se busca que individuos que tienen características similares sean tratados de manera similar.
- **Equidad grupal**: Diferentes grupos de personas definidos con distintos factores demográficos son tratados de manera similar.

El concepto de *equidad grupal* nos lleva a considerar qué grupos de individuos tienen un mayor riesgo de experimentar daños por parte del sistema. Para realizar una evaluación de equidad algorítmica en un sistema con un enfoque grupal, es crucial seguir un enfoque sistemático:

1. Identificar los posibles daños que el sistema podría causar.
2. Identificar grupos afectados.
3. Definir métricas adecuadas para realizar la evaluación de equidad.
4. Aplicar las métricas a los datos.

En las siguientes subsecciones, se hablará más en detalle de estos puntos, proporcionando un análisis de cómo identificar y mitigar los posibles daños, así como evaluar la equidad algorítmica en sistemas de ciencia de datos e inteligencia artificial.

2.3.1. Grupos Afectados

Los *grupos afectados* o *grupos protegidos* deben ser considerados con atención al realizar la evaluación de equidad. En el trabajo de Beltramelli et al [37] se definen como *"aquellas poblaciones que comparten una característica por la que podrían eventualmente sufrir las consecuencias negativas de los sesgos si los hay. Ejemplos de grupos protegidos pueden ser las mujeres en la variable género, las personas no blancas en la variable raza, etc."*. En estos ejemplos, las variables mencionadas son denominadas *atributos sensibles*. La determinación de los grupos afectados debe surgir de un análisis profundo del contexto y del impacto potencial del sistema en diversas poblaciones. En una investigación [43] sobre diversos conjuntos de datos empleados para estudios de equidad en sistemas de inteligencia artificial, se identificaron los atributos sensibles utilizados para evaluar la equidad y su relación con la variable a predecir. Los conjuntos de datos analizados se organizaron según su dominio de aplicación (criminología, finanzas, salud y educación). Una de las principales conclusiones del estudio es que los atributos protegidos más utilizados en los estudios de

equidad revisados son el género, la raza, la edad y el estado civil. Aunque ninguno de estos estudios se llevó a cabo en Latinoamérica, algunos atributos sensibles relevantes en esta región podrían incluir la ruralidad o la inmigración.

Es importante notar que no implica que estos atributos no deban ser utilizados para hacer predicciones. Muchas veces, estos atributos están implícitos en otras variables, lo que facilita la reconstrucción del atributo incluso si se elimina del conjunto de entrenamiento. Aunque no se utilice el atributo sensible directamente (por ejemplo, por razones legales), otros atributos pueden actuar como *proxy* del mismo. Un ejemplo de esto es la herramienta *Gender Guesser* [32], que infiere el género de una persona a partir de su nombre. Si un sistema no utiliza el género para realizar una predicción, pero sí emplea el nombre, el género puede ser inferido indirectamente. Sin embargo, el nombre no siempre refleja con precisión el género de una persona.

En general, estos estudios tienden a analizar un atributo a la vez de forma independiente, no considerando la interseccionalidad. El concepto de *interseccionalidad* explica cómo distintas formas de discriminación (género, raza, etnicidad, estatus socio-económico, etc.) se intersectan de manera compleja y única. Kimberle Crenshaw definió este término por primera vez en 1989 [15]. Para ilustrar las dificultades del concepto, ejemplificó con tres distintos casos judiciales, incluido el caso *DeGraffenreid vs General Motors* [21]: Emma DeGraffenreid, una madre trabajadora afro-descendiente, demandó a General Motors por discriminación al no ser contratada en una fábrica de la compañía. Aunque la fábrica había contratado a hombres afro-descendientes para trabajos industriales y a mujeres blancas para trabajos de secretaría, no contrataba a mujeres afro-descendientes para ningún rol, lo que evidenciaba una discriminación basada en la intersección de género y raza juntos. El tribunal desestimó el caso, argumentando que la empresa no discriminaba por raza o género individualmente. Esto llevó a la pregunta de Crenshaw sobre la discriminación basada en raza y género juntas, una experiencia interseccional que va más allá de la suma.

La definición se extiende a los algoritmos de ciencia de datos. A la hora de detectar y mitigar sesgos, los atributos sensibles no deben ser considerados individualmente, sino abordados y analizados de forma simultánea. Un estudio relevante en este contexto es el realizado por Joy Buolamwini y Timnit Gebru [9], quienes investigaron sesgos en tres modelos comerciales de algoritmos automatizados de análisis facial. Los resultados revelaron notables disparidades de precisión, destacando la mayor diferencia de exactitud en la clasificación entre hombres de tez clara y mujeres de tez oscura. Es esencial reconocer que la interseccionalidad no solo es relevante en el ámbito jurídico y social, sino que también juega un papel crucial en el desarrollo y la implementación de algoritmos de ciencia de datos. Al abordar los sesgos y la equidad en los sistemas, es fundamental considerar cómo diferentes formas de discriminación se entrelazan y afectan a las personas.

Sobre la elección de los grupos afectados, Ghosh, Genuit y Reagan [34] exponen un desafío relacionado con la fragmentación de los conjuntos de datos en una gran cantidad de subgrupos. Esta división puede resultar en subgrupos muy pequeños, lo que complica el análisis y afecta la validez de las conclusiones. En el estudio de Beltramelli et al. [37] también se menciona esta problemática, donde los autores destacan la importancia de que la selección de los atributos que componen los grupos afectados se realice de manera crítica, problematizando cómo fueron escogidos.

2.3.2. Tipos de Sesgos en un Algoritmo

En el estudio de Beltramelli et al. [37] se define *sesgo* como *"el perjuicio sistemático que las salidas de estos sistemas de decisión basados en datos producen sobre ciertas poblaciones en comparación con otras"*. En esta sección, se mencionarán algunos tipos de sesgos que pueden introducirse en algoritmos de inteligencia artificial. Es importante tener en cuenta que esta lista no es exhaustiva y que existen otros tipos de sesgos no mencionados. Los sesgos en los algoritmos pueden surgir de diversas formas y tener diferentes impactos en los resultados. Algunos tipos de sesgos mencionados por Suresh y Guttag [56] incluyen:

- **Sesgo de representación**: Este sesgo ocurre cuando una parte de la población está subrepresentada, lo que provoca que el modelo no generalice bien para un subconjunto de la población de uso.

- **Sesgo histórico**: Este sesgo se produce cuando los datos históricos utilizados para entrenar el modelo contienen sesgos que reflejan las desigualdades y discriminaciones pasadas. Un ejemplo ilustrativo es el caso de una herramienta de contratación que utilizó Amazon [16], que mediante inteligencia artificial asignaba puntuaciones a candidatos. La herramienta, entrenada con datos de currículums de los últimos 10 años, mostró sesgos de género. Al provenir la mayoría de estos currículums de hombres, el sistema desarrolló la preferencia por candidatos masculinos, penalizando por ejemplo, términos como "women's" y desfavoreciendo estudiantes graduadas de colegios exclusivamente femeninos.
- **Sesgo en la medida**: Este sesgo ocurre al elegir, recopilar o calcular características y etiquetas para usar en un problema de predicción. Se produce cuando las características y etiquetas registradas no reflejan con precisión la realidad que intentan representar. Un ejemplo ilustrativo es el algoritmo *Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)* [27] de Northpointe, que intenta medir el riesgo de reincidencia de infractores.

  El algoritmo asume que la reincidencia se define como *"un nuevo arresto por delito menor o delito grave dentro de dos años"*. Sin embargo, esta definición presenta limitaciones significativas, ya que un arresto no implica necesariamente culpabilidad. Puede haber arrestos que no sean crímenes reales (falsos positivos) o crímenes reales que no resulten en arrestos (falsos negativos). Este enfoque ha llevado a sesgos significativos, especialmente hacia las personas afro-descendientes, como se evidencia en el análisis realizado por ProPublica [42].
- **Sesgo de agregación**: Este sesgo ocurre cuando se aplica un modelo *"único para todos"* a datos que contienen grupos que deberían ser tratados de manera diferenciada. En un estudio sobre las admisiones a programas de posgrado en la Universidad de California, Berkeley, en 1975 [8], se observó que, al analizar las solicitudes y admisiones por género, los hombres presentaban una tasa de admisión superior a la de las mujeres. Sin embargo, un análisis más detallado, considerando los departamentos, reveló que las mujeres tendían a postularse a departamentos con tasas de admisión más bajas, mientras que los hombres solicitaban ingreso a departamentos menos competitivos con tasas de admisión más altas. En resumen, el estudio mostró que las diferencias en las tasas de admisión no se debían al género, sino a las variaciones entre departamentos y las tendencias de solicitud por género.

Los distintos tipos de sesgos que pueden afectar a los algoritmos de aprendizaje automático destacan la importancia de medir la equidad de estos sistemas. Las métricas de equidad permiten identificar cuándo un sistema no trata de manera equitativa a todos los grupos. A continuación, se presentan algunas de estas métricas que permiten evaluar la equidad en los algoritmos.

2.3.3. Métricas de Equidad

El objetivo estas métricas es evaluar la equidad grupal del sistema y detectar sesgos mediante posibles desigualdades en su rendimiento entre distintos grupos afectados. Estas métricas se calculan sobre un conjunto da datos que no haya sido utilizado para entrenar. A continuación, se presentan algunas métricas utilizadas para este propósito, siguiendo la notación extraída de [1]:

- $\mathcal{A}$ es el conjunto de grupos afectados definidos en la Sección 2.3.1, $\mathcal{A} = \{a_1, \ldots, a_p\}$.
- A es la variable aleatoria asociada a la distribución de los grupos afectados.
- h es el clasificador.
- $h(X)$ es la predicción de X del clasificador.
- Y es la etiqueta de X.

Para calcular las métricas considerando a todo un conjunto de datos, se definen *radio* y *diferencia*, propuestas por Fairlearn [29], un proyecto de código abierto que ofrece herramientas en Python para la evaluación y mitigación de sesgos en algoritmos.

2.3.3.1. Paridad Demográfica

El objetivo de *paridad demográfica* o *paridad estadística* [26] es asegurar que las predicciones del modelo sean independientes en cada grupo. En un problema binario esto significa que la tasa de selección (las predicciones positivas sobre el total) sea igual para cada grupo. Matemáticamente, para un problema de clasificación esto es equivalente a buscar que:

$$\mathbb{E}(h(X)|A = a) = \mathbb{E}(h(X)) \quad \forall a \in \mathcal{A}.$$

La métrica *paridad demográfica* puede ser utilizada para identificar posibles *daños de asignación*, definidos en la Sección 2.3. Por ejemplo, si se considera un sistema que decide si un candidato es contratado o no para un trabajo y se define el género como atributo sensible, la métrica exige que la tasa de contrataciones sea igual para todos los géneros considerados.

Para computar esta métrica en todo el conjunto de datos se definen:

- **Diferencia de paridad demográfica**: Es el valor de la diferencia absoluta entre el valor de tasa de selección más alta y más baja entre los grupos, un valor de 0 indica que no hay diferencia entre los grupos, alcanzando la paridad demográfica. *Fairlearn* computa esta diferencia de la siguiente manera:

  $$\text{máx}\,(\text{TS}) - \text{mín}\,(\text{TS})$$

  donde TS es la tasa de selección del atributo.
- **Radio de paridad demográfica** Es el cociente entre el valor de tasa de selección más alta y más baja entre los grupos. Un valor de 1 indica igualdad entre los grupos en términos de tasas de aceptación (es decir, no hay disparidad). *Fairlearn* computa el radio de la siguiente forma:

  $$\frac{\text{mín}(\text{TS})}{\text{máx}(\text{TS})}$$

  donde TS es la tasa de selección del atributo.

El problema con esta métrica, es que solo se enfoca en igualar las tasas de predicción sin considerar las etiquetas verdaderas. Esto puede ser problemático si existen razones legítimas para que haya diferencias en las tasas positivas y negativas, como cuando algunos atributos sensibles están correlacionados con la etiqueta. En estos casos, exigir igual tasa de predicción positiva en todos los grupos podría impedir que el modelo alcance su utilidad prevista.

2.3.3.2. Probabilidades Igualadas

El objetivo de *probabilidades igualadas* [38] es asegurar que el sistema funcione igualmente bien para diferentes grupos. A diferencia de la *paridad demográfica*, no solo exige que las predicciones sean independientes de la pertenencia a un grupo sensible, sino que también requiere que las tasas de falsos positivos y verdaderos positivos sean iguales para todos los grupos. Matemáticamente, para un problema de clasificación esto es equivalente a buscar que:

$$\mathbb{E}(h(X)|A = a, Y = y) = \mathbb{E}(h(X)|Y = y) \quad \forall a \in \mathcal{A}, y$$

.

Para computar esta métrica se definen:

- **Diferencia de probabilidades igualadas** calcula la diferencia entre el valor más alto y más bajo separado de los falsos positivos y verdaderos positivos, y retorna la mayor de las diferencias. Mientras más pequeña sea esta medida el modelo será equitativo para las clases dentro de ese grupo. *Fairlearn* computa esta diferencia de la siguiente manera:

  $$\text{máx}\,\{\text{máx}\,(\text{TVP}) - \text{mín}\,(\text{TVP}), \text{máx}\,(\text{TFP}) - \text{mín}\,(\text{TFP})\}$$

  siendo TVP la tasa de verdaderos positivos y TFP la tasa de falsos positivos para un atributo dado.

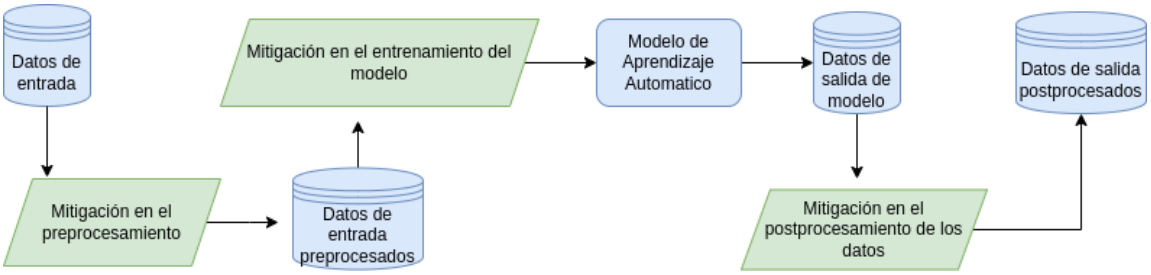


Figura 2.2: Diagrama de etapas de mitigación, basado en la publicación de AI Fairness 360 [40]

- **Radio de probabilidades igualadas** funciona de manera similar a la métrica anterior pero emplea el cociente en lugar de la diferencia, si esta métrica es cercana a 1 indica que el modelo es más equitativo. *Fairlearn* computa el radio de la siguiente forma:

$$\text{mín}\left\{\frac{\text{mín(TVP)}}{\text{máx(TVP)}}, \frac{\text{mín(TFP)}}{\text{máx(TFP)}}\right\}$$

  siendo TVP la tasa de verdaderos positivos y TFP la tasa de falsos positivos para un atributo dado.

La métrica *probabilidades igualadas* puede ser utilizada para identificar posibles *daños de asignación*, al igual que la *paridad demográfica*, y también para identificar *daños de representación*, como se presenta en la Sección 2.3. Esta métrica asegura que el sistema funcione igualmente bien (o mal) en distintos grupos, lo que implica que cada grupo tenga la misma tasa de clasificación errónea.

2.3.3.3. Igualdad de Oportunidades

La métrica *igualdad de oportunidades* [38] es una versión menos restrictiva de *probabilidades igualadas*. Solamente exige la igualdad de tasas para las etiquetas positivas $Y = 1$, considerándola como el resultado beneficioso.

2.3.4. Mitigación de Sesgos en Algoritmos

En esta sección, se exploran técnicas de mitigación de sesgos en algoritmos. El objetivo es mitigar sesgos sin comprometer las métricas de rendimiento del sistema. Es importante tener en cuenta que mitigar sesgos no significa eliminarlos por completo, puede resultar en la introducción de otros sesgos, no garantiza un resultado equitativo en todos los casos.

En la Figura 2.2 se presentan las etapas donde se puede aplicar mitigación de sesgos algorítmicos, en las siguientes subsecciones se explica cada etapa. Es recomendable emplear las técnicas de preprocesamiento siempre que sea posible modificar los datos. En caso de que sea viable modificar el algoritmo utilizado, se pueden aplicar técnicas durante el procesamiento. Sin embargo, si solo es posible modificar los datos de salida del modelo, será necesario recurrir a las técnicas de postprocesamiento.

2.3.4.1. Técnicas en el Preprocesamiento

Para mitigar sesgos en los datos de entrenamiento del algoritmo, el primer paso en el preprocesamiento es explorar los datos, visualizando la distribución, para identificar cualquier grupo subrepresentado o sobrerrepresentado. Asegurarse de que los datos sean diversos y representativos de todas las poblaciones relevantes es esencial. Esto puede implicar un análisis de las distribuciones de los datos según los grupos sensibles. Una técnica común para abordar la representación desigual de los grupos es el submuestreo de

datos de grupos sobrerrepresentados y el sobremuestreo de datos de grupos subrepresentados. Esto puede ayudar a equilibrar las distribuciones de datos e intentar que el algoritmo no esté sesgado hacia grupos mayoritarios. En algunos casos, puede ser útil generar datos artificiales para aumentar la representación de grupos minoritarios. Esto puede hacerse mediante técnicas como el aumentado de datos, que genera nuevas muestras similares a las existentes, o mediante el uso de modelos generativos como las Redes Generativas Adversarias (GANs). En particular, para los datos tabulares se puede utilizar la arquitectura CTGAN [59] (*Conditional Tabular GAN*). Una Red Generativa Adversaria (GAN) [35] es un tipo de modelo generativo que consiste en dos redes neuronales, un generador y un discriminador, que compiten entre sí. El generador crea muestras sintéticas, mientras que el discriminador intenta distinguir entre las muestras reales y las generadas. Este proceso de competencia mejora la capacidad del generador para producir muestras que son indistinguibles de las reales. En el caso de las CTGAN, la información condicional se utiliza para generar datos tabulares con ciertas características específicas. Esto es útil en aplicaciones donde se desean generar datos sintéticos que mantengan la estructura y las características de los datos reales.

2.3.4.2. Técnicas en el Entrenamiento del Modelo

Una técnica es la introducción de restricciones, donde se impone una métrica de equidad durante el entrenamiento del modelo. La restricción se aplica añadiendo un término de penalización por violaciones de equidad al objetivo de entrenamiento del modelo. Para la clasificación binaria, un enfoque es el método de *reducción*, como se describe en el trabajo de Agarwal et al. (2018) [1]. El método de reducción ajusta los pesos de clasificación para reducir las disparidades entre los grupos protegidos. Este enfoque incluye varios pasos:

1. **Inicialización**: Se comienza con un clasificador base y se define una función de pérdida que incorpora tanto la precisión del clasificador como las restricciones de equidad.
2. **Ajuste de Pesos**: Las observaciones de grupos subrepresentados o que no cumplen con las restricciones de equidad reciben un mayor peso, mientras que las observaciones de grupos sobrerrepresentados reciben un menor peso.
3. **Optimización**: Utilizando un algoritmo de optimización, se ajustan los parámetros del clasificador para minimizar la función de pérdida ajustada, que penaliza las disparidades entre los grupos.
4. **Iteración**: El proceso de ajuste de pesos y optimización se repite iterativamente hasta que el clasificador cumpla con las restricciones de equidad o se alcance un equilibrio deseado entre precisión y equidad.

De manera similar, existe un enfoque para la regresión que introduce restricciones de equidad durante el entrenamiento del modelo, como se describe en el trabajo de Agarwal et al. (2019) [2].

Otra técnica, presentada en el trabajo de Zhang, B. H., Lemoine, B., Mitchell, M. [61], es el *aprendizaje por adversario*. Esto implica la construcción de un modelo predictor y uno adversario. El modelo adversario intenta predecir los atributos sensibles a partir de las predicciones del modelo principal. El objetivo del modelo principal es engañar al adversario para que no pueda predecir la variable protegida, forzando así a que las predicciones no estén relacionadas de manera directa con los atributos sensibles.

2.3.4.3. Técnicas en el Postprocesamiento

Cuando solo se tiene acceso a los datos a la salida del sistema, es necesario realizar un procesamiento a posteriori de los resultados para que no presenten sesgos. La estrategia se centra en ajustar las etiquetas predecidas por el modelo para reducir los sesgos detectados en los resultados. La técnica de *Optimizador de umbral*, que describe Hardt, M., Price, E., Srebro, N. [38], busca encontrar el umbral que mejor se ajusta con un objetivo y una restricción impuesta. Esta técnica ajusta el umbral de decisión del modelo para minimizar las diferencias en las métricas de equidad entre los grupos. Sin embargo, una desventaja de esta técnica es que se debe conocer que instancias pertenecen a los grupos sensibles o grupos afectados para aplicarla eficazmente.

2.3.5. Desafíos y Dilemas en la Búsqueda de la Equidad en Ciencia de Datos

Hasta el momento se presentaron métodos y métricas que pretenden evaluar y mitigar algunos tipos de sesgos algorítmicos. Estos son solamente una parte de los dilemas que se enfrentan en este campo. Miceli, Posada y Yang [46] argumentan que es necesario adoptar un enfoque más amplio que el de los sesgos algorítmicos, ya que están profundamente influenciados por las dinámicas de poder dentro de las organizaciones y las estructuras sociales que intervienen en la producción de datos. Las soluciones que simplemente buscan mitigar los sesgos son insuficientes si no se abordan las relaciones de poder que configuran los sistemas de datos. Es crucial examinar cómo estas asimetrías de poder influyen en los resultados y destaca la necesidad de fomentar un diálogo interdisciplinario para producir un enfoque que considere no solo los sesgos en los datos, sino también el contexto social más amplio en el que estos se desarrollan. Estos argumentos se alinean con los principios del feminismo de datos tratados en la Sección 2.2.

Caton y Haas [12] destacan desafíos en el ámbito de la investigación. Aún existen dilemas que demandan un análisis más profundo, entre ellos el dilema del *compromiso entre el desempeño del modelo y la equidad*. En ocasiones, la reducción de sesgos en el modelo implica una disminución en su rendimiento según otras métricas, especialmente cuando el sesgo fue el factor inicial que aumentó dicho rendimiento. Otro dilema importante es el de la *incompatibilidad de las definiciones de equidad*. Las métricas de equidad suelen enfocarse ya sea en la equidad individual o en la grupal, pero no logran combinar ambas y muchas veces se contradicen como especifican Friedler S.A., Scheidegger C., Venkatasubramanian S. [30]. También se habla sobre la necesidad de tener una mayor conciencia de los aspectos sociales a la hora de buscar desarrollar algoritmos equitativos. Los investigadores deben interactuar de manera más efectiva con los involucrados y participar proactivamente en debates abiertos sobre políticas y estandarización. Aunque se observa un aumento en los grupos de trabajo sobre ética, sesgo y equidad, se destaca la urgencia de un mayor impulso a nivel nacional e internacional.

Capítulo 3

Relevamiento de Aplicaciones de Ciencia de Datos en la Educación Uruguaya

En este capítulo se registra y analiza los resultados obtenidos de un relevamiento en el que se contactó a diversos actores involucrados en proyectos relacionados con la ciencia de datos en la educación en Uruguay. El propósito de este relevamiento es centralizar la información para hacerla accesible al público, con el fin de aumentar la conciencia pública y respaldar futuras investigaciones en este ámbito. Los proyectos relevados tienen como objetivo formar un SAT que pretenden prevenir un evento no deseado en la trayectoria educativa del estudiante, como se explica en la Sección 2.1.3. Realizar este relevamiento busca aplicar los conceptos de feminismo de datos mencionados en la Sección 2.2, específicamente los principios *hacer visible el trabajo* y *abrazar el pluralismo*, teniendo en cuenta las particularidades del contexto educativo en Uruguay. Otro de los objetivos del relevamiento fue conseguir los datos necesarios para realizar un estudio de caso que permita evaluar la equidad en un sistema de predicción, siguiendo los pasos mencionados en la Sección 2.3. Para contextualizar los proyectos presentados en esta investigación, se comienza por una breve descripción del sistema educativo en Uruguay.

3.1. Sistema Educativo en Uruguay

En Uruguay, la educación formal se organiza en varias etapas: educación inicial de tres años de edad, educación inicial de cuatro y cinco años de edad, educación primaria, educación media básica, educación media superior y educación terciaria. El Ministerio de Educación y Cultura (MEC) es la entidad gubernamental responsable de diseñar y gestionar las políticas educativas, culturales y tecnológicas en Uruguay. Este ministerio opera de manera independiente, regulando tanto el sector público como el privado. En el ámbito de la educación pública, la Administración Nacional de Educación Pública (ANEP) administra el sistema educativo hasta el nivel de educación media superior. La educación inicial de tres a cinco años, así como la educación primaria, están bajo la dirección del Consejo de Educación Inicial y Primaria (CEIP), mientras que la educación media es gestionada por el Consejo de Educación Secundaria y la Dirección General de Educación Técnico Profesional – UTU. En cuanto a la educación terciaria, en el ámbito público existen instituciones como la Universidad de la República (UdelaR) y la Universidad Tecnológica. Además, hay una amplia oferta de carreras terciarias no universitarias, como las ofrecidas por la Universidad del Trabajo del Uruguay (UTU), el Instituto de Profesores Artigas y el Consejo de Formación en Educación, entre otros.

En Uruguay, son de carácter obligatorio las siguientes etapas del sistema educativo: educación inicial para niños de cuatro y cinco años, educación primaria y educación media (básica y superior) [57]. Los

proyectos analizados en este estudio se basan en planes de estudio previos a la reforma educativa de 2023, la cual introdujo cambios importantes en los programas y métodos de evaluación. A continuación se especifican los niveles educativos y métodos de evaluación previos a la reforma educativa de 2023:

La educación primaria comienza a los seis años y se divide en seis niveles. Cada estudiante progresa a través de estos niveles y al finalizar cada uno, se determina si está habilitado para avanzar al siguiente o si debe repetir el nivel. Los criterios de evaluación en esta etapa se centran en la evaluación de competencias, donde no es suficiente evaluar el conocimiento, las habilidades y las actitudes de manera aislada, sino que se requiere una evaluación integral [5]. El sistema de calificaciones utiliza una escala del 1 (indicado como Deficiente Regular) al 12 (indicado como Sobresaliente), siendo 6 (indicado como Bueno) la calificación mínima para aprobar. Además, se registra el número de ausencias del estudiante y si éste supera un límite establecido, también puede ser requerido que repita el nivel.

La educación secundaria se divide en dos ciclos: ciclo básico y bachillerato. El ciclo básico que corresponde a la educación media básica se inicia generalmente a los doce años, tras la finalización de la etapa de educación primaria y abarca tres niveles. En este ciclo las áreas de estudio se dividen en doce asignaturas. Cada asignatura se evalúa de manera independiente y las calificaciones se otorgan en una escala del 1 al 12, siendo 6 la nota mínima para aprobar. Los docentes participan en dos reuniones evaluativas conjuntas denominadas "reuniones docentes", donde se discute el desempeño de los estudiantes y se asignan las calificaciones correspondientes. Al final del curso, se realiza una reunión final para determinar el resultado definitivo de cada estudiante. Además de las calificaciones, se considera el registro de inasistencias del alumno, que puede ser motivo de repetición. Los resultados finales se clasifican en tres categorías: aprobado, cuando el estudiante demuestra un desempeño satisfactorio con una calificación igual o superior a seis; reprobado por rendimiento, si no alcanza el desempeño necesario con una calificación igual o inferior a cinco; y reprobado por inasistencias, si el estudiante excede el límite de veinticinco inasistencias.

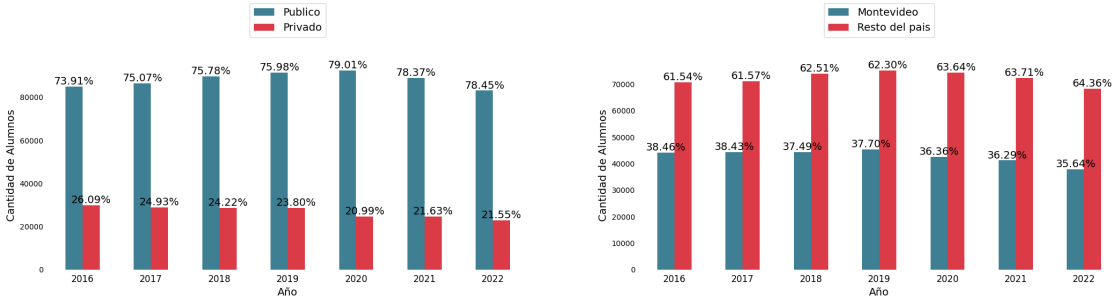
Aquellos estudiantes que reprueben en más de la mitad de las asignaturas o que excedan el límite de inasistencias se consideran repetidores, es decir, deben volver a cursar todo el nivel [17]. El ciclo de Bachillerato correspondiente a la Educación Media Superior. En este nivel los estudiantes pueden elegir una orientación académica específica, y el sistema de evaluación siguen en general los mismos parámetros que en el ciclo básico, con algunas diferencias menores. En la Educación Terciaria los métodos de evaluación varían según la institución, carrera y curso. En el caso de la UdelaR el sistema de créditos se utiliza para reflejar la dedicación requerida en cada curso. Al aprobar un curso se asignan créditos al estudiante, que son acumulados para completar el programa académico.

Para realizar una comparación con los datos utilizados en cada proyecto relevado, en las secciones posteriores se presentarán datos del Anuario Estadístico en Educación que realiza el MEC. Este documento se publica anualmente y brinda información de la distribución de estudiantes en el sistema educativo.

3.2. Relevamiento Realizado

El relevamiento se llevó a cabo entre agosto de 2023 y abril de 2024. El proceso se desarrolló de la siguiente manera:

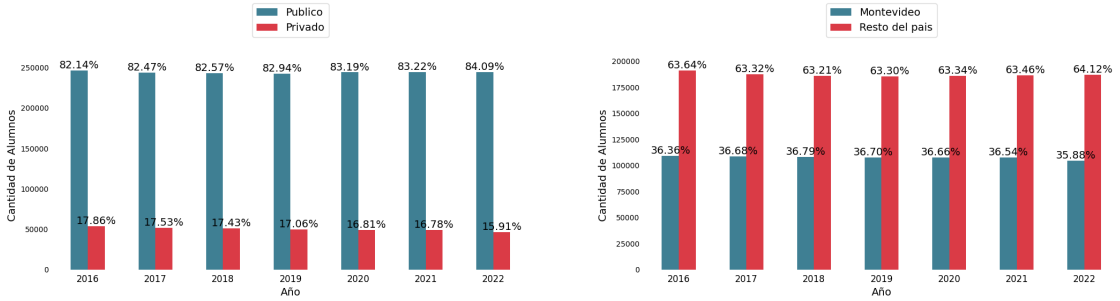
1. **Contacto**: Se identificaron y contactaron personas involucradas en proyectos relacionados con el tema de estudio.
2. **Reuniones iniciales**: Se realizaron reuniones para presentar los objetivos del proyecto y se consultó a los participantes sobre su experiencia previa en la temática o sobre proyectos específicos de interés.
3. **Recopilación de información**: Los participantes proporcionaron información sobre los proyectos. Se solicitó acceso a datos relevantes, código y modelos utilizados en los proyectos. En algunos casos no fue posible acceder a los datos.
4. **Análisis de datos**: Se llevaron a cabo análisis en profundidad de los datos proporcionados. Durante este proceso, se realizaron preguntas adicionales para clarificar detalles y obtener una comprensión más completa de los proyectos.



(a) Cantidad de estudiantes de educación inicial según el tipo de educación.

(b) Cantidad de estudiantes de educación inicial según su locación.

Figura 3.1: Distribución de estudiantes de educación inicial.



(a) Cantidad de estudiantes de educación primaria según el tipo de educación.

(b) Cantidad de estudiantes de educación primaria según su locación.

Figura 3.2: Distribución de estudiantes de primaria.

Dentro del conjunto de sistemas que se relevaron, se presentan según el nivel educativo: Educación inicial y primaria, Educación secundaria y Educación terciaria.

3.2.1. Educación Inicial y Primaria

A continuación se presentan los proyectos relevados vinculados a Educación Inicial y Primaria. Los organismos encargados de brindar los datos para estos proyectos están vinculados a ANEP y CEIP, y los proyectos tienen información sobre estudiantes de entre cinco y doce años de edad. A partir de los resultados del Anuario Estadístico de Educación mencionado en la Sección 3.1 se muestra en la Figura 3.1 la distribución de estudiantes de educación inicial y en la Figura 3.2 la de estudiantes en educación primaria desde los años 2016 a 2022. En ambos casos se diferencia entre sistemas de carácter público o privado, y además entre Montevideo y el resto del país.

3.2.1.1. Detección temprana del riesgo escolar. Predicción de trayectorias de rezago en la educación primaria en Uruguay mediante técnicas de *Machine learning* (DTER-ML)

El proyecto DTER-ML [11] tiene como objetivo identificar el riesgo escolar en las trayectorias de los estudiantes durante sus primeros tres años en la educación primaria. El riesgo escolar se define como una trayectoria académica afectada por al menos un evento de repetición. Los datos analizados corresponden a estudiantes que ingresaron al primer grado en 2017. Sin embargo, para este análisis se considera únicamente a aquellos estudiantes que realizaron la Evaluación Infantil Temprana (EIT) en 2016, lo que implica aproximadamente seis mil estudiantes. Esta cantidad de estudiantes para el año 2016 representa representa el 2,4 % de los estudiantes de educación publica (Figura 3.1a).

La EIT evalúa habilidades cognitivas que están vinculadas con la preparación para la escuela y cada estudiante es clasificado en una de tres categorías: verde, amarillo o rojo, las cuales indican el nivel de dificultad que el estudiante presentó al realizar la evaluación. Además se utiliza información proporcionada por ANEP, como los fallos y calificaciones escolares, así como la evaluación adaptativa (SEA+). También se integran datos del Ministerio de Desarrollo Social (MIDES) sobre hogares beneficiarios de los programas sociales AFAMPE y TUS, y del Ministerio de Salud Pública (MSP) sobre certificados de nacido vivo, los cuales aportan información sobre riesgos sociosanitarios asociados al período de gestación y al nacimiento.

El propósito del análisis es determinar la influencia que tiene la EIT en la estimación del desempeño de los estudiantes a lo largo de su educación primaria y determinar a partir de la misma el riesgo escolar. Este proyecto se llevo adelante entre la UdelaR y ANEP.

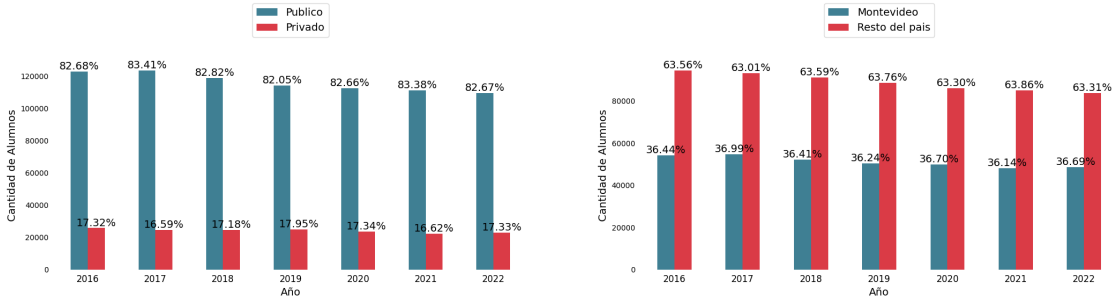
- **¿Qué busca predecir?:** La repetición del estudiante, indicada mediante una variable binaria que asume el valor 1 si el alumno experimenta al menos un evento de repetición entre 2017 y 2019.
- **Población analizada:** Estudiantes de primer año de escuelas de todo el territorio uruguayo en 2017.
- **Cantidad de estudiantes:** 6000.
- **Atributos del conjunto estudiantes utilizados:** Condiciones de gestación, contexto socioeconómico de las familias, información sociodemográfica de las escuelas, información sobre aprobación, repetición y desvinculación de estudiantes, resultados de la EIT de 2016.
- **Técnicas utilizadas:** Se entrenan tres modelos distintos, Regresión Logística, Redes Bayesianas y árbol de clasificación.
- **Resultados:** Modelo predictivo con porcentaje de acierto de 80 % y una sensibilidad de 62 % a 64 %. El atributo que más peso tiene en la predicción son los resultados de la EIT.

3.2.1.2. Lexiland (LXLD)

El proyecto Lexiland [62] [63], realizado por la UdelaR y el Centro Interdisciplinario en Cognición para la Enseñanza y el Aprendizaje (CICEA), examina el desarrollo y la contribución de la conciencia fonológica a las habilidades de lectura temprana en español. El estudio comenzó con niños en su último año de jardín de infantes y continuó siguiéndolos durante su primer y segundo año de escuela primaria. Los niños fueron reclutados de veintiséis escuelas públicas en Montevideo, Uruguay. En lo que respecta a la población estudiada, representa sólo el 1,4 % de la cantidad de estudiantes de educación primaria de Montevideo según datos del Anuario Estadístico de Educación(Figura 3.1b).

Para evaluar la conciencia fonológica en los niños, se utilizó una aplicación para *tablets* llamada Lexiland con formato de videojuego, diseñada específicamente para el estudio. Los niños realizan tareas que miden distintos indicadores de habilidades de lectura a través de la aplicación. Además se obtuvieron datos demográficos y socioeconómicos de ANEP, incluyendo la edad, el género y el nivel máximo de educación de la madre de cada niño. En base a la información obtenida, se realizó un análisis y se entrenó un modelo de regresión lineal para predecir si el niño presentará dificultades en la lectura en uno o dos años posteriores.

- **¿Qué busca predecir?:** Signos tempranos de dificultad en la lectura. Es una variable binaria que distingue entre lectores típicos y lectores con dificultades en la lectura.



(a) Cantidad de estudiantes de educación media básica según el tipo de educación.

(b) Cantidad de estudiantes de educación media básica según su locación.

Figura 3.3: Distribución de estudiantes de educación media básica.

- **Población analizada:** Estudiantes de nivel 5 de escuelas públicas de Montevideo, Uruguay.
- **Cantidad de datos estudiantes:** 388.
- **Atributos del conjunto de estudiantes utilizados:** Edad, Género, Nivel educativo de la madre, Vocabulario, IQ [58], Memoria a corto plazo (verbal y no verbal), Conocimiento de las letras, Conocimiento fonológico, *Rapid Automatized Naming* (RAN).
- **Técnicas utilizadas:** Regresión Lineal.
- **Resultados:** Modelo predictivo con porcentaje de acierto de 89 % y una especificidad de 70 % para un 90 % de sensibilidad. Los atributos más importantes para la predicción fueron: Nivel educativo de la madre, Memoria a corto plazo (no verbal), Conocimiento de las letras, Conocimiento fonológico.

Para este proyecto fue posible obtener los datos y el sistema implementado, se realizó un análisis en más profundidad en el Anexo A.

3.2.2. Educación Secundaria

En el caso de Educación Secundaria se obtiene únicamente un proyecto, vinculado a la Educación Media Básica o Ciclo Básico. La importancia de este proyecto es que esta etapa es de carácter obligatorio para los estudiantes, como se menciono en la Sección 3.1. Según datos del Anuario estadístico, la distribución de estudiantes de educación Media Básica se presenta en la Figura 3.3 dividiendo según carácter público o privado, y además entre Montevideo y el resto del país.

3.2.2.1. Using Data Mining techniques to follow students trajectories in secondary schools of Uruguay (DM-STU)

Este proyecto tiene como objetivo la predicción de desvinculación en estudiantes de enseñanza media en Uruguay [44], en particular se centra en estudiantes de ciclo básico. Se utiliza un conjunto de datos que consiste en 13500 estudiantes de ciclo básico de los años 2015 a 2016. Además se busca tener una referencia de los factores de riesgo asociados a la desvinculación. El conjunto de estudiantes que se evaluaron son de educación publica, representan el 10 % de los estudiantes de esa época, según datos del anuario estadístico (Figura 3.3a).

A partir de los resultados obtenidos como conclusión se obtienen los principales factores de riesgo a la desvinculación, estos factores son: cinco ausencias injustificadas, cuatro ausencias justificadas, 50 % de calificaciones promedio menores a cinco, estudiantes que estén realizando de nuevo el año (son reprobados

del año anterior) y tienen el 50 % de las notas menores que seis en la primera reunión. Este proyecto se llevó a cabo en conjunto con la Universidad Federal de Santa Catarina (UFSC), la Universidad Federal do Rio Grande do Sul (UFRGS), la Escuela Superior Politécnica del Litoral (ESPOL), UdelaR y ANEP.

- **¿Qué busca predecir?**: Repetición en estudiantes de secundaria. Clasifica a los estudiantes según el fallo final (promovido, repite por rendimiento y repite por inasistencia).
- **Población analizada:** Estudiantes de ciclo básico de liceos dentro del territorio uruguayo en el periodo 2015-2016.
- **Cantidad de estudiantes:** 13500.
- **Atributos del conjunto de estudiantes utilizados:** Se utilizan los atributos de cantidad de materias que el estudiante cursó, edad del estudiante, porcentaje de materias en la cual el estudiante obtiene una calificación menor respecto a la primear reunión, cantidad de inasistencias no justificadas en la primera reunión, cantidad de inasistencias justificadas en la primera reunión.
- **Técnicas utilizadas:** K-Means [6] (k=3).
- **Resultados:** Porcentaje de acierto de 68,62 % en primer año, 62,22 % en segundo y 61,61 % en tercer año.

3.2.3. Educación Terciaria

A diferencia de los proyectos de secciones anteriores, la educación terciaria no es de carácter obligatorio en Uruguay. Los proyectos que se presentan a continuación son implementados sobre la educación terciaria universitaria, no es el único tipo de educación terciaria que existe en Uruguay, como se menciona en 3.1. En la Figura 3.4 se presenta la distribución de estudiantes de educación terciaria universitaria, dividido según si son de universidad pública o privada. Los proyectos relevados todos son con datos de educación terciaria universitaria pública, esta población representa históricamente una mayoría frente a la población universitaria privada.

3.2.3.1. Modelado de Trayectorias Académicas de Estudiantes Universitarios mediante Técnicas de Analítica de Aprendizaje (MTA-Ap)

El proyecto MTA-Ap [45] fue realizado en el marco del proyecto de fin de carrera de Ingeniería en Computación en la UdelaR. En el mismo se busca explicar la desvinculación estudiantil en base a los datos del sistema de bedelías. Para ello se realizó un análisis de las variables que inciden en la desvinculación de los estudiantes y se desarrolló un modelo predictivo que pudiera explicar la desvinculación. Además se automatizaron procesos de realización de informes de la UEFI (Unidad de Enseñanza de Facultad de Ingeniería).

Este proyecto utilizó datos del sistema de bedelías de la carrera de Ingeniería en Computación de 1997 a 2019. En principio se realiza un análisis de los datos descriptivo del cual se extraen características vinculadas a la desvinculación estudiantil.

- **¿Qué busca predecir?:** Desvinculación en estudiantes de Ingeniería en Computación.
- **Población analizada:** Estudiantes de Ingeniería en Computación del periodo 1997 a 2019.
- **Cantidad de estudiantes:** 5644.
- **Atributos del conjunto de estudiantes utilizados:** Edad del estudiante al ingreso (en meses al primero de marzo del año de la generación), estrato social a partir de la dirección de vivienda del estudiante, subsistema de egreso preuniversitario, sexo del estudiante, cantidad de créditos que obtuvo ese estudiante en el primer semestre de su ingreso a la facultad.
- **Técnicas utilizadas:** Random Forest.

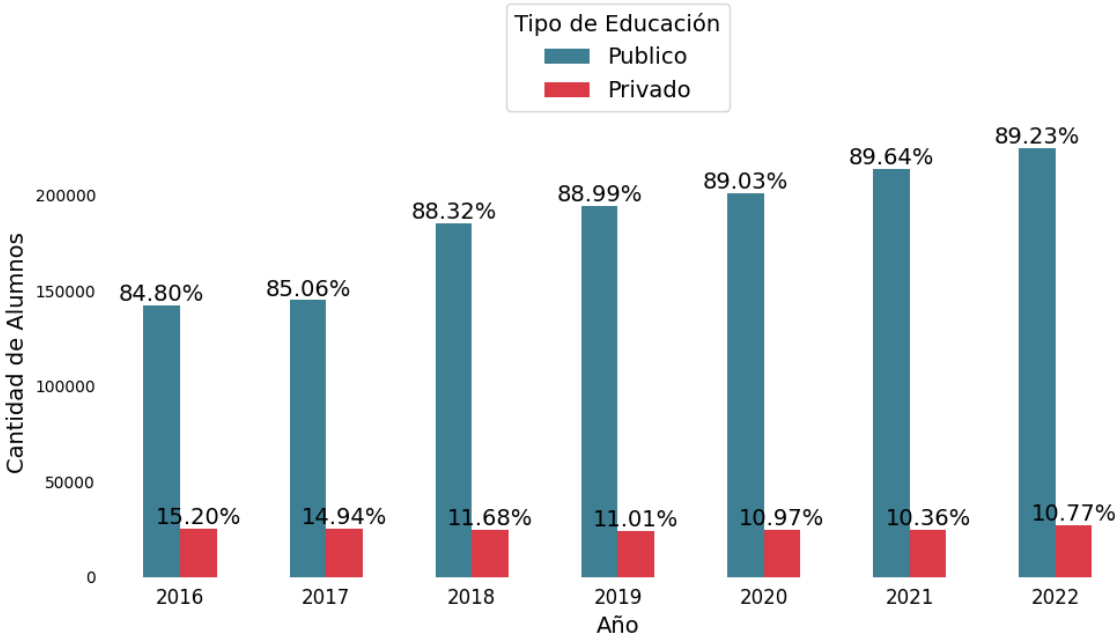


Figura 3.4: Distribución de estudiantes de educación terciaria universitaria.

- **Resultados:** Porcentaje de acierto de 86 %. Los atributos más importantes para la predicción fueron los créditos obtenidos en el primer semestre y la edad del estudiante al ingreso. En segundo lugar, el estrato social y el tipo de institución previa.

3.2.3.2. Propuesta de modelización predictiva del rendimiento académico en el corto plazo para estudiantes de Ingeniería (MPRI)

El proyecto MPRI [3] surge como tesis de Maestría en Ingeniería Matemática de la UdelaR y analiza la capacidad de varios modelos de aprendizaje automático para predecir el éxito académico (si al final del primer año tiene más del 50 % de los créditos) de estudiantes de ingeniería de primer año en la UdelaR. Se utiliza la información sociodemográfica obtenida de los registros de Bedelía de la Facultad de Ingeniería y la información académica proveniente de la Herramienta Diagnóstica al Ingreso (HDI), que realizalos estudiantes al inicio del primer año de la carrera, para entrenar y probar modelos como regresión logística, máquinas de soporte vectorial, árboles de decisión, y métodos de ensemble.

- **¿Qué busca predecir?:** Aprobación de más de la mitad de los créditos al final del primer año.
- **Población analizada:** Generación ingresante a la Facultad de Ingeniería de la UdelaR en 2016.
- **Cantidad de estudiantes:** 694.
- **Atributos del conjunto de estudiantes utilizados:** Edad al ingreso, sexo, subsistema de egreso preuniversitario, lugar del centro preuniversitario de procedencia, primer carrera a la que se inscribió, puntajes de la HDI, separados por subpruebas de matemática y comprensión lectora.
- **Técnicas utilizadas:** Regresión Logística, Support Vector Machines [13], Clasificador Bayesiano, CART, Random Forest, AdaBoost.

- **Resultados:** Modelo con una porcentaje de acierto del 80 %, 77 % de precisión. Los atributos más importantes para la predicción fueron puntaje en matemática de HDI, edad al ingreso, lugar de finalización de estudios preuniversitarios y subsistema de finalización de estudios preuniversitarios.

Para este proyecto fue posible obtener los datos y el sistema implementado, se realizó un análisis en más profundidad en el Anexo B.

3.2.3.3. Using Virtual Learning Environment Data for the Development of Institutional Educational Policies (VLE-EP)

El proyecto VLE-EP [51] presentado por el Instituto Federal do Rio Grande do Sul, UdelaR, Universidade Federal de Santa Catarina y Universidade Federal do Rio Grande do Sulen, se enfoca en la utilización de datos generados en el entorno virtual de aprendizaje (EVA) para desarrollar políticas educativas a nivel institucional en la UdelaR. Este trabajo combina técnicas de ciencia de datos y minería de datos educativos para identificar patrones de comportamiento y generar modelos predictivos tempranos que ayuden a mejorar el rendimiento académico de los estudiantes.

- **¿Qué busca predecir?:** Éxito en el curso (estudiantes que aprobaron sin necesidad de exámenes) y la predicción del éxito en los exámenes finales.
- **Población analizada:** Estudiantes de segundo año matriculados en cursos de tres facultades diferentes en el año 2017, de la siguiente manera: Facultad de Información y Comunicación, Facultad de Enfermería y Facultad de Ciencias en la Universidad de la República.
- **Cantidad de estudiantes:** 4529.
- **Atributos del conjunto de estudiantes utilizados:** Interacciones de los estudiantes en el EVA (por semana, por tipo de interacción - foro, url, cuestionarios, carpeta, etc), materias en las que el estudiante está matriculado, Rendimiento académico en las materias, número de reprobaciones previas en cada materia, respuestas de la encuesta FormA-Estudiantes de Grado (111 atributos sobre antecedentes socio-demográficos, socioeconómicos, estudios, situación laboral, idiomas, becas, etc.).
- **Técnicas utilizadas:** Random Forest, Ada Boost, Regresión Logística.
- **Resultados:** Modelo con un porcentaje de acierto del 90 % para predecir la aprobación del curso y 87 % para predecir el éxito en el examen final. Los atributos más importantes para la predicción fueron las materias matriculadas por estudiante, educación de la madre, interacciones totales, promedio, en foros y en urls en el EVA en la semana 2, semana 4, semana 6, lugar de residencia, y si el estudiante está cursando una única materia.

3.3. Reflexiones Sobre el Relevamiento

El relevamiento de sistemas realizado ha permitido identificar y analizar una variedad de enfoques y técnicas aplicadas en distintos niveles educativos en Uruguay. Cada proyecto fue analizado de manera independiente y se extrajeron las principales características para obtener una visión más completa de los trabajos realizados.

La Tabla 3.1 presenta un resumen detallado de los modelos de aprendizaje automático utilizados, junto con la estructura de los datos en cada proyecto. Se observa que el porcentaje de acierto en estos proyectos es consistentemente alto, encontrándose todos por encima del 60 %. Esto lleva a concluir que los atributos considerados por los proyectos son de utilidad para predecir y buscar indicadores sobre la repetición, desvinculación o desempeño en la educación. Sin embargo, la cantidad de datos empleada en cada proyecto es limitada siendo poco representativa de las poblaciones y difiere según cada proyecto. Esto puede generar un daño en la representación (presentado en la Sección 2.3) ya que existen grupos no representados. En general no son parte de un muestreo balanceado, por ejemplo en el proyecto LXLD los datos utilizados fueron de

veintiséis escuelas públicas de Montevideo, esto deja afuera al resto de las escuelas del país. Esta selección no fue intencional, sino que se debió a la disponibilidad y acceso de las escuelas donde se pudieron realizar las pruebas. En cuanto a los modelos empleados, la Regresión Logística y Random Forest son los más recurrentes. Se destacan por su simplicidad en la implementación y por la capacidad de proporcionar valores que reflejan la importancia de cada atributo en la toma de decisiones. Entre los atributos más influyentes identificados en los diferentes proyectos, se destacan la edad del estudiante, el nivel educativo de la madre y en el caso de los estudios terciarios, el subsistema de egreso preuniversitario (liceo privado, liceo público o UTU).

La Tabla 3.2 muestra las diversas fuentes de datos utilizadas en los proyectos. En los estudios de primaria y secundaria, se recurre en gran medida a datos de ANEP, mientras que para los estudiantes de nivel terciario, la información proviene principalmente de los sistemas de Bedelía. En Uruguay, no existe una fuente de datos unificada que englobe toda la información relacionada con la educación. Como resultado, los proyectos deben recurrir a múltiples fuentes de datos, las cuales varían según el alcance y los objetivos de cada estudio, y generalmente requieren permisos de autoridades para su acceso. Cada institución maneja y almacena sus propios datos, sin que haya un intercambio fluido de información entre ellas. Esta distribución fragmentada de los datos educativos genera dificultades en su acceso y utilización, complicando la realización de estudios integrales sobre el sistema educativo uruguayo. Esta característica forma parte de las particularidades de los datos educativos mencionados en la Sección 2.1.3.

Proyecto	**¿Qué busca predecir?**	**Cantidad de estudiantes**	**Modelos Utilizados**	**Atributos más influyentes**	**Porcentaje de acierto (%)**
3.2.1.1 DTER-ML	Repetición	6000	Regresión Logística, Clasificador Bayesiano y Árbol de clasificación	Resultados de la EIT	80
3.2.1.2 LXLD	Dificultad en la lectura	388	Regresión Logística	Nivel educativo de la madre, Memoria a corto plazo, conocimiento de las letras y conocimiento fonológico	89
3.2.2.1 DM-STU	Repetición	13500	K-Means	Cantidad de faltas injustificadas, Edad del estudiante, Calificaciones en la primera reunión.	69
3.2.3.1 MTA-Ap	Desvinculación	5644	Random Forest	Créditos obtenidos en primer semestre, edad al ingreso, estrato social y tipo de institución previa	86
3.2.3.2 MPRI	Aprobación de más de la mitad de los créditos	694	Regresión logística, SVM, Clasificador Bayesiano, Árbol de clasificación, Random Forest, AdaBoost	Puntaje en Matemática en HDI, Edad al ingreso, tipo de institución previa	80
3.2.3.3 VLE-EP	Éxito en el curso	4529	Random Forest, AdaBoost, Regresión Logística	Cantidad de materias matriculadas por estudiante, Educación de la madre, Interacciones con EVA	90

Tabla 3.1: Estructura de datos y modelo de cada proyecto.

Cabe destacar que todos los proyectos relevados están basados en la educación pública y fueron llevados a cabo por instituciones públicas. Otra característica a destacar de los proyectos relevados es que todos permanecen en una etapa de estudio y no fueron puestos en producción. Además, teniendo en cuenta el principio de *Abrazar el pluralismo* mencionado en la Sección 2.2, algunos de estos proyectos se realizaron de manera interdisciplinaria, con la colaboración de distintas áreas del conocimiento. Otros se llevaron a

Proyecto	Institución que realiza el proyecto	Características de estudiantes	Fuente de los datos
3.2.1.1 DTER-ML	UdelaR, ANEP-CODICEN	Estudiantes de primaria	EIT, ANEP, MSP, MIDES
3.2.1.2 LXLD	UdelaR, CICEA	Estudiantes de primaria	Recolección propia y ANEP
3.2.2.1 MATA-Ap	UdelaR, ANEP, Universidade Federal de Santa Catarina, Universidade Federal do Rio Grande do Sul, Escuela Superior Politécnica del Litoral	Estudiantes de ciclo básico	ANEP
3.2.3.1 DM-STU	UdelaR	Estudiantes de Ingeniería en Computación	Sistema de Bedelía
3.2.3.2 MPRI	UdelaR	Estudiantes de Facultad de Ingeniería	Sistema de Bedelía y HDI
3.2.3.3 VLE-EP	Instituto Federal do Rio Grande do Sul, UdelaR, Universidade Federal de Santa Catarina, Universidade Federal do Rio Grande do Sul	Estudiantes de segundo año de nivel terciario	EVA y encuesta FormA-Estudiantes de Grado

Tabla 3.2: Contexto sobre los datos utilizados en proyectos.

cabo en colaboración con instituciones de Brasil, como es el caso de los proyectos DM-STU y VLE-EP. En DM-STU también se encuentra la participación de investigadores de Ecuador. Esto permite enriquecer el análisis y abordar los desafíos educativos desde múltiples perspectivas. A su vez no son consideradas en el análisis las instituciones privadas.

Otra de las características que comparten los proyectos relevados es que utilizan en su mayoría la categorización de Femenino y Masculino para referirse a género o sexo. En el proyecto DTER-ML se refiere a niñas y varones, en LXLD utilizan la distinción de género masculino y femenino, en MTA-Ap género y sexo se tratan como el mismo atributo, con las categorías femenino y masculino. En los demás proyectos, no se realiza ninguna mención específica a género o sexo. Como se mencionó en el principio de *Replantear binarios y jerarquías* 2.2, que los individuos sean encasillados dentro de estas dos categorías no es representativo de la población ya que existe un conjunto más amplio de identidades de género reconocidas. La ausencia de otras identidades de género en los conjuntos de datos podría deberse a que no hubo casos registrados en los datos utilizados o a que simplemente no fueron consideradas en el análisis.

Se obtuvo acceso a los sistemas de predicción y datos para los proyectos LXLD y MPRI. No fue posible

realizar el estudio de caso planteado con estos proyectos debido a la insuficiencia de datos disponibles. La falta de un conjunto de datos independiente no utilizado en el entrenamiento de los modelos no permitió realizar una evaluación de equidad más allá de un análisis de datos de entrada.

Realizar este relevamiento fue fundamental para comprender el panorama actual de la aplicación de la ciencia de datos en la educación en Uruguay. Al identificar los modelos más utilizados, los problemas que intentan resolver, las instituciones que los llevan a cabo, las fuentes de datos disponibles y las limitaciones presentes, se obtiene una visión clara de las áreas en las que se han logrado avances y de los desafíos que aún persisten.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 4

Estudio de Caso: Desvinculación Académica

El objetivo de este capítulo es aplicar el principio de *desafiar el poder* mencionado en la Sección 2.2, mediante el análisis de un algoritmo de predicción, así cómo la familiarización con las herramientas de análisis y mitigación de sesgos algorítmicos. Originalmente, se buscaba seleccionar un proyecto de Uruguay basado en el relevamiento realizado en el Capítulo 3. Sin embargo, debido a la falta de un proyecto adecuado que cumpliera con los requisitos para este análisis en términos de disponibilidad y cantidad de datos, se decidió trabajar con el proyecto *"Student Dropout Analysis for School Education"* [53] ya que fue posible reproducir el sistema inicial. El estudio está enfocado en reducir la desvinculación académica en la educación superior del Instituto Politécnico de Portalegre, Portugal. El objetivo es predecir el estado del estudiante, clasificándolo como "Desvinculado", "Matriculado" y "Egresado", según su información académica. Para simplificar el análisis y la reproducción del modelo, se transformó el problema en una clasificación binaria, considerando únicamente las instancias de estudiantes "Desvinculado" o "Egresado", excluyendo a aquellos que aún continúan matriculados. El estudio incluye una descripción y análisis de los datos, una evaluación de sesgos algorítmicos, y la implementación de técnicas para mitigar los sesgos algorítmicos identificados.

4.1. Conjunto de Datos

Siguiendo la práctica *datasheets for datasets* descrita en el principio *Considerar el contexto* mencionado en la Sección 2.2, se responden las siguientes preguntas para comprender mejor el contexto del conjunto de datos. La información que responde a estas preguntas es extraída de un informe sobre las características del conjunto de datos [52].

- **¿Quién creó el conjunto de datos y en nombre de qué entidad?** El conjunto de datos fue creado por el Instituto Politécnico de Portalegre, Portugal, específicamente por un equipo de docentes e investigadores.
- **¿Quién fundó la creación del conjunto de datos?** Este conjunto de datos se creó bajo el programa SATDAP - Capacitação da Administração Pública bajo la subvención POCI-05-5762-FSE-000191, Portugal.
- **Qué representan las instancias que componen el conjunto de datos?** Los datos representan a estudiantes y sus registros académicos, incluyendo información demográfica, socioeconómica, macroeconómica, y el rendimiento académico al final del primer y segundo semestre. Estos datos fueron recolectados de estudiantes entre 2008 y 2019, incluyendo diecisiete orientaciones distintas (por ejemplo agronomía, diseño, enfermería, entre otros). Los datos se extraen de fuentes que son independientes y posteriormente se combinan en un único conjunto de datos. La información demográfica y socioeconómica se obtiene del Sistema de Gestión Académica. La información sobre el estudiante al

ingresar al instituto se obtienen del Concurso Nacional de Acceso a la Educación Superior, mientras que los datos del rendimiento académico se obtienen de la Dirección General de Educación Superior. Finalmente los datos macroeconómicos se obtienen de la base de datos de Portugal.

- **¿El conjunto contiene datos que podrían considerarse confidenciales?** Se menciona que los datos fueron anonimizados, asegurando el cumplimiento del Reglamento General de Protección de Datos [28].
- **¿Hay una etiqueta u objetivo asociado con cada instancia?** El objetivo asociado con cada instancia es la clasificación del estudiante en una de las dos categorías: "Desvinculado" o "Egresado". Para clasificar a un estudiante como "Egresado", se consideraron aquellos que completaron sus estudios dentro de la duración estándar del curso (tres años, excepto en el caso de Enfermería, donde la duración es de cuatro años). Los estudiantes que no se desvincularon ni egresaron dentro del plazo establecido se consideran aún matriculados, y no fueron incluidos en este estudio.
- **¿Se realizó algún preprocesamiento, limpieza o etiquetado de los datos?** En la documentación se menciona que se realizó un riguroso preprocesamiento de datos para manejar anomalías, valores atípicos y valores faltantes. Además, para realizar este estudio se utilizaron 27 de los 36 atributos del conjunto de datos.
- **¿Cuántas instancias hay en total?** Hay 3630 instancias. Se dividieron los datos en el conjunto de entrenamiento (80 %) y en el de prueba (20 %).

4.1.1. Atributos

En esta sección se presenta el conjunto de datos con el que se trabajará. El conjunto de datos utilizado en este proyecto contiene 3630 instancias que contienen atributos con información sobre estudiantes. Además se asigna el atributo *Etiqueta* el cual indica si el estudiante egresó o se desvinculó (-1 si se desvinculó, 1 si egresó). Los atributos se dividen entre información demográfica, socioeconómica y macroeconómica y académica.

- **Información demográfica**:
  - **Estado civil**: Estado civil del estudiante (1 - soltero, 2 - casado, 3 - viudo, 4 - divorciado, 5 - unión de hecho, 6 - separado legalmente).
  - **Género**: Género del estudiante (1 - género masculino, 0 - género femenino).
  - **Desplazado**: Si el estudiante es una persona desplazada (1 - sí, 0 - no).
  - **Edad**: Edad del estudiante en el momento de la inscripción.
- **Información socioeconómica**:
  - **Ocupación de la madre**: Ocupación de la madre del estudiante. (0 - Estudiante, 1 - Representantes del Poder Legislativo y Órganos Ejecutivos, Directores y Gerentes Ejecutivos, 2 - Especialistas en Actividades Intelectuales y Científicas, 3 - Técnicos y Profesionales de Nivel Intermedio, 4 - Personal Administrativo, 5 - Trabajadores de Servicios Personales, Seguridad y Vigilancia, y Vendedores, 6 - Agricultores y Trabajadores Calificados en Agricultura, Pesca y Silvicultura, 7 - Trabajadores Calificados en la Industria, Construcción y Artesanía, 8 - Operadores de Instalaciones y Máquinas y Trabajadores de Ensamblaje, 9 - Trabajadores No Calificados, 10 - Profesiones de las Fuerzas Armadas, etc.).
  - **Ocupación del padre**: Ocupación del padre del estudiante (mismas que la madre).
  - **Deudor**: Si el estudiante es deudor (1 - sí, 0 - no).
  - **Matrícula al día**: Si las tasas de matrícula del estudiante están al día (1 - sí, 0 - no).
  - **Becado** : Si el estudiante es titular de una beca (1 - sí, 0 - no).

- **Información macroeconómica**:
  - **GDP**: Producto interno bruto de la región (número entre −4,100 y 3,500).
- **Información académica**:
  - **Información al momento de inscripción**:
    - **Modalidad de aplicación**: Modalidad de aplicación utilizada por el estudiante. Los valores numéricos representan diferentes categorías, como 1 - aplicación en la primera fase general o 2 - especial, 15 - la aplicación internacional, entre otras modalidades específicas según la normativa correspondiente.
    - **Orden de aplicación**: Orden en el que el estudiante aplicó. Números entre 0 - primera opción y 9 - última opción.
    - **Curso**: Identificador del curso tomado por el estudiante. (1 - Tecnologías de Producción de Biocombustibles, 2 - Animación y Diseño Multimedia, 3 - Servicio Social (turno vespertino), 4 - Agronomía, 5 - Diseño de Comunicación, 6 - Enfermería Veterinaria, 7 - Ingeniería Informática, 8 - Equinocultura, 9 - Gestión, 10 - Servicio Social, 11 - Turismo, 12 - Enfermería, 13 - Higiene Bucal, 14 - Gestión Publicitaria y de Marketing, 15 - Periodismo y Comunicación, 16 - Educación Básica, 17 - Gestión (turno vespertino)).
    - **Asistencia diurna/vespertina**: Si el estudiante asiste a clases durante el día o por la noche (1 - diurno, 0 - vespertino).
    - **Calificación previa**: Calificación obtenida por el estudiante antes de inscribirse en educación superior (1 - Educación secundaria, 2 - Licenciatura, 3 - Grado, 4 - Maestría, 5 - Doctorado, etc.).
  - **Información del primer semestre**:
    - **Unidades curriculares 1er semestre (acreditadas)**: Número de unidades curriculares acreditadas en el primer semestre del estudiante.
    - **Unidades curriculares 1er semestre (matriculadas)**: Número de unidades curriculares en las que se inscribió el estudiante en el primer semestre.
    - **Unidades curriculares 1er semestre (evaluaciones)**: Número de evaluaciones de unidades curriculares en el 1er semestre tomadas por el estudiante.
    - **Unidades curriculares 1er semestre (aprobadas)**: Número de unidades curriculares aprobadas por el estudiante en el primer semestre.
    - **Unidades curriculares 1er semestre (nota)**: Promedio de calificaciones en el primer semestre (entre 0 y 20).
  - **Información del segundo semestre**:
    - **Unidades curriculares 2do semestre (acreditadas)**: Número de unidades curriculares acreditadas en el segundo semestre del estudiante.
    - **Unidades curriculares 2do semestre (matriculadas)**: Número de unidades curriculares en las que se inscribió el estudiante en el segundo semestre.
    - **Unidades curriculares 2do semestre (evaluaciones)**: Número de evaluaciones de unidades curriculares en el 2do semestre tomadas por el estudiante.
    - **Unidades curriculares 2do semestre (aprobadas)**: Número de unidades curriculares aprobadas por el estudiante en el segundo semestre.
    - **Unidades curriculares 2do semestre (nota)**: Promedio de calificaciones en el segundo semestre (entre 0 y 20).
    - **Unidades curriculares 2do semestre (sin evaluaciones)**: Número de unidades curriculares tomadas en el segundo semestre sin evaluaciones.

4.2. Análisis de los Datos

En esta Sección se desglosa un análisis de los atributos del conjunto de datos de entrada y su vinculación. En la Figura 4.1 se muestran las correlaciones entre los atributos. En la última fila se puede ver la correlación con la etiqueta (1: Egresado, -1: Desvinculado). Los valores positivos indican una mayor relación con la etiqueta de Egresado, mientras que los valores negativos indican una mayor relación con la etiqueta de Desvinculado. Además, el valor numérico de la correlación proporciona información sobre la fuerza de esta relación: cuanto más cercano a 1 (positivo) o -1 (negativo), mayor es la relación entre el atributo y la variable objetivo. Los atributos que están más correlacionados con la etiqueta son: *Modalidad de aplicación, Deudor, Matrícula al día, Género, Becado, Edad y Unidades curriculares 1er sem (aprobadas), Unidades curriculares 1er sem (nota).*

Existe una alta correlación entre los atributos del primer y segundo semestre, lo que indica una fuerte vinculación entre el rendimiento en ambas mitades del año académico. La Edad se correlaciona notablemente con el Estado civil, el Curso, y la Asistencia diurna o vespertina. También se observa una relación significativa entre los atributos Deudor y Matrícula al día.

En la Figura 4.2a se muestra la distribución de etiquetas sobre los estudiantes, donde se observa una mayor proporción de estudiantes egresados que desvinculados. Sobre el atributo género, como se muestra en la Figura 4.2b, se tiene que un 65 % de estudiantes que se identifica con género femenino y 35 % con género masculino.

En la Figura 4.3a se presenta la distribución de etiquetas según los atributos Género, Deudor, Matrícula al día y Becado, evidenciando un desbalance en los datos. En el caso del atributo Género existe una mayor proporción de egresadas del género femenino que del género masculino. No se presentan instancias de género no binario, aunque tampoco se especifica si estos datos fueron recolectados o eliminados. Para el atributo Deudor, el 90 % de los estudiantes en el conjunto de datos no son deudores, y presentan un porcentaje de desvinculación menor que aquellos estudiantes deudores. Para esta categoría, podría pasar que el modelo interprete que si el estudiante es deudor, se va a desvincular de sus estudios. Podría ser un caso de alta tasa de falsos negativos para los estudiantes deudores. Con el atributo Matrícula al día, se obtiene un resultado similar. Podría ser un caso de alta tasa de falsos negativos para los estudiantes que no tienen sus matrículas al día. No se especifica bien la diferencia entre estas categorías, aun así tienen una correlación alta, como se puede ver en la matriz de correlación en la Figura 4.1. Con respecto a las becas un 27 % de los estudiantes tienen una beca de los cuales un 86 % son egresados, mientras que los estudiantes que no tienen una beca, solo un 51 % son egresados.

Si ahora se observa la distribución según el atributo Género (Figura 4.3b) se muestra que la distribución de deudores y no deudores es similar para género masculino y femenino. Mientras que los estudiantes de género masculino tienen menor porcentaje de matrículas al día que las de género femenino. Respecto a la distribución de becas según género, se distribuyen de manera que el 80 % de las estudiantes de género femenino tienen becas y 20 % para el género masculino.

La Figura 4.4 muestra que la proporción de egresados y la cantidad de alumnos egresados disminuyen con la edad al ingreso. Sólo se analizaron los alumnos menores de treinta años, ya que representan el 85 % de los datos.

En la Figura 4.5 se muestra la relación entre la ocupación de la madre y padre y la etiqueta. Se utilizaron solo las primeras catorce categorías, ya que representaban la mayor cantidad de datos disponibles. No se observó una relación significativa entre las ocupaciones y el egreso. Para la madre las categorías con mayor cantidad de datos (67 % del total) son:

- 10 - *Profesiones de las Fuerzas Armadas*
- 5 - *Trabajadores de Servicios Personales, Seguridad, Protección y Vendedores*
- 6 - *Agricultores y Trabajadores Especializados en Agricultura, Pesca y Silvicultura*

Para el padre las categorías con mayor cantidad de datos (59 % del total) son:

- 10 - *Profesiones de las Fuerzas Armadas*

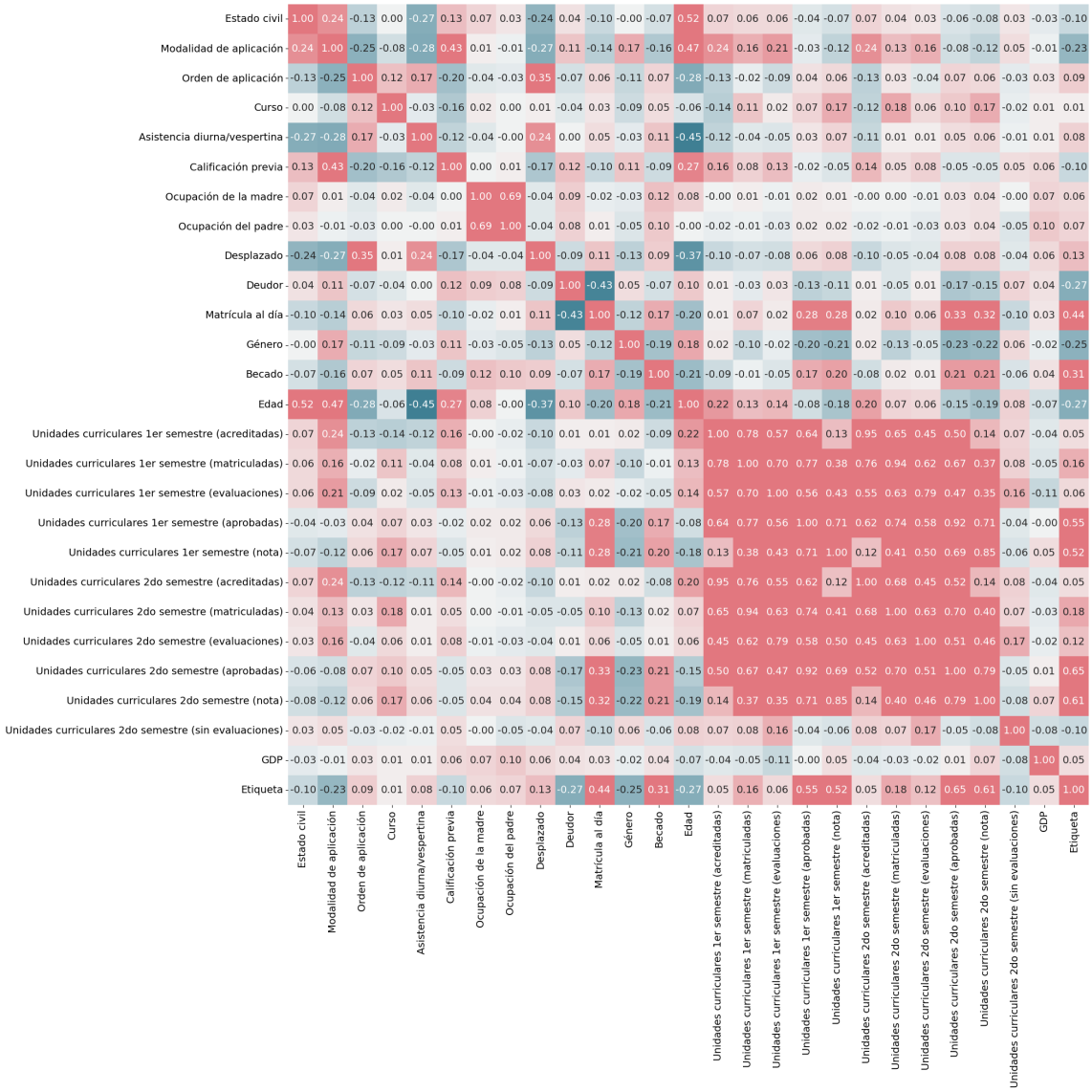
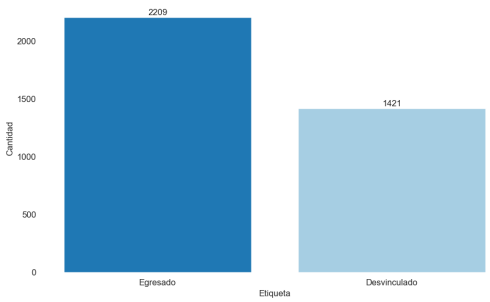
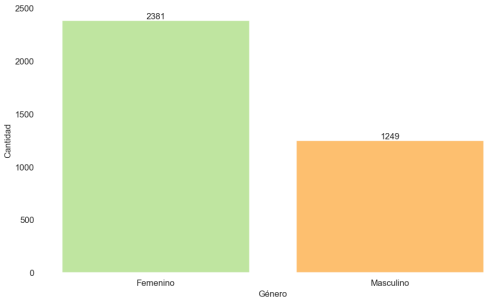


Figura 4.1: Matriz de correlación de los datos

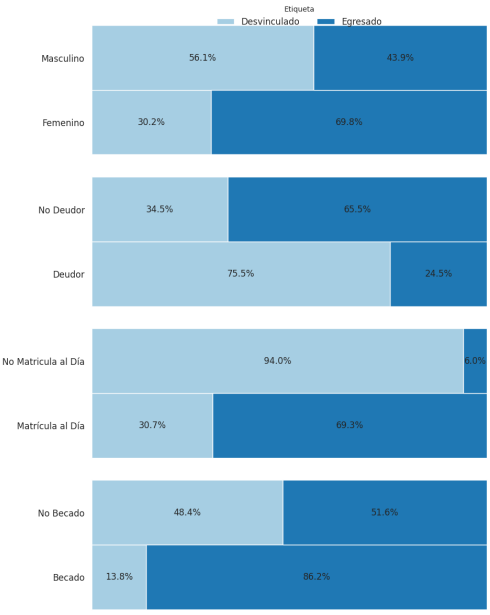


(a) Distribución de los estudiantes según su estatus académico: "Desvinculado" o "Egresado".

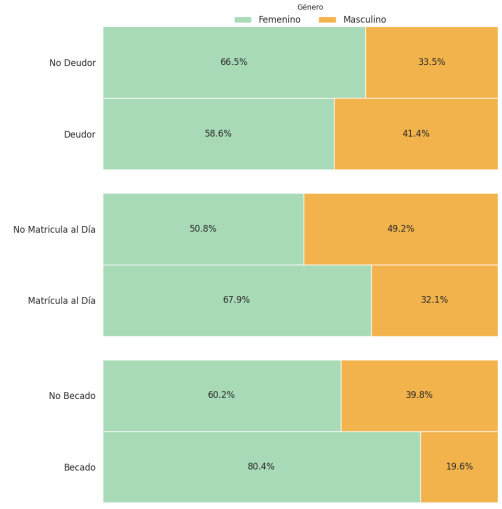


(b) Distribución de los estudiantes según su género: "Femenino" o "Masculino".

Figura 4.2: Distribuciones de atributos Género, Deudor, Matrícula al día y Becado según etiqueta asignada (a) y género (b).



(a) Distribución de etiquetas según atributos Género, Deudor, Matrícula al día y Becado.



(b) Distribución de Género respecto a los atributos Deudor, Matrícula al día y Becado.

Figura 4.3: Distribuciones de atributos Género, Deudor, Matrícula al día y Becado según etiqueta asignada (a) y género (b).

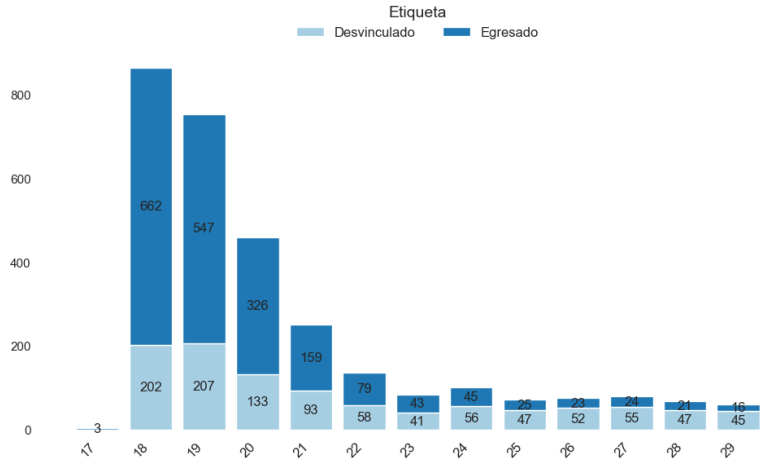


Figura 4.4: Distribución de los estudiantes según la edad de ingreso y etiqueta.

- 8 - *Operadores de Máquinas e Instalaciones y Trabajadores de Ensamblaje*
- 6 - *Agricultores y Trabajadores Especializados en Agricultura, Pesca y Silvicultura*
- 4 - *Personal Administrativo*

No se obtuvo información para justificar por qué la mayor parte de la ocupación de padres y madres de los estudiantes son Profesionales de las Fuerzas Armadas. Otro de los atributos que es de interés analizar como se distribuyen los estudiantes según el curso/carrera al cual se inscriben. La distribución de estudiantes "Desvinculados" o "Egresados" según el curso/carrera se presenta en la Figura 4.6, donde hay una tendencia a la desvinculación en el área de Tecnologías de Producción de Biocombustibles, seguido por Ingeniería Informática.

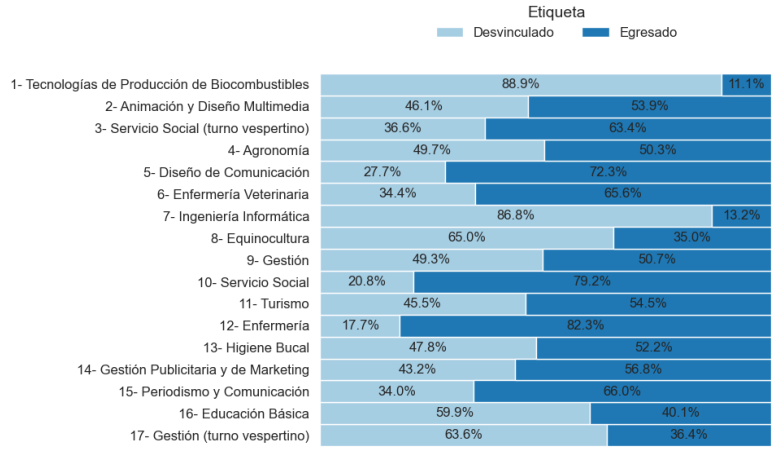
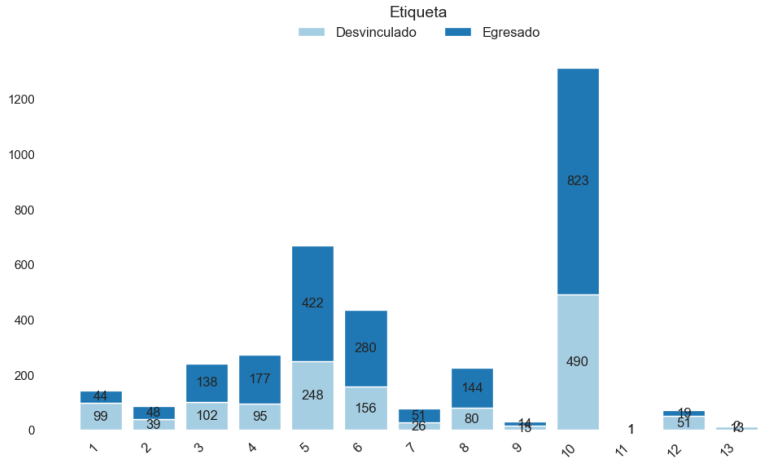
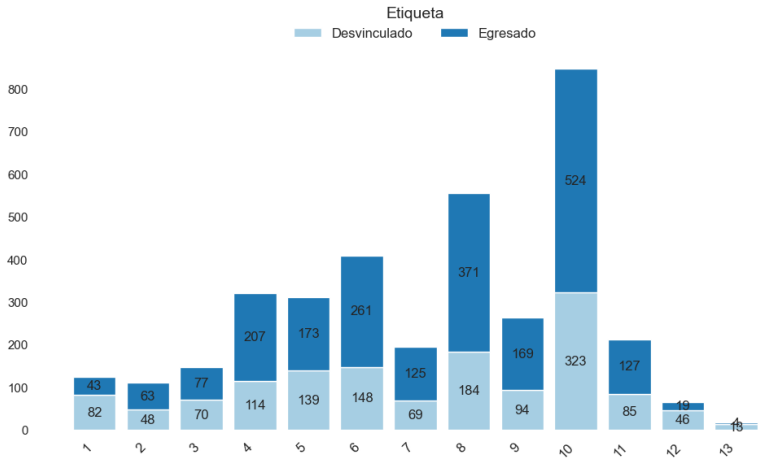


Figura 4.6: Distribución de los estudiantes según los cursos que toman y el estatus académico.



(a) Ocupación de la madre vs estatus académico.



(b) Ocupación del padre vs estatus académico.

Figura 4.5: Relación entre la ocupación de los padres y el estatus académico de los estudiantes.

4.3. Modelo

El modelo utilizado es un clasificador por votación (*Voting Classifier*), este consiste en predecir basado en las predicciones de un conjunto de modelos dado. En este caso se utilizó en su modalidad *soft*, donde cada modelo brinda una probabilidad de elegir las clases y la clase con mayor probabilidad de ocurrencia es la predicción final. Los modelos que son parte del conjunto de selección son Random Forest, Regresión Logística y Adaptive Boosting, explicados en la Sección 2.1.1.

En la Figura 4.7 se presenta la importancia de los atributos para cada modelo. Para el caso del modelo de regresión logística se indican los coeficientes que se asocian a cada atributos. Los atributos más importantes son similares en *Random Forest* y *Adaptive Boosting* siendo las materias aprobadas del primer y segundo semestre y la nota de las materias del segundo semestre. En lo que corresponde al modelo de regresión

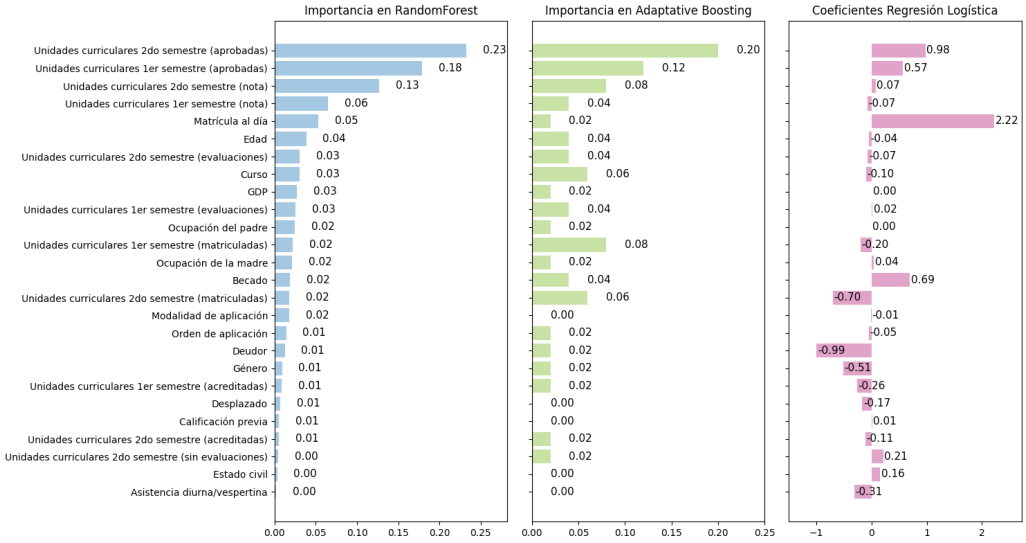


Figura 4.7: Importancia de los atributos según cada Random Forest, Adaptive Boosting y coeficientes de Regresión logística.

logística los coeficientes con más relevancia son los de los atributos Matrícula al día, materias aprobadas del semestre, y Deudor y Género como coeficientes negativos.

Para entrenar se utilizó la biblioteca *scikit-learn* [54] en Python, se dividió el conjunto de datos en un conjunto de entrenamiento (80 %) y un conjunto de prueba (20 %). El modelo clasifica si el estudiante se desvincula (-1) o egresa (1) de sus estudios. Para el conjunto de prueba, el modelo tiene un 91,74 % de tasa de acierto y la matriz de confusión se muestra en la Figura 4.8, donde se observa que 47 estudiantes fueron incorrectamente clasificados como egresados siendo desvinculados (falsos positivos) y 13 egresados fueron incorrectamente clasificados como desvinculados (falsos negativos). Los conceptos de falsos positivos, falsos negativos, matriz de confusión y tasa de acierto se explican en la Sección 2.1.

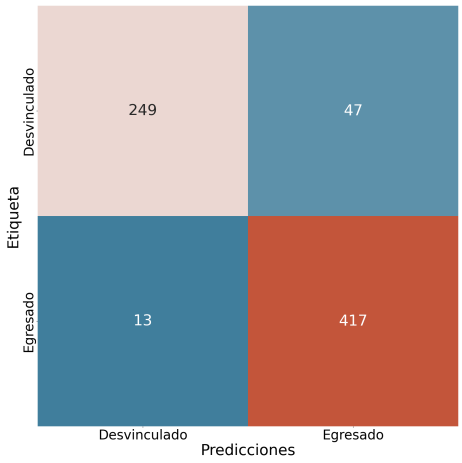


Figura 4.8: Matriz de confusión del modelo.

4.4. Análisis de Sesgos

En esta sección se presentará el análisis de sesgos algorítmicos realizado sobre el modelo, siguiendo el procedimiento y aplicando conceptos de el Capítulo 2. Se identifican los riesgos que este contiene el sistema frente a los atributos sensibles y se analiza como interfieren los sesgos de manera interseccional.

4.4.1. Identificación de Riesgos

El algoritmo asigna la etiqueta positiva (de valor 1) a los estudiantes egresados y la etiqueta negativa (de valor -1) a los estudiantes que se desvincularon. Esta clasificación binaria permite analizar los impactos negativos del sistema según su caso de uso. Un falso negativo ocurre cuando el modelo predice que un estudiante se desvinculó, pero en realidad egresó. De manera similar, un falso positivo se da cuando el modelo predice que un estudiante se egresó, cuando en realidad se desvinculó. Como no se especifica el caso de uso de este sistema, se pueden tener dos escenarios:

- **Caso de uso 1**: Si el objetivo es favorecer a los estudiantes con alta probabilidad de egresar, entonces un falso positivo sería beneficioso para los individuos, mientras que un falso negativo sería perjudicial.
- **Caso de uso 2**: Si el objetivo es incentivar a los estudiantes con alta probabilidad de abandonar, entonces un falso negativo beneficiaría al estudiante, mientras que un falso positivo sería perjudicial.

Como no está definido el caso de uso del modelo, se toma como supuesto el caso de uso 1.

4.4.2. Identificando Posibles Grupos Afectados

Dado el análisis de los datos realizado en la Sección 4.2, se seleccionaron los atributos Género, Edad, Ocupación de la madre, Ocupación del padre, Matrícula al día, Deudor y Beca para realizar el análisis de sesgos. Estos atributos fueron seleccionados debido a su alta correlación con la etiqueta y su significativo desbalance, como se mencionó en la Sección 4.2. Además, basado en el estudio realizado por [43], que analiza y releva conjuntos de datos utilizados en investigaciones sobre la equidad en la inteligencia artificial. En la sección dedicada a los conjuntos de datos relacionados con la educación, se destacan Edad y Género como atributos protegidos estudiados.

Para el análisis del atributo Edad, se consideraron únicamente los estudiantes menores de treinta años, ya que el subgrupo de estudiantes mayores de esa edad era demasiado reducido. En cuanto a los atributos Ocupación de la madre y Ocupación del padre, se analizaron los subgrupos de las categorías del 1 al 13, ya que en el resto contiene pocas instancias.

4.4.3. Análisis Utilizando Paridad Demográfica

Como se explica en la Sección 2.3.3.1, la *paridad demográfica* busca que las tasas de selección sean iguales en cada subgrupo de los atributos sensibles. La Figura 4.9 muestra el radio de paridad demográfica para cada atributo seleccionado, siendo los atributos que muestran mayor disparidad: Deudor, Ocupación de la madre, Ocupación del padre, Matrícula al día y Edad. Por otra parte Becado y Género muestran resultados más equitativos. Las diferencias para los atributos de Edad, Ocupación del padre y Ocupación de la madre se explican porque contienen categorías con muy pocos valores, como ya se observó en la Sección 4.2.

Para los atributos con menor radio de paridad demográfica se realiza un análisis más profundo analizando las tasas de selección de los subgrupos en la Tabla 4.1. Aquellos estudiantes que no son deudores obtienen una tasa de selección positiva mayor frente a los deudores, siendo favorecidos. Los estudiantes que cuentan con matrícula al día tienen una mayor tasa de selección frente a los que no, generando nuevamente un desfavorecimiento para aquellos estudiantes que no tienen matrícula al día. Hay una leve diferencia entre la tasa de selección de aquellos estudiantes que son de género masculino frente a quienes son de género femenino, siendo mayor la tasa de selección para el género femenino.

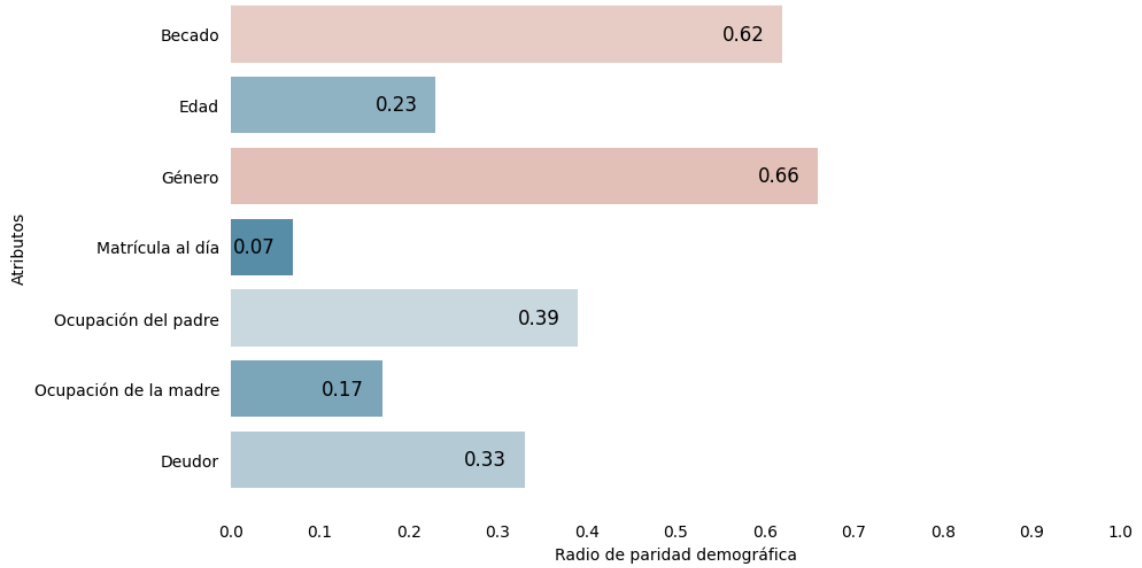


Figura 4.9: Radio de paridad demográfica para los atributos analizados. Un índice cercano a 1 en el radio indica paridad demográfica, mientras que un valor cercano a 0 sugiere una mayor disparidad.

Grupo	**Tasa de selección**
No deudor	0,70
Deudor	0,23
No tiene matrícula al día	0,05
Tiene matrícula al día	0,73
Femenino	0,71
Masculino	0,47

Tabla 4.1: Tasas de selección para los atributos Deudor, Matrícula al día y Género.

4.4.4. Análisis Utilizando Probabilidades Igualadas

La métrica de *probabilidades igualadas*, como se explicó en la Sección 2.3.3.2, busca que las tasas de verdaderos y falsos positivos sean iguales entre todos los subgrupos. En la Figura 4.10 se presentan los resultados del radio de probabilidades igualadas. Los atributos con valores más bajos de radio son Ocupación de la madre, Ocupación del padre, Edad, dando la mayor disparidad entre el conjunto de atributos analizados. Este fenómeno se debe a que existen categorías dentro de estos atributos que no cuentan con falsos positivos, por lo tanto al realizar el calculo del radio de probabilidades igualadas dan un valor nulo. Sin tener en cuenta estos atributos, Becado y Matrícula al día son los que obtienen peor desempeño según la métrica seguidos por Género y Deudor, que tienen un mejor rendimiento en este sentido.

Para el caso de aquellos estudiantes que reciben una beca se tiene un radio de 0,27 este resultado se desglosa en la Tabla 4.2, donde se presentan las tasas de verdaderos positivos y falsos positivos para cada categoría dentro de este atributo. Se observa que aquellos estudiantes que están becados presentan una mayor tasa de falsos positivos, lo cual significa que son calificados positivamente con una probabilidad mayor que aquellos estudiantes que no cuentan con una beca. Este análisis es importante para ver la

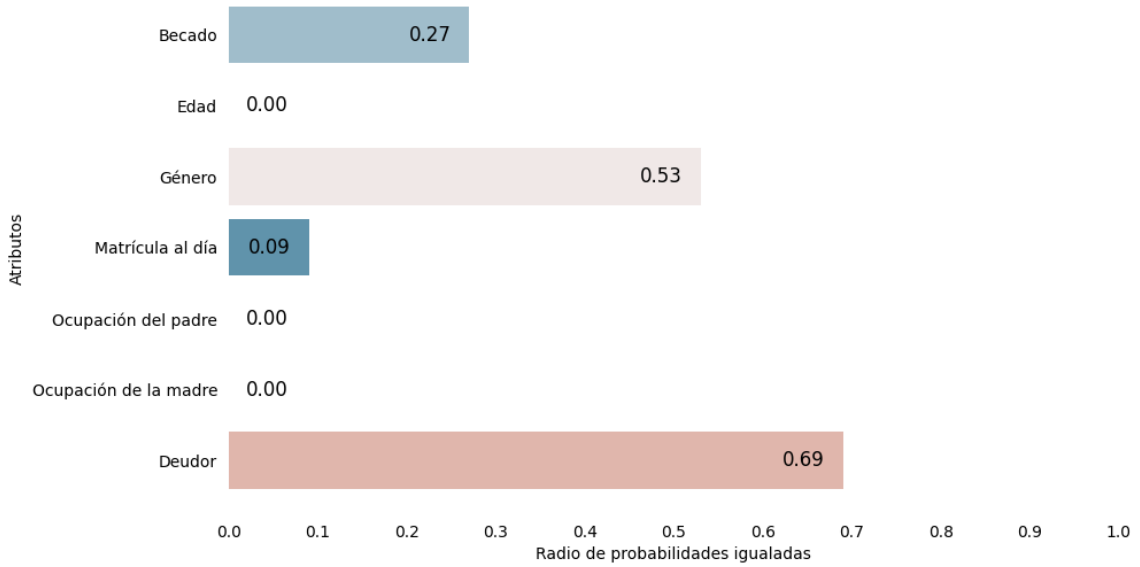


Figura 4.10: Radio de probabilidades igualadas para los atributos analizados. Un valor cercano a 1 del radio indica un mejor desempeño de la métrica que un valor cercano a 0.

diferencia entre las métricas, ya que este mismo atributo medido frente a la métrica de paridad demográfica no parecía presentar sesgos en la decisión.

Grupo	**Tasa de verdaderos positivos**	**Tasa de falsos positivos**
No becado	0,96	0,12
Becado	0,98	0,46
No tiene matrícula al día	0,75	0,02
Tiene matrícula al día	0,97	0,22
Femenino	0,98	0,20
Masculino	0,92	0,10

Tabla 4.2: Tasas de verdaderos positivos y falsos positivos para los atributos Becado, Matrícula al día y Género.

Al observar los valores de la Tabla 4.2 que presenta la tasa de verdaderos positivos y falsos positivos para el atributo de Matrícula al día, se nota que los estudiantes sin la matrícula al día tienen una tasa de falsos negativos casi nula a diferencia de los que no tienen la matrícula al día, lo cual puede indicar un sesgo en la predicción relacionado con este atributo. Por último se observa con detenimiento el atributo Género, donde las tasas de falsos positivos y verdaderos positivos son similares para ambos géneros.

4.4.5. Análisis con Enfoque Interseccional

Dada la importancia de realizar un análisis con enfoque interseccional como se explica en la Sección 2.3.1, se analizan los atributos en conjunto. Para esto se redujo el conjunto de atributos a aquellos que son binarias y que arrojaron resultados desfavorables según el análisis anterior. Los atributos seleccionados son Deudor, Matrícula al día, Género y Beca.

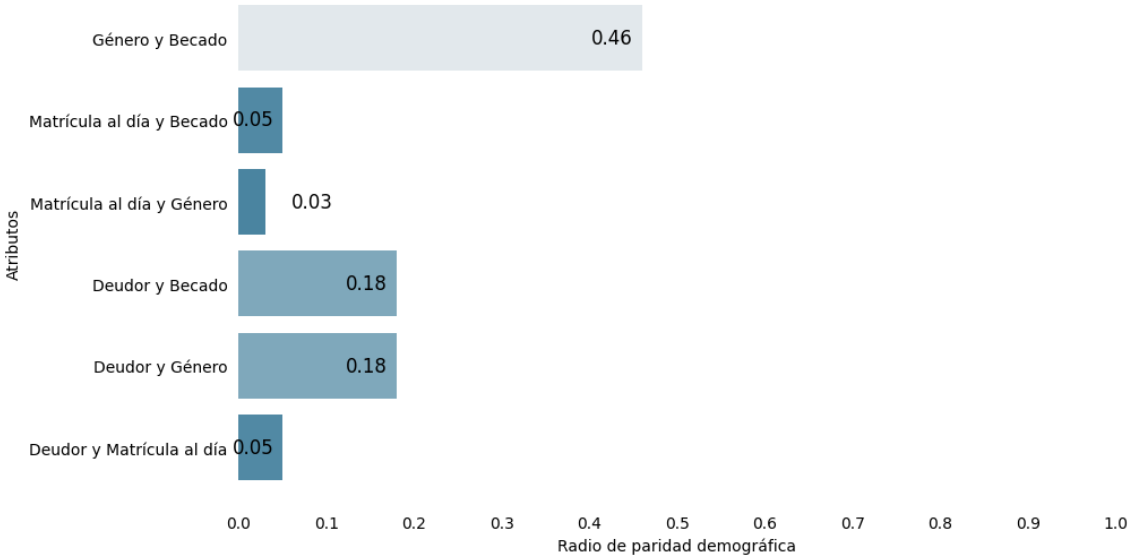


Figura 4.11: Radio de paridad demográfica interseccional tomando de a dos atributos analizados.

En la Figura 4.11 se presentan los resultados de evaluar la paridad demográfica interseccionalmente, tomando de a dos atributos por vez. A partir de los resultados se destaca que la presencia de los atributos Deudor y Matrícula al día tiene peso negativo sobre los demás, produciendo una disminución en el radio de paridad demográfica. En particular la tupla de atributos Matrícula al día y Género obtienen el peor desempeño dentro del conjunto. En la Tabla 4.3 se presentan los resultados para la tupla de atributos Matrícula al día y Género, donde se observa una alta disparidad demográfica, principalmente dada por contar o no con la matrícula al día. Además la tasa de selección para aquellos que tienen matrícula al día es mayor para los estudiantes de género femenino que masculino pero la diferencia en cantidad de estudiantes es baja.

Matrícula al día	**Género**	**Tasa de selección**	**Cantidad de estudiantes**
No	Femenino	0,07	56
	Masculino	0,02	46
Sí	Femenino	0,80	431
	Masculino	0,58	193

Tabla 4.3: Tasas de selección para los atributos Matrícula al día y Género.

Respecto a la métrica de probabilidades igualadas, la presencia del atributo Matrícula al día genera un radio de valor nulo (Figura 4.12). Esto se explica por qué no existen falsos positivos en alguna de las categorías, como en el caso de el atributo Género, en la Tabla 4.4 se detecta que la tasa de falsos positivos es nula para los estudiantes de género masculino sin matrícula al día. Este resultado desfavorece a estudiantes de género femenino respecto a la métrica descrita, ya que esta categoría representa una mayoría en el conjunto.

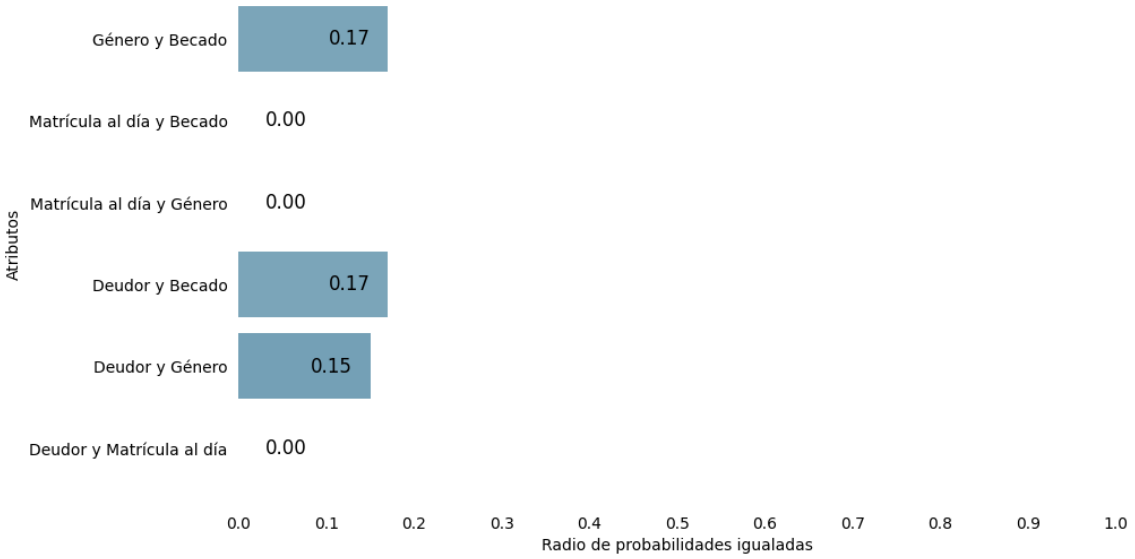


Figura 4.12: Radio de probabilidades igualadas tomando de a dos atributos analizados.

Matrícula al día	Género	TVP	TFP	Cantidad de estudiantes
No	Femenino	0,67	0,04	56
	Masculino	1	0	46
Sí	Femenino	0,99	0,28	431
	Masculino	0,92	0,16	193

Tabla 4.4: Tasas de verdaderos positivos (TVP) y falsos positivos (TFP) para el atributo Género y Matrícula al día.

4.4.6. Resultados del Análisis de Sesgos Algorítmicos

En principio se realizó un análisis de cada atributo por separado, en este análisis se refleja que los atributos que tienen un desbalance en los datos de entrada son los que generan un gran impacto sobre el modelo como es el caso de Edad, Ocupación de la madre y Ocupación del padre que son grupos en donde existe una gran disparidad demográfica producto de la diversidad de los datos a la entrada. Se identifica el sesgo de representación porque el modelo no logra generalizar bien para estas categorías por contar con pocos ejemplos. Además estos atributos se identifica una limitación de los métodos estudiados para analizar e identificar sesgos algorítmicos, ya que cuando existen muchas clases, es común que el radio sea cero si en uno de los grupos el valor de la métrica es cero. Por otra parte dentro del conjunto de atributos, Matrícula al día es el que refleja una diferencia mayor entre aquellos estudiantes que se predicen como egresados y como desvinculados. Respecto a este atributo, la tasa de selección tiene valores más bajos para aquellos estudiantes que no tienen matrícula al día frente a los que sí. La misma situación se repite para la tasa de falsos positivos de los grupos dentro de este atributo, dejando en evidencia un desfavorecimiento del grupo de estudiantes que no tienen matrícula al día.

A continuación, se realizó un análisis interseccional para visualizar cómo inciden los sesgos al combinar múltiples atributos. Considerando los atributos por sí solos no siempre muestran una desfavorecimiento para algún grupo, por eso es que en conjunto se pueden visualizar ciertas diferencias que en el análisis individual no se observan. En el caso del atributo Género por sí solo no representa un desfavorecimiento

para ninguna de las clases, sin embargo, al realizarse el análisis interseccional en conjunto con el atributo Matrícula al día se puede visualizar un desfavorecimiento de aquellos estudiantes de género masculino.

Este enfoque también permite identificar atributos que operan de manera independiente, ya que, incluso al ser analizados en conjunto, siguen mostrando una alta tasa de desfavorecimiento. Tal es el caso de Matrícula al día, un atributo que disminuye significativamente el valor tanto del radio de paridad demográfica como de las probabilidades igualadas cuando se combina con otros atributos. Además, este atributo presentó los peores resultados en ambas métricas, lo que resalta su impacto negativo en el análisis de equidad.

Los sesgos encontrados pueden generar daños de asignación ya que el modelo repite sistemáticamente selecciones desfavorables para estudiantes dentro de grupos sensibles. Por otra parte en este análisis se evidencia la eficiencia de las métricas seleccionadas para medir sesgos. Por un lado la paridad demográfica presenta buenos resultados al momento de identificar si existe diferencias en las tasas de selección en distintos grupos. Mientras que la métrica de probabilidades igualadas proporciona una métrica de qué tanto se esta equivocando en la predicción en distintos grupos pero no es efectiva para los casos en que se tienen pocos datos en alguna de las categorías como es el caso de los atributos Edad, Ocupación de la madre y Ocupación del padre.

4.5. Mitigación de Sesgos

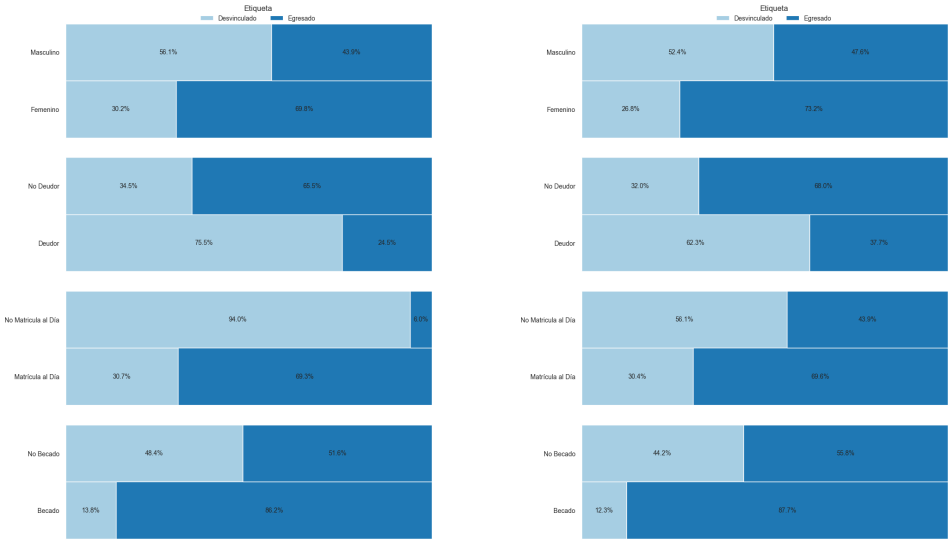
En esta sección se implementan distintos métodos de mitigación de sesgos algorítmicos aplicados al problema presentado. Basado en el análisis de sesgos realizado en la sección anterior, se destaca el atributo Matrícula al día como determinante para el desempeño del sistema. En esta etapa de mitigación se busca atacar el desbalance en el conjunto de entrenamiento y la incidencia del atributo Matrícula al día en la toma de decisiones del modelo intentando no modificar significativamente el desempeño del sistema.

La elección de aplicar las técnicas únicamente en el atributo Matrícula al día se debe a que fue el que mostró un desempeño más deficiente en el análisis de sesgos. Además, al intentar aplicar métodos de mitigación a múltiples atributos simultáneamente, los resultados no fueron satisfactorios. Al introducir numerosas condiciones en los algoritmos, no se logró un equilibrio adecuado entre el rendimiento del sistema y las métricas de equidad.

4.5.1. Mitigación en el Preprocesamiento

Tal como se presento en la Sección 4.2, se presenta un desbalance el grupo de entrenamiento. Para mitigar este desbalance se entrenó un modelo generativo con la arquitectura CTGAN mencionada en la Sección 2.3.4.1, con la biblioteca *YData Synthetic* [60], para generar datos sintéticos con los datos de entrenamiento. Se tomó una muestra de 2000 datos sintéticos y luego de verificar que su distribución coincidía con los datos reales, se seleccionaron los que correspondían a estudiantes sin la matrícula al día y egresados. Se concatenaron los datos de entrenamiento con 256 datos nuevos, en total obteniendo un conjunto de entrenamiento de 3160 datos. En la Figura 4.13 se muestra la diferencia en la distribución del conjunto de entrenamiento original y conjunto de entrenamiento balanceado. Se resalta que la tasa de estudiantes que no tienen la matrícula al día y se egresaron aumentó de 6 % a un 44 %. Además la distribución del resto de los atributos no se vio afectada significativamente por este agregado de datos. Con este nuevo conjunto de entrenamiento se entrenó el mismo clasificador mencionado en la Sección 4.3 y se obtuvo una tasa de aciertos de 90,36 %.

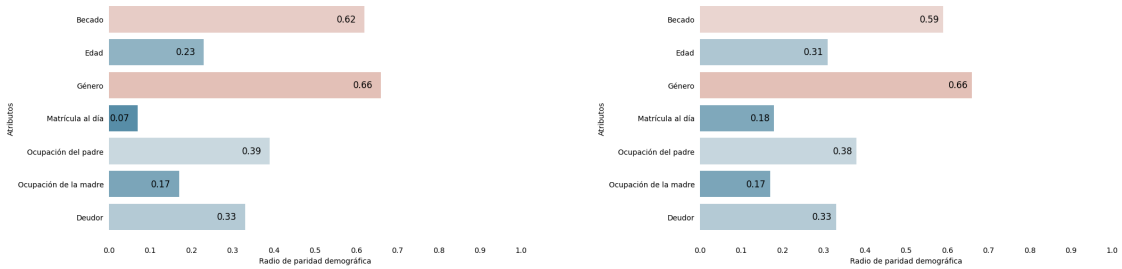
En lo que respecta a las medidas de sesgos algorítmicos del modelo, en la Figura 4.14 se presentan los resultados de paridad demográfica antes y después de aplicar la mitigación. En dicha figura se puede apreciar que, respecto al conjunto inicial, el atributo Matrícula al día mejora su desempeño de un radio de 0,07 a uno de 0,18. La aplicación de esta técnica no afecta significativamente a los demás atributos de forma individual. Se presentan en la Figura 4.15 mejoras en la mayoría de los grupos tomando los atributos de manera interseccional, habiendo casos en los que disminuye en un pequeño porcentaje.



(a) Distribución original de atributos según etiqueta.

(b) Distribución de atributos según etiqueta luego de aplicar balanceo.

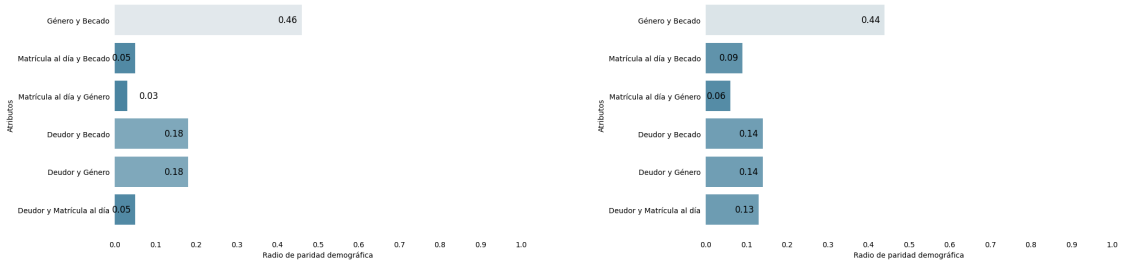
Figura 4.13: Distribución de los atributos Género, Deudor, Matrícula al día y Becado según etiquetas antes y después del balanceo.



(a) Radio de paridad demográfica para los atributos analizados en el modelo original.

(b) Radio de paridad demográfica luego de aplicar balanceo de datos.

Figura 4.14: Radio de paridad demográfica antes y después de aplicar balanceo de datos.



(a) Radio de paridad demográfica interseccional para los atributos analizados en el modelo original.

(b) Radio de paridad demográfica interseccional luego de aplicar balanceo de datos.

Figura 4.15: Radio de paridad demográfica interseccional antes y después de aplicar balanceo de datos.

(a) Radio de probabilidades igualadas para los atributos analizados en el modelo original.

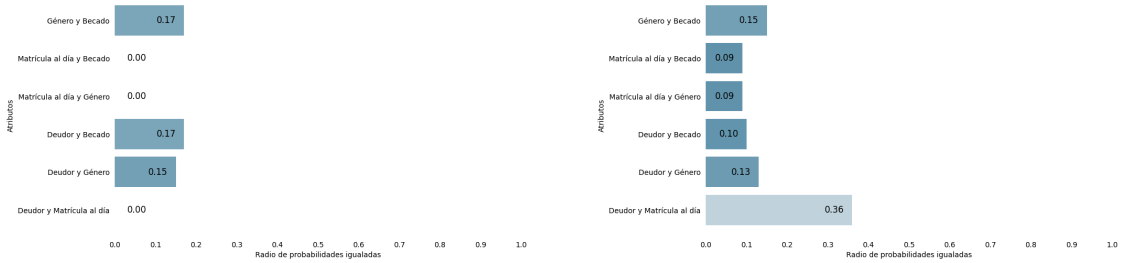
(b) Radio de probabilidades igualadas después del balanceo.

Figura 4.16: Radio de probabilidades igualadas antes y después de aplicar balanceo de datos.

Para el caso de la métrica de probabilidades igualadas se obtienen los resultados de la Figura 4.16, donde el atributo Matrícula al día pasa a tener un radio cinco veces mayor que el original. Esto se explica porque luego del balanceo hay una mayor cantidad de datos en el nuevo conjunto de entrenamiento de personas que no tienen la matrícula al día y están graduadas, entonces, luego de entrenar el modelo se refleja en un aumento en la tasa de verdaderos y falsos positivos. La Tabla 4.5 presenta los resultados específicos para cada categoría dentro del atributo. En la Figura 4.17, se puede observar que de manera individual no se ven afectados los atributos pero sí su desempeño interseccional cuando se analiza el desempeño en conjunto con el atributo Matrícula al día se obtienen mejoras en los resultados pero los demás atributos se ven desfavorecidos de manera interseccional.

Matrícula al día	**TVP**	**TFP**
No tiene matrícula al día	1,00	0,09
Tiene matrícula al día	0,96	0,22

Tabla 4.5: Tasas de verdaderos positivos (TVP) y falsos positivos (TFP) para el atributo Matrícula al día.



(a) Radio de probabilidades igualadas interseccional para los atributos analizados en el modelo original.

(b) Radio de probabilidades igualadas interseccional después del balanceo.

Figura 4.17: Radio de probabilidades igualadas interseccional antes y después de aplicar balanceo de datos.

4.5.2. Mitigación en el Entrenamiento

Se explora el método de *reducción* descrito en la Sección 2.3.4.2. Para ello, se utilizó la implementación de la biblioteca *Fairlearn*. El proceso consistió en proporcionar el modelo que se desea entrenar, el mismo que se explicó en la Sección 4.3. Además, se especificó la métrica de equidad a emplear, que en este caso fue probabilidades igualadas y como resultado se obtuvo una tasa de aciertos de 88,02 %. La matriz de confusión se muestra en la Figura 4.18.

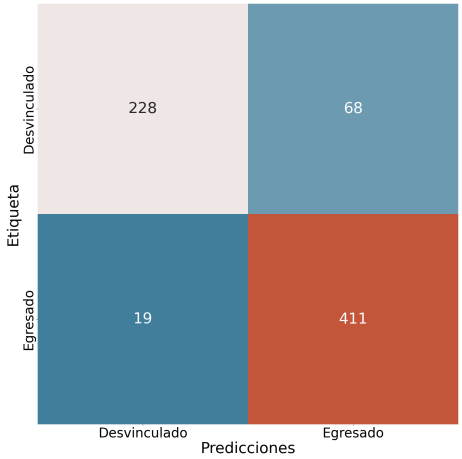
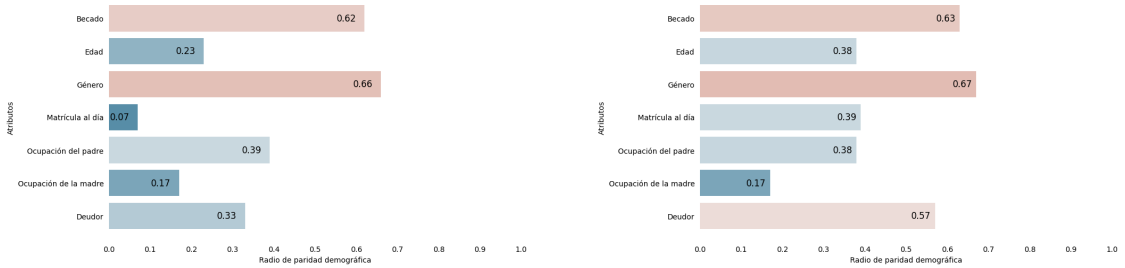


Figura 4.18: Matriz de confusión para la mitigación en el entrenamiento.

Respecto al análisis de sesgos en la Figura 4.19 se presenta el radio de paridad demográfica antes y después de aplicar mitigación en el entrenamiento. Se observa que los atributos que varían significativamente su valor sobre esta métrica son Matrícula al día y Deudor obteniendo un mejor rendimiento. El resto de los atributos mantienen sus niveles de tasas de selección. En el caso de la métrica probabilidades igualadas se observa que el atributo sensible Matrícula al día, aumenta su radio de 0,07 a 0,85, mejorando el desempeño de la métrica para este atributo (Figura 4.20).

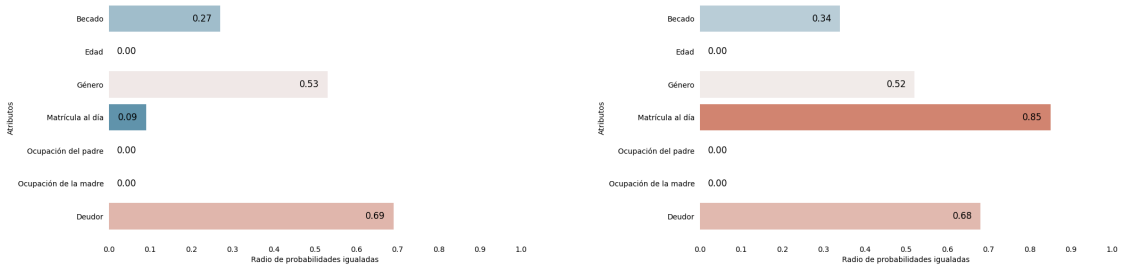
Realizando un análisis con enfoque interseccional se obtienen resultados favorables respecto al caso original como se muestra en la Figura 4.21. Se analiza la inferencia sobre la métrica de probabilidades



(a) Radio de paridad demográfica para los atributos analizados en el modelo original.

(b) Radio paridad demográfica luego de aplicar mitigación en el entrenamiento.

Figura 4.19: Radio de paridad demográfica antes y después de aplicar mitigación en el entrenamiento.

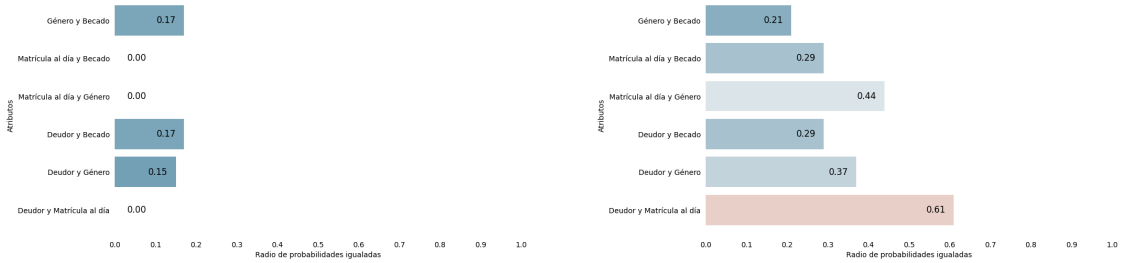


(a) Radio de probabilidades igualadas para los atributos analizados en el modelo original.

(b) Radio de probabilidades igualadas después de aplicar mitigación en el entrenamiento.

Figura 4.20: Radio de probabilidades igualadas antes y después de aplicar mitigación en el entrenamiento.

igualadas, pues es en base a dicha métrica que se aplica la mitigación en el entrenamiento. A diferencia del caso original, ahora todas los atributos que se acompañan de Matrícula al día tienen un radio de probabilidades igualadas de más de 0,29 lo cual indica un mejor desempeño del atributo respecto a la métrica en todo el conjunto, considerándolo tanto de manera individual como interseccionalmente.



(a) Radio de probabilidades igualadas interseccional para los atributos analizados en el modelo original.

(b) Radio de probabilidades igualadas para atributos de manera interseccional después de aplicar mitigación en el entrenamiento.

Figura 4.21: Radio de probabilidades igualadas para atributos de manera interseccional antes y después de aplicar mitigación en el entrenamiento.

4.5.3. Mitigación en el Postprocesamiento

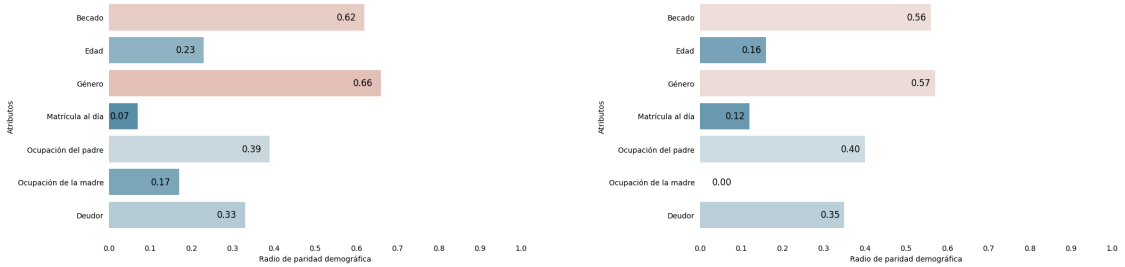
Con el fin de aplicar un método de mitigación alternativo para disminuir los sesgos algorítmicos existentes en el modelo original se exploró con la técnica de mitigación en el postprocesamiento llamado optimizador de umbral, mencionado en la Sección 2.3.4.3. Para aplicar dicha técnica se utilizó la implementación de la biblioteca *Fairlearn* en donde se requiere primero identificar los atributos sensibles y las métricas de equidad que se buscan mitigar. Dado el análisis realizado en la Sección 4.2 se identifica el atributo Matrícula al día como atributo sensible y probabilidades igualadas como métrica de equidad. Finalmente esta técnica de mitigación logra un porcentaje de acierto del 89,53 %.

En lo que respecta al análisis de sesgos, se observa que para la métrica de paridad demográfica como resultado se presenta una mejora en la medida del atributo Matrícula al día, pero los atributos de Becado, Edad y Ocupación de la madre se ven afectados negativamente (Figura 4.22). En particular este último pasa a tener una diferencia en la tasa de selección de manera tal que el radio de la diferencia es 0. Para la métrica de probabilidades igualadas presentadas en la Figura 4.23 se observa una mejora respecto a los valores originales en el atributo Matrícula al día, que pasa a tener un radio de 0,26 cuando originalmente tenía un valor de 0,09. Sin embargo el desempeño de esta métrica respecto al atributo de Género disminuye, lo cual indica un desfavorecimiento de una de las categorías dentro del atributo. En la Tabla 4.6 se presentan las tasas de verdaderos positivos y falsos positivos para el atributo Género después de aplicar el optimizador de umbral, se observa que disminuye la tasa de verdaderos positivos y falsos positivos para el género masculino lo cual lleva a la disminución del radio de probabilidades igualadas.

Género	**TVP**	**TFP**
Femenino	0,94	0,16
Masculino	0,79	0,05

Tabla 4.6: Tasas de verdaderos positivos (TVP) y falsos positivos (TFP) para el atributo Género.

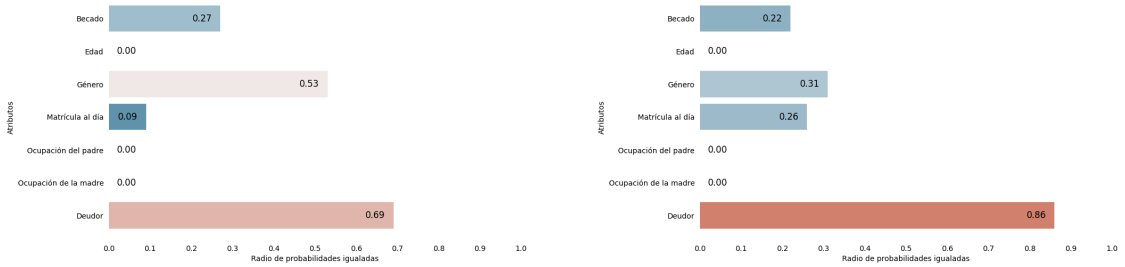
En el análisis interseccional, los resultados se muestran en la Figura 4.24, donde se observaron mejoras en los casos de los atributos Deudor y Matrícula al día, así como en los atributos Matrícula al día y Becado, en comparación con el caso original. Para los otros grupos, la técnica de mitigación redujo o no altero el radio de probabilidades igualadas.



(a) Radio de paridad demográfica para los atributos analizados en el modelo original.

(b) Radio paridad demográfica luego de aplicar Optimizador de Umbral.

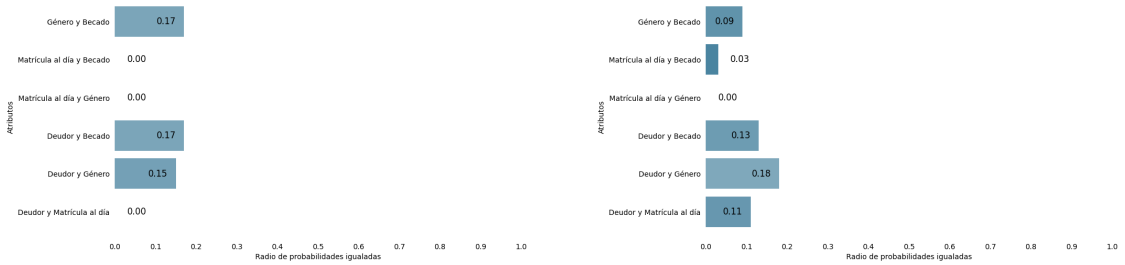
Figura 4.22: Radio de paridad demográfica antes y después de aplicar Optimizador de Umbral.



(a) Radio de probabilidades igualadas para los atributos analizados en el modelo original.

(b) Radio de probabilidades igualadas después de aplicar Optimizador de Umbral.

Figura 4.23: Radio de probabilidades igualadas antes y después de aplicar Optimizador de Umbral.



(a) Radio de probabilidades igualadas interseccional para los atributos analizados en el modelo original.

(b) Radio de probabilidades igualadas para atributos de manera interseccional después de aplicar Optimizador de Umbral.

Figura 4.24: Radio de probabilidades igualadas para atributos de manera interseccional antes y después de aplicar Optimizador de Umbral.

4.5.4. Resultados de la Aplicación de Técnicas de Mitigación

En la Tabla 4.7 se presentan los resultados de las tasas de aciertos obtenidos para las mitigaciones aplicadas. Se muestra que el balanceo de los datos es la técnica que menos afecta al sistema original en lo que respecta a porcentaje de acierto. Mientras que la técnica de reducción obtiene el peor desempeño. En lo que respecta a las métricas de sesgos algorítmicos se presentan los resultados comparativos en las Tablas 4.8 y 4.9. Estas técnicas fueron implementadas con el objetivo de disminuir el sesgo producido por el atributo Matrícula al día, presentando resultados más equitativos para este atributo implementando cualquiera de las técnicas respecto al modelo original sin mitigación. Para los atributos de Edad, Ocupación de la madre y Ocupación del padre pasan a tener un peor desempeño respecto al original logrando valores nulos en la paridad demográfica. Otra observación de degradación importante sucede con el atributo Género en la técnica de optimizador de umbral, presentando una mayor brecha en el radio en comparación con el modelo original.

La técnica que presenta mejores resultados respecto a las métricas de equidad es la reducción, cuando se aplica esta técnica aumentan los valores de radio de todos los atributos. Si bien es la técnica con menor tasa de acierto, no obtiene una diferencia significativa frente al modelo original.

Técnica	**Tasa de aciertos (%)**
Sin mitigación	91,74
Balanceo de datos	90,36
Reducción	88,02
Optimizador de umbral	89,53

Tabla 4.7: Tasa de aciertos para las técnicas de mitigación aplicadas en comparación del modelo sin mitigación.

	Radio de Paridad Demográfica						
Técnica	Becado	Edad	Género	Matrícula al día	Deudor	Ocupación Madre	Ocupación Padre
Sin mitigación	0,62	0,23	0,66	0,07	0,33	0,17	0,39
Balanceo de datos	0,59	0,31	0,66	0,18	0,38	0,17	0,33
Reducción	0,63	0,38	0,67	0,39	0,38	0,17	0,57
Optimizador de umbral	0,53	0,16	0,57	0,12	0,35	0	0,40

Tabla 4.8: Radio de paridad demográfica para para las técnicas de mitigación aplicadas en comparación del modelo sin mitigación.

	Radio de Probabilidades Igualadas						
Técnica	Becado	Edad	Género	Matrícula al día	Deudor	Ocupación Madre	Ocupación Padre
Sin mitigación	0,27	0	0,53	0,09	0,69	0	0
Balanceo de datos	0,23	0	0,41	0,42	0,61	0	0
Reducción	0,34	0	0,52	0,85	0,68	0	0
Optimizador de umbral	0,22	0	0,31	0,26	0,86	0	0

Tabla 4.9: Radio de probabilidades igualadas para las técnicas de mitigación aplicadas en comparación del modelo sin mitigación.

4.6. Conclusiones del Estudio de Caso

En el capítulo, se analizó el proyecto "Student Dropout Analysis for School Education" [53] con el fin de familiarizarse con herramientas de evaluación de sesgos algorítmicos y mitigación. Se examinó el conjunto de datos y su distribución, detectando un notable desbalance en varios de sus atributos. Se evaluó el desempeño

del modelo y se extrajeron los atributos más relevantes. Se identificaron algunos sesgos presentes mediante dos métricas de equidad: el radio de probabilidades igualadas y el radio de paridad demográfica. El análisis reveló que ambas métricas arrojan resultados distintos, lo que subraya la importancia de comprender qué mide cada una y seleccionar la métrica adecuada en función del objetivo del estudio.

Además se seleccionaron grupos de atributos a analizar. El atributo que presentó mayor desigualdad entre sus categorías, de acuerdo con las métricas de equidad evaluadas, fue Matrícula al día. Al analizar de manera interseccional el conjunto de atributos sensibles se identificaron sesgos algorítmicos que individualmente no eran detectados. Por ejemplo, el análisis conjunto de Matrícula al día y Género evidenció un peor rendimiento para el género femenino mientras que si se considera solo el atributo Género no se presenta esta disparidad.

Una limitación identificada al usar métricas de equidad fue que, en algunos casos, el radio de probabilidades igualadas resultó nulo. Esto ocurre cuando se analizan subgrupos pequeños, ya que es posible que algunos no contengan suficientes casos, lo que distorsiona la métrica y la hace poco útil para la evaluación global. . Esta observación coincide con lo discutido en la Sección 2.3.1.

Finalmente, se exploraron tres técnicas de mitigación de sesgos algorítmicos, buscando mitigar los sesgos algorítmicos en el atributo Matrícula al día. Una se aplicó en la etapa de preprocesamiento, intentando balancear el conjunto de datos en relación con la distribución del atributo Matrícula al día. La segunda se aplicó al momento del entrenamiento, aplicando una restricción teniendo en cuenta la métrica de probabilidades igualadas. Otra técnica utilizada fue un optimizador de umbral aplicado en la etapa de postprocesamiento, que busca modificar el umbral de las las predicciones para mejorar el desempeño de la métrica de equidad.

Las técnicas mostraron resultados similares en cuanto al desempeño del modelo, mejorando las métricas de equidad para el atributo Matrícula al día. Teniendo en cuenta los dilemas presentados en la Sección 2.3.5 se logro aplicar técnicas de mitigación sin comprometer significativamente el desempeño del modelo. Sin embargo, también tuvieron efectos negativos en otros grupos de atributos, lo que sugiere que la mitigación de sesgos algorítmicos puede no ser efectiva si perjudica a otros grupos. Se observó que la técnica de mitigación más eficiente para la métrica de equidad que se busca compensar no resulto la más efectiva para el desempeño general del sistema, se entiende que tiene que existir en estos casos un compromiso entre no agravar desigualdades y el desempeño final del modelo.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 5

Conclusiones y Trabajo Futuro

El objetivo principal del proyecto fue analizar la equidad en sistemas de ciencia de datos en el ámbito educativo, con un enfoque basado en el feminismo de datos, para fomentar una visión más inclusiva y equitativa. Para cumplir este objetivo, se introdujeron conceptos clave sobre ciencia de datos y equidad, destacando la importancia del feminismo de datos y la necesidad de una mayor diversidad y representación en los modelos y sus datos. También se presentaron las métricas de equidad empleadas en el análisis y la importancia de seleccionar las métricas adecuadas para cada estudio. Se exploraron las técnicas de mitigación de sesgos algorítmicos más frecuentes en el área y se discutieron los desafíos y dilemas encontrados.

Con el objetivo de relevar proyectos de ciencia de datos aplicados a la educación en Uruguay, se realizó una investigación en la que se contactó a los autores de dichos proyectos y se llevó a cabo un análisis detallado de los mismos. Durante este proceso se identificaron limitaciones en cuanto al acceso a los datos y su representatividad. Además, se evidenció la fragmentación de las fuentes de datos y las dificultades para realizar estudios integrales. A pesar de estas limitaciones, los proyectos revisados mostraron un uso consistente de modelos como la Regresión Logística y Random Forest, con un buen desempeño predictivo, aunque con desafíos relacionados con la equidad y la consideración de otras identidades de género. Aún asi estos proyectos presentaron un alto porcentaje de acierto, evidenciando la eficiencia de los atributos que se utilizan para predecir. Es importante destacar que ninguno de los sistemas relevados se encuentra actualmente en uso en entornos de producción; la mayoría de ellos está en fases piloto o experimentales. Esto implica que no ha sido posible evaluar el cumplimiento efectivo de los objetivos que estos sistemas buscaban resolver ni las consecuencias potenciales de su implementación en un contexto real. Otro aspecto identificado es la falta de representatividad de los conjuntos de datos utilizados, lo que se agravó por la escasez de datos. Este análisis se profundiza para dos proyectos en los Anexos A y B, donde no fue posible implementar técnicas de análisis de sesgos algorítmicos por las características de los proyectos. Las limitaciones identificadas no tienen como objetivo criticar los proyectos analizados sino más bien resaltar desafíos comunes en el ámbito de la ciencia de datos aplicada a la educación en Uruguay.

Con el fin de aplicar las técnicas de análisis y mitigación de sesgos algorítmicos estudiadas, se abordó un estudio de caso centrado en el análisis de desvinculación estudiantil. En este capítulo, se identificaron desbalances significativos en los datos y se evaluaron sesgos algorítmicos utilizando dos métricas de equidad: el radio de probabilidades igualadas y el radio de paridad demográfica. Se observó que la métrica de equidad seleccionada afecta los resultados, lo que resalta la importancia de comprender qué se busca medir. Se encontró que las métricas de sesgos seleccionadas no son buenas para identificar sesgos al evaluar en subgrupos pequeños. Además, se probaron técnicas de mitigación de sesgos, se encontró que la técnica que más mejoró con respecto a las métricas de equidad fue el método de reducción. En los otros casos mejoraron la equidad en algunos aspectos y generaron efectos negativos en otros grupos, lo que plantea la necesidad de una evaluación cuidadosa en la implementación de estas técnicas. Otra conclusión a destacar de este

análisis es que aunque los sistemas tengan un buen desempeño, se debe tener en cuenta el compromiso que existe entre la equidad y el desempeño del sistema.

En conjunto, el proyecto resalta la importancia de analizar y mitigar los sesgos algorítmicos en sistemas de ciencia de datos aplicados a la educación. Aunque teniendo en cuenta el principio de desafiar el poder, identificar y mitigar sesgos algorítmicos no es suficiente. Es fundamental incorporar una perspectiva crítica desde el inicio del diseño del sistema, en lugar de depender únicamente de auditorías retroactivas para abordar los sesgos.

Las conclusiones apuntan a la necesidad de contar con mejores fuentes de datos y una mayor representación en los conjuntos de datos para garantizar la equidad en los resultados de los modelos. Se destaca la importancia de seleccionar adecuadamente las métricas de equidad y de aplicar técnicas de mitigación de manera cuidadosa, para evitar impactos negativos en grupos específicos. Se destaca que es necesario especificar el contexto sobre el cual se extraen los datos y el proceso de implementación del sistema en general para poder lograr transparencia en dichos sistemas.

Finalmente el trabajo sienta las bases para investigaciones futuras, enfatizando la relevancia de continuar evaluando la equidad en ciencia de datos, especialmente en contextos educativos, con el fin de mejorar los resultados y promover un sistema más equitativo e inclusivo para todos los estudiantes. Para un trabajo futuro, sería valioso realizar un estudio de un caso en producción, lo que permitiría evaluar de manera más concreta los impactos de la implementación de estos sistemas en entornos reales. Además, sería interesante abordar las limitaciones identificadas en el proyecto.

Apéndice A

Lexiland

A.1. Introducción

El proyecto Lexiland [62] [63] examina el desarrollo y la contribución de la conciencia fonológica a las habilidades de lectura temprana en español. El estudio comenzó con niños en su último año de jardín de infantes y continuó siguiendo a estos niños durante su primer y segundo año de escuela primaria. Los niños fueron reclutados de veintiséis escuelas públicas en Montevideo, Uruguay.

Para evaluar la conciencia fonológica en los niños, se utilizó una aplicación para *tablets* llamada Lexiland, con formato de videojuego, diseñada específicamente para el estudio. Los niños realizan tareas que miden distintos indicadores de habilidades de lectura a través de la aplicación. Además, se obtuvieron datos demográficos y socioeconómicos de la Administración Nacional de Educación Pública de Uruguay (ANEP), incluyendo la edad, el género y la educación materna de los niños.

Con base en la información obtenida, se realizó un análisis y se entrenó un modelo de regresión lineal para predecir si el niño presentará dificultades en la lectura en uno o dos años posteriores.

A.2. Conjunto de Datos

La muestra consistió en 388 instancias de estudiantes, donde se clasifica el desempeño de los estudiantes como lectores en dos grupos: lectores típicos y lectores con dificultades de lectura. Para la recolección de los datos que forman estas instancias de estudiantes, fue necesario realizar las evaluaciones usando la aplicación de Lexiland y, posteriormente, unir esta evaluación con los datos de ANEP para lograr instancias que permitan predecir el desempeño de los estudiantes como lectores. El proceso consiste en primero buscar escuelas que puedan realizar esta actividad, luego presentar un planteo a los docentes sobre cómo deben llevar adelante las evaluaciones. Esas evaluaciones se suben a una plataforma donde se unen con datos demográficos de los estudiantes y, posteriormente, pasan por el modelo predictivo. Con base en los resultados, los expertos se reúnen con los docentes responsables de realizar la evaluación para discutir los resultados de la predicción. Debido a la complejidad de la recolección de los datos, en esta instancia solo se cuenta con información de veintiséis escuelas de Montevideo.

Los atributos que se utilizan sobre cada estudiante para la clasificación son los siguientes:

- **Género**: Indica el género binario del niño.
- **Educación de la madre**: Refleja el nivel máximo de educación alcanzado por la madre del niño, con las categorías: secundaria incompleta o menos, secundaria completa y terciario.
- **Edad**: La edad del niño en el momento de la evaluación.

- **Estatus Socioeconómico**: Nivel de Contexto Sociocultural [4]. Se refiere al quintil de la escuela a la que asiste el niño, indicando el nivel desde el más bajo hasta el más alto.

- **Cociente intelectual**: Mide la inteligencia del niño utilizando la Escala de Inteligencia de Wechsler [58].

- **Memoria verbal a corto plazo**: Capacidad del niño para recordar información verbal de manera inmediata después de una breve presentación.

- **Memoria no verbal a corto plazo**: Habilidad para recordar información no verbal o visual-espacial de forma inmediata.

- **Vocabulario**: Mide el conocimiento de palabras del niño y su capacidad para definirlas o reconocer su significado.

- **Rapid Automatized Naming (RAN)**: Evalúa la rapidez con la que el niño puede nombrar en voz alta series de objetos, colores, letras o números familiares.

- **Conocimiento de letras**: Evalúa el conocimiento del niño sobre el alfabeto y su capacidad para reconocer y nombrar letras.

- **Conocimiento fonológico**: Habilidad para manipular sonidos del habla, como identificar y crear rimas, segmentar palabras en sílabas o fonemas.

Los niños fueron evaluados en su desempeño lector mediante la puntuación compuesta, que se calculó como la media aritmética de las puntuaciones z en decodificación, fluidez y comprensión lectora. Aquellos con una puntuación compuesta por debajo del percentil 16 (z = -1.3) fueron categorizados como lectores con dificultades de lectura (n = 64), mientras que aquellos por encima de ese umbral fueron considerados lectores típicos (n = 324).

A.3. Análisis de los Datos

Se comienza realizando un análisis exploratorio de los datos, donde se evalúa la correlación de los atributos con la etiqueta adjudicada. El atributo Etiqueta indica si el estudiante tiene dificultades de lectura o es un lector típico. La matriz de correlación se presenta en la Figura A.1, donde se observa que los atributos que favorecen la asignación de la Etiqueta de Lector Típico son Conocimiento de letras, Educación de la madre, Memoria verbal a corto plazo y Memoria no verbal a corto plazo. Mientras que dentro de los atributos que afectan negativamente se encuentra el valor de RAN. Todos los atributos que se extraen a partir de la evaluación están fuertemente correlacionados.

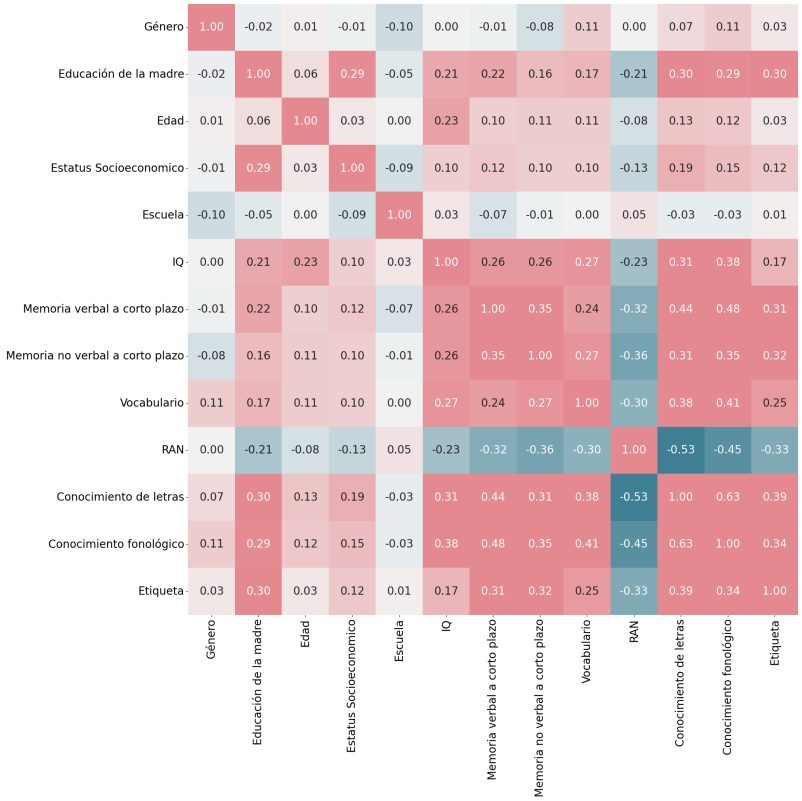


Figura A.1: Matriz de correlación

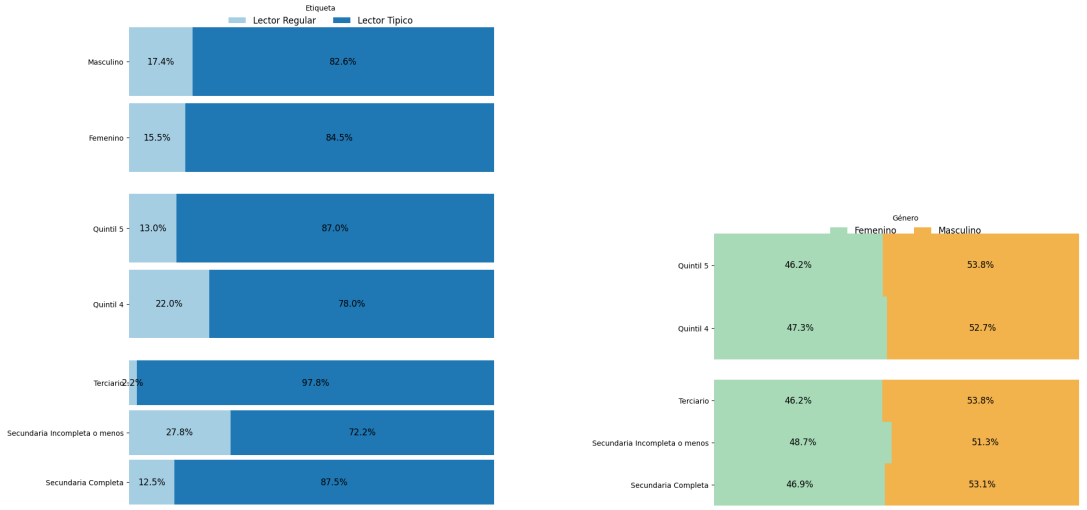
La distribución de los atributos Género, Educación de la madre y Estatus Socioeconómico según la Etiqueta se presenta en la Figura A.2a. En la misma se puede observar un porcentaje alto de estudiantes calificados como lectores típicos.

En el caso del atributo Género, se cuentan con 207 estudiantes de género masculino y 181 de género femenino. Se destaca que no existe la categoría no binario, pero tampoco se especifica si este dato fue recabado o no. Además, la distribución de Lectores típicos y Lectores con dificultades de lectura indica que pertenecer a una clase u otra es independiente del género. Esto es consecuencia de que la Etiqueta se encuentra distribuida de igual manera para ambos géneros.

En el caso del Estatus socioeconómico o quintil de la escuela, el conjunto de datos tiene 238 estudiantes del 4to quintil y 150 estudiantes del 5to quintil. Se observa que no se tomaron datos en escuelas de los quintiles 1, 2 y 3. Esto fue una restricción determinada por la disponibilidad y accesibilidad de las instituciones. Aún así, esta ausencia de datos plantea una limitación en la generalización de los resultados obtenidos. Por otra parte, al analizar el atributo de Educación de la madre, cuando la madre posee un nivel educativo elevado, hay una disminución en el porcentaje de individuos clasificados con dificultades de lectura.

Este atributo se divide en tres categorías según el grado máximo de educación alcanzado a la hora de la inscripción del estudiante al centro educativo, por lo que puede no representar la realidad actual de las familias. La cantidad de instancias de cada categoría tiene la siguiente distribución, 158 estudiantes tienen madre con Secundaria incompleta o menos, 64 con Secundaria completa y 91 con madres que cuentan con educación Terciaria.

En la Figura A.3 se muestra la relación entre el quintil de la escuela y el nivel de estudios de la madre,



(a) Distribución de atributos Género, Estatus socioeconómico, Educación de la madre según el tipo de lector clasificado.

(b) Distribución de Estatus socioeconómico y Educación de la madre según el atributo Género.

Figura A.2: Distribuciones de atributos según etiqueta asignada (a) y género (b).

dos atributos que están fuertemente correlacionados. Se puede observar que en la categoría de madres con estudios terciarios hay una mayor proporción de estudiantes de escuelas del quintil 5.

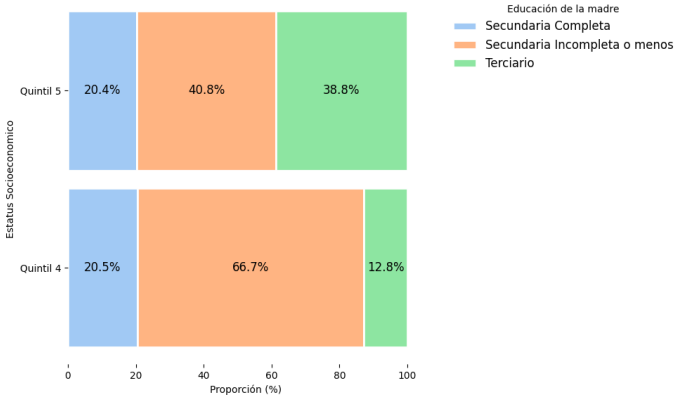


Figura A.3: Distribución de quintil según el atributo Estudio de la madre.

Por último para evaluar como afectan los atributos de manera interseccional se toman los atributos Género, Educación de la madre, y Etiqueta representado en la Figura A.4. Se observa que entre los estudiantes clasificados como lectores típicos, la distribución según género y nivel educativo de la madre es la misma, siendo en su mayoría estudiantes con madre que tiene secundaria completa o más. En el caso de los estudiantes clasificados con bajas habilidades lectoras encontramos mayor porcentaje hombres y a la vez que tienen madre con secundaria incompleta o menos.

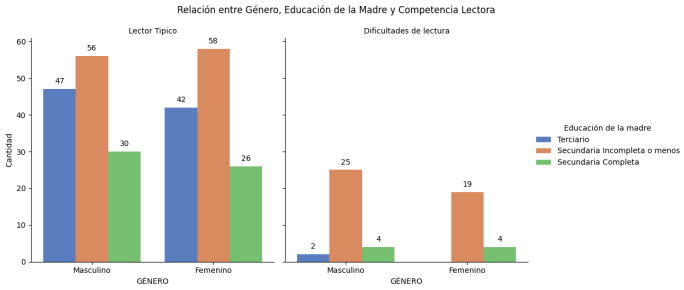


Figura A.4: Clasificación según Educación de la madre y Género

A.4. Modelo

Se entrenaron dos modelos de regresión lineal mediante el método de validación cruzada, el cual establece una relación entre la variable de comprensión lectora y otras variables independientes. El primer modelo, el cual se denomina modelo extendido, tiene los parámetros definidos en la tabla A.1.

Atributo	**Coeficiente**
Constante	−3,40
Edad	0,09
Género	0,47
Estatus socioeconómico	−1,76
Estatus socioeconómico como factor cuadrático	−0,87
IQ	−0,01
Memoria verbal a corto plazo	−0,33
Memoria no verbal	−0,53
Vocabulario	−0,17
RAN	0,32
Conocimiento de las letras	−0,84
Conocimiento fonológico	−0,81

Tabla A.1: Parámetros de modelo extendido.

Dentro de las variables que mayor aporte generan al modelo se encuentran: Memoria no verbal, Conocimiento de letras, Conocimiento fonológico, Género y Estatus socioeconómico.

Posteriormente tomando las variables con coeficientes más altos se realiza un segundo modelo reducido que tiene los parámetros de la Tabla A.2.

Para entrenar y evaluar el modelo, debido a la baja cantidad de datos, se utilizó el método de validación cruzada separando el conjunto en 70 % de instancias para entrenar y 30 % de instancias para validarlo, iterando 1000 veces. Luego se selecciona como modelo, aquel que performe mejor. Los coeficientes utilizados fueron obtenidos en función del área debajo de la curva de ROC (0,88), logrando una especificidad de 76 % para una sensibilidad del 90 %.

Atributo	**Coeficiente**
Constante	−3,20
Estatus socioeconómico	−1,76
Estatus socioeconómico como factor cuadrático	−0,83
Memoria no verbal	−0,61
Conocimiento de las letras	−1,02
Conocimiento fonológico	−1, 17

Tabla A.2: Parámetros de modelo reducido.

A.5. Análisis de Sesgos

Para realizar un análisis de sesgos de este sistema se tuvo una etapa de traducción de lenguaje de programación original R a un lenguaje donde se puedan analizar los sesgos con las herramientas que se tienen, en este caso Python. Dadas las características que presenta la implementación realizar un análisis de sesgos algorítmico sobre este sistema no fue posible. Para evaluar sesgos algorítmicos con las herramientas de análisis de sesgos estudiadas es necesario contar con un conjunto de datos que no haya participado del entrenamiento. Dadas las características de la validación cruzada, la cantidad de iteraciones realizadas de este proceso y la cantidad de instancias reales con las que cuenta el proyecto, no es posible contar con un conjunto de instancias que no haya sido participe de la etapa de entrenamiento. Por lo tanto, no es posible evaluar los sesgos algorítmicos.

A.6. Conclusiones

En este Anexo se estudia el caso de Lexiland, un sistema realizado en Uruguay que proporciona una clasificación de estudiantes según sus habilidades de lectura. Se realizó en principio un análisis de datos exploratorio del que se extraen tres atributos para el análisis que son Género, Educación de la madre y Estatus socioeconómico. De estos atributos se encuentra una correlación grande entre los atributos Educación de la madre y Estatus socioeconómico. Además el porcentaje de madres con secundaria incompleta o menos representa la mayoría de la población. Por otro lado destacar que quienes tienen madre con estudios terciarios pertenecen en su mayoría a quintiles más altos. El comportamiento de estos atributos es importante además porque forman parte directa o indirectamente del modelo propuesto. En particular, el Estatus Socioeconómico es el único atributo que aparece de forma cuadrática en el modelo generando una gran influencia sobre el mismo. Finalmente no fue posible realizar el análisis de sesgos algorítmicos ya que la implementación se hace a través de un modelo que utiliza el método de validación cruzada, dejando al sistema sin un conjunto de prueba. Esta particularidad se debe a la dificultad de obtención de datos, es por esto que además no se tiene información de escuelas de quintiles 1, 2 y 3. Se espera que en etapas futuras del proyecto el mismo sea aplicado a escuelas de todo el país, momento en el cual van a existir más datos para poder realizar futuras investigaciones en este sentido.

Apéndice B

Propuesta de modelización predictiva del rendimiento académico en el corto plazo para estudiantes de Ingeniería

B.1. Introducción

El proyecto *Propuesta de modelización predictiva del rendimiento académico en el corto plazo para estudiantes de Ingeniería* [3] busca predecir el rendimiento de los estudiantes de Facultad de Ingeniería de UdelaR en el primer año. Para ello evalúa el conjunto de instancias de estudiantes bajo distintos modelos de aprendizaje automático para poder encontrar cuál se ajusta mejor al problema. Se busca tener una herramienta para identificar a aquellos estudiantes con posibles dificultades para cursar el primer año lectivo y de esta manera brindarles apoyo para minimizar las consecuencias negativas que puede tener.

B.2. Datos Utilizados

La muestra consistió en 694 datos, de la generación ingresante durante el primer semestre de Facultad de Ingeniería en el año 2016. Se extrae información de los registros de Bedelía obtenidos a la hora de inscripción del estudiante y de la Herramienta Diagnóstica al Ingreso (HDI). La HDI es una prueba escrita de régimen obligatorio para todos los estudiantes que ingresan a la Facultad de Ingeniería de UdelaR. El fin de la prueba es obtener un diagnóstico sobre cada generación y en la cual se miden habilidades en áreas como Física, Matemáticas, Química, Comprensión Lectora, expresión escrita, entre otras. Finalmente se integran los datos provenientes de estos dos sistemas para generar instancias que tienen los siguientes atributos:

- **Sexo**: Catalogado como Masculino (M) o Femenino (F).
- **Institución de origen**: Subsistema educativo de precedencia, puede ser Público, Exterior, Privado o UTU.
- **Lugar de origen**: Puede ser Montevideo, Interior o sin datos.
- **Carrera**: Indica la primer carrera a la que se inscribió el estudiante. (IngCop - Ingeniería en Computación, IngCivil - Ingeniería Civil, , IngElect - Ingeniería Eléctrica, IngIMec - Ingeniería Industrial Mecánica, IngProd - Ingeniería de Producción, Agrim - Agrimensura, IngNaval - Ingeniería Naval, IngAlim - Ingeniería en Alimentos, LicCsAtm - Licenciatura en Ciencias de la Atmósfera).

- **HDIm**: Puntaje en matemática de la HDI, número del 0 al 15.
- **HDIcl**: Puntaje en comprensión lectora de la HDI, número del 0 al 15.
- **HDITotal**: Puntaje total de la HDI, es un número del 0 al 50.
- **Edad al ingreso**: Edad en números decimales del estudiante al ingreso de la Facultad.
- **Aprobado 50p primer semestre**: Es la variable que se pretende predecir. Se clasifica como 1 si al final del primer año tiene más del 50 % de los créditos y 0 en caso contrario.

B.3. Análisis de los Datos

El análisis comienza evaluando cómo se vinculan los distintos atributos del sistema, donde en la Figura B.1 se presenta la matriz de correlación de los datos del modelo. Se resalta la fuerte vinculación entre la aprobación del 50 % de los créditos del primer semestre con el puntaje en matemática alto en la HDI. Dentro de los atributos que son de carácter sociodemográficos, el lugar de origen resulta de correlación negativa respecto a la aprobación de los créditos. A su vez obtener una puntuación alta en la HDI en total esta vinculada directamente con el atributo Sexo del estudiante.

En lo que respecta a la relación más fuerte representada en la matriz de correlación, los puntajes totales de la HDI y su vinculación con la aprobación de la mitad de los créditos en el primer semestre se presentan en la tabla B.1. En la tabla se puede identificar que aquellos estudiantes que obtienen mayores resultados en la HDI, obtienen en su mayoría el 50 % de los créditos del primer semestre. Por lo que obtener un puntaje alto en esta evaluación representa un buen desempeño a futuro para el estudiante. En este caso se toma en conjunto los puntajes en matemática y comprensión lectora.

Total en HDI	**No aprueba 50 % de créditos**	**Aprueba 50 % de créditos**	**Total**
0-10	8	3	11
10-20	90	165	255
20-30	63	226	289
30-40	7	113	120
Más de 40	0	19	19

Tabla B.1: Resultados de estudiantes que aprueban el 50 % de los créditos en el primer semestre respecto al resultado total de la HDI.

En la Figura B.2 se muestra la distribución de los atributos Sexo, Institución de origen, Lugar de origen y Carrera según la etiqueta. El atributo Lugar de origen se divide en tres categorías: Montevideo, Interior y Sin Datos. Se observa que los estudiantes provenientes de Montevideo tienen una mayor tasa de aprobación en comparación con aquellos que no pertenecen a esta categoría. Los que figuran como Sin Datos son en su mayoría los estudiantes que tienen como institución de origen Exterior, teniendo casi la misma tasa de aprobación. En particular, el 80 % de los estudiantes de Montevideo obtiene los créditos del primer semestre, mientras que en el resto de la población solo el 68,75 % logra aprobar los créditos. Estos resultados se resumen en la Tabla B.2.

La distribución del atributo Sexo se presenta en la Tabla B.3, donde se puede observar que el sexo masculino representa el 72,3 % de la población estudiada. Sin embargo si se analiza detenidamente cómo se distribuyen esos estudiantes en la predicción de aprobación en el primer semestre, la cantidad de aprobados sigue la misma distribución independientemente del sexo, obteniendo cerca de 75 % de aprobación en cada caso (Figura B.2). Esto indica que si bien las poblaciones se ven desfavorecidas en cantidad de instancias, el porcentaje de repuestas positivas (Aprobar 50 % de los créditos en primer semestre) es el mismo para ambas poblaciones.

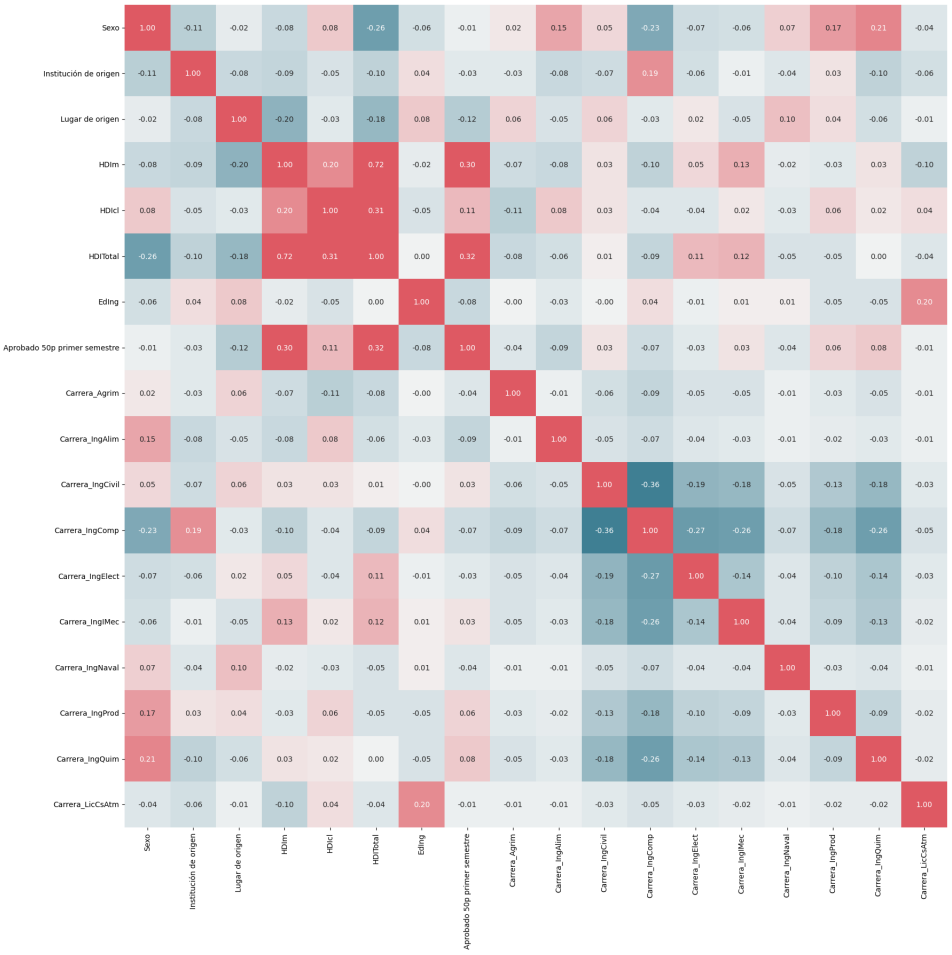


Figura B.1: Matriz de correlación.

Si se analiza la distribución de los estudiantes según su edad, como se muestra en la Figura B.3, se observa que la mayoría ingresan a la facultad a los 17 y 18 años, coincidiendo con la edad típica de finalización de la educación media. La figura muestra los estudiantes hasta los 20 años, que agrupan el 90 % de los datos, aunque también hay estudiantes de hasta 47 años.

B.4. Modelo

El proyecto se enfoca en la implementación de varios modelos de predicción con el objetivo de predecir el desempeño de los estudiantes durante el primer semestre. Los modelos utilizados incluyen: Regresión Logística, Máquinas de Vectores de Soporte (SVM), Random Forest, Clasificador Bayesiano, Árboles de Clasificación, CART, y AdaBoost. En cada iteración, los modelos se ajustan y se obtienen sus respectivas métricas: porcentaje de acierto, sensibilidad, precisión y especificidad.

Luego de entrenar los modelos, se exploraron distintas métricas de ensamblado como se menciona en la

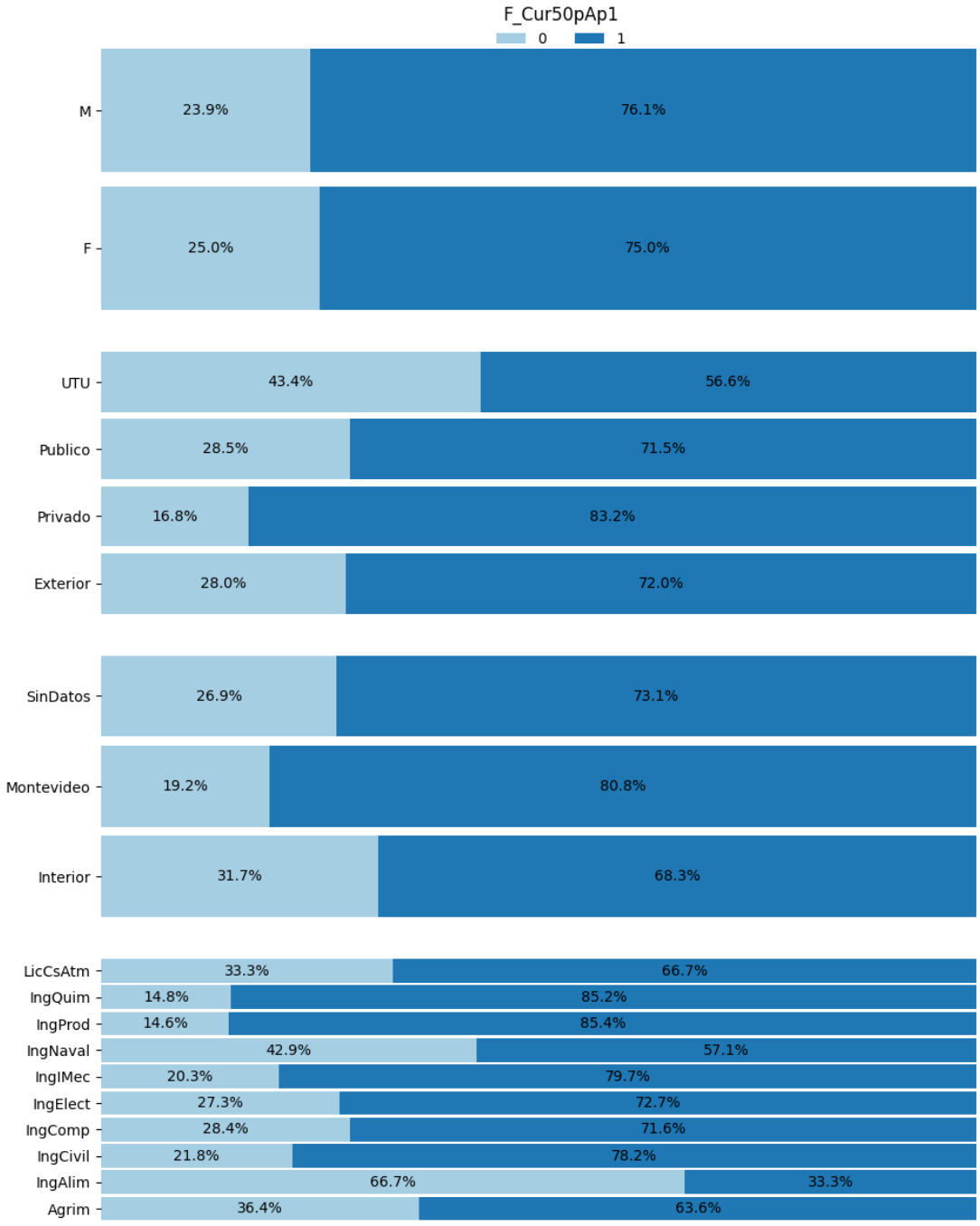


Figura B.2: Distribución de atributos Sexo, Institución de origen, Lugar de origen y Carrera según la etiqueta.

Lugar de origen	No aprueba 50 % de créditos	Aprueba 50 % de créditos	Total
Montevideo	328	78	406
Interior	179	83	288
Sin Datos	19	7	26

Tabla B.2: Resultados de estudiantes que aprueban el 50 % de los créditos en el primer semestre respecto al lugar de origen.

Masculino	Femenino
502	192

Tabla B.3: Distribución de sexo

Sección 2.1.1.4. La técnica de promedio ponderado de las salidas de los modelos obtuvo las métricas de la Tabla B.4.

Métrica	Valor
Porcentaje de acierto	0,800
Precisión	0,775
Especificidad	0,921
Sensibilidad	0,458

Tabla B.4: Métricas obtenidas del promedio ponderado

Los atributos utilizados para el entrenamiento de los modelos son Puntaje en Matemática del HDI (HDIm), Edad al ingreso (EdIng), Lugar de finalización de estudios preuniversitarios (OrLug) y Subsistema de finalización de estudios preuniversitarios (OrTip).

B.5. Análisis de Sesgos

Siguiendo la misma metodología presentada en el Anexo A, se llevó a cabo un proceso de conversión del código de R a Python con el objetivo de reproducir el modelo y realizar un análisis de sesgos. Sin embargo, las características del sistema, en particular el uso de validación cruzada, impidieron la evaluación de sesgos algorítmicos en el modelo. Esta limitación resalta la necesidad de considerar enfoques alternativos.

B.6. Conclusiones

En este anexo se examinó el proyecto titulado *Propuesta de modelización predictiva del rendimiento académico en el corto plazo para estudiantes de Ingeniería*, que tiene como objetivo evaluar el rendimiento de los estudiantes de la Facultad de Ingeniería de UdelaR durante su primer año. A través de este proyecto, se analizaron diversos modelos para caracterizar a los estudiantes y detectar patrones asociados con el bajo rendimiento académico. Se llevó a cabo un análisis exhaustivo de todos los atributos involucrados en el modelo, lo que permitió evidenciar las características de la población estudiantil, predominantemente compuesta por jóvenes de sexo masculino provenientes de Montevideo. Es importante señalar que el sexo no se identificó como un atributo relevante que afecte la obtención final de créditos. Sin embargo, en relación al análisis de sesgos algorítmicos, no fue factible realizar una evaluación debido a las particularidades de los datos recopilados y de la implementación del sistema.

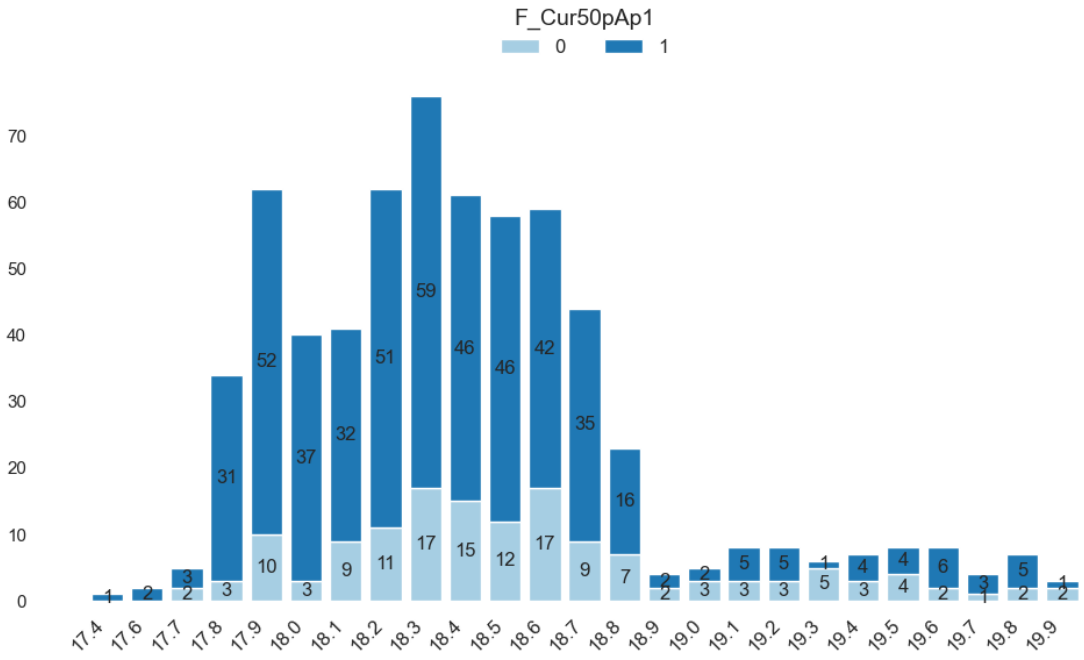


Figura B.3: Distribución de las edades de los estudiantes según la etiqueta.

Referencias

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

[3] Daniel Alessandrini, Paola Bermolen, and Mathias Bourel. *Propuesta de modelización predictiva del rendimiento académico en el corto plazo para estudiantes de Ingeniería.* Tesis de maestría, Universidad de la República (Uruguay). Facultad de Ingeniería, 202o.

[4] ANEP. Definiciones. `https://www.anep.edu.uy/monitor/servlet/definiciones`. Acceso 11/10/2024.

[5] ANEP. Diseño basado en competencias. programa escolar. `https://www.dgeip.edu.uy/documentos/normativa/programaescolar/DBAC-mayo-2017.pdf`, 2017. Acceso 11/10/2024.

[6] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[7] Lucila Berniell, Cecilia Llambí, Romina Paola Durán, Margarita Olivera, Leopoldo Javier Ontivero, and Paula Ortega Grebenc. Alertas tempranas para prevenir el abandono escolar: el caso de la provincia de Mendoza, 2023.

[8] Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[10] Judith Butler and Gender Trouble. Feminism and the Subversion of Identity. *Gender trouble*, 3(1):3–17, 1990.

[11] S Cardozo Politi, A Silveira, and B Fonseca. Detección temprana del riesgo escolar. Predicción de trayectorias de rezago en la educación primaria en Uruguay mediante técnicas de machine learning. *Revista Latinoamericana de Estudios Educativos*, 52(2):30, 2022.

[12] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

[13] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

[14] Kate Crawford. The Trouble with Bias - NIPS 2017 Keynote. Keynote presentation at NeurIPS 2017, 2017. #NIPS2017.

[15] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge, 2013.

[16] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[17] Dirección General de Educación Secundaria. Reglamento de evaluación y promoción de estudiantes. `https://www.dges.edu.uy/sites/default/files/normative/2956.pdf`, 2020. Acceso 11/10/2024.

[18] Centro de Estudios Interdisciplinarios Feministas UdelaR. Observatorio para la Igualdad de Género. `https://www.ceifem.ei.udelar.edu.uy/observatorio/`, 2024. Acceso 11/10/2024.

[19] Gobierno de la Provincia de Buenos Aires. Sistema de Indicadores con perspectiva de género de la provincia de Buenos Aires. `https://sipg.ec.gba.gov.ar/`. Acceso 11/10/2024.

[20] Intendencia de Montevideo. Informe de resultados de la Encuesta a Personas No Binarias 2022. `https://montevideo.gub.uy/sites/default/files/biblioteca/informederesultadosencuestaapersonasnobinarias2022.pdf`, 2022. Acceso 11/10/2024.

[21] Emma et al. DeGraffenreid. DeGraffenreid v. GENERAL MOTORS ASSEMBLY DIV., 1976.

[22] Nicholas Diakopoulos. Algorithmic accountability reporting: On the investigation of black boxes. *Tow Center for Digital Journalism, Columbia University*, 2014.

[23] Catherine D'Ignazio, Isadora Cruxên, Helena Suárez Val, Angeles Martinez Cuba, Mariel García-Montes, Silvana Fumega, Harini Suresh, and Wonyoung So. Feminicide and counterdata production: Activist efforts to monitor and challenge gender-related violence. *Patterns*, 3(7), 2022.

[24] Catherine D'ignazio and Lauren F Klein. *Data Feminism.* MIT press, 2023.

[25] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[27] Equivant. Practitioner's Guide to COMPAS Core. `https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf`, 2019. Acceso 11/10/2024.

[28] Comisión Europea. La protección de datos en la UE. `https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_es`. Acceso 11/10/2024.

[29] Fairlearn Community. Fairlearn: A toolkit for assessing and improving fairness in ai. `https://fairlearn.org/`. Acceso 11/10/2024.

[30] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

[31] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[32] Gender Guesser. Gender Guesser. `https://gender-guesser.com/`. Acceso 11/10/2024.

[33] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc., Sebastopol, CA, USA, 2nd edition, 2019.

[34] Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR, 2021.

[35] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[36] Joel Grus. *Data science from scratch: first principles with python*. O'Reilly Media, 2019.

[37] Noelia Beltramelli Gula, Camila Ferro, María Goñi Mazzitelli, Lorena Etcheverry, and Martín Rocamora. UN CONCEPTO VIAJERO. *Novos Rumos Sociológicos*, 10(18):152–180, 2022.

[38] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[39] Helena Suarez Val. Feminicidio en Uruguay. `https://sites.google.com/view/feminicidiouruguay/`. Acceso 11/10/2024.

[40] IBM. Ai fairness 360. `https://info.watsonadvertising.ibm.com/rs/765-YGI-327/images/AI%20Fairness%20360.pdf`. Acceso 11/10/2024.

[41] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[42] Surya Mattu Julia Angwin, Jeff Larson and ProPublica Lauren Kirchner. Machine Bias. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, 2016. Acceso 11/10/2024.

[43] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

[44] Luiz Antonio Macarini, Cristian Cechinel, Henrique Lemos dos Santos, Xavier Ochoa, Virginia Rodés, Guillermo Ettlin Alonso, and Alén Pérez Casas. Using Data Mining Techniques to Follow Students Trajectories in Secondary Schools of Uruguay. In *2018 XIII Latin American Conference on Learning Technologies (LACLO)*, pages 307–314, 2018.

[45] P. Martínez Ben, Ó. Montañés Soleri, and J. Serralta Gascue. *Modelado de trayectorias académicas de estudiantes universitarios mediante técnicas de analítica de aprendizaje*. Tesis de grado, Universidad de la República (Uruguay). Facultad de Ingeniería, 2021.

[46] Milagros Miceli, Julian Posada, and Tianling Yang. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–14, 2022.

[47] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[48] Rosalind C Morris. *Can the subaltern speak?: Reflections on the history of an idea*. Columbia University Press, 2010.

[49] J Perusia and A Cardini. Sistemas de alerta temprana en la educación secundaria: prevenir el abandon escolar en la era del covid-19. *Documento de Política Pública Nº233. Buenos Aires: CIPPEC*, 2021.

[50] Philip J Piety, Daniel T Hickey, and MJ Bishop. Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 193–202, 2014.

[51] Emanuel Marques Queiroga, Carolina Rodríguez Enríquez, Cristian Cechinel, Alén Perez Casas, Virgínia Rodés Paragarino, Luciana Regina Bencke, and Vinicius Faria Culmant Ramos. Using virtual learning environment data for the development of institutional educational policies. *Applied Sciences*, 11(15):6811, 2021.

[52] Valentim Realinho, Jorge Machado, L Baptista, and MV Martins. Predict students' dropout and academic success. *UCI Machine Learning Repository*, 10:C5MC89, 2021.

[53] Jeevabharathi S. Student Dropout Analysis for School Education. `https://www.kaggle.com/code/jeevabharathis/student-dropout-analysis-for-school-education/notebook`. Acceso 11/10/2024.

[54] scikit-learn Community. scikit-learn. `https://scikit-learn.org/stable/`. Acceso 11/10/2024.

[55] Michael Skirpan and Micha Gorelick. The Authority of "Fair" in Machine Learning. *arXiv preprint arXiv:1706.09976*, 2017.

[56] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

[57] Uruguay. Ley N.º 18.437, 2008.

[58] D. Wechsler, S.C. Gregorio, F. Sánchez-Sánchez, D.A. Águila, P.S. Fernández, and I.F. Pinto. *WPPSI-III, Escala de Inteligencia de Wechsler para Preescolar y Primaria-III: Manual técnico y de interpretación*. TEA Ediciones, S.A., 2009.

[59] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[60] YDataAI. Ydata synthetic. `https://github.com/ydataai/ydata-synthetic`. Acceso 11/10/2024.

[61] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[62] Camila Zugarramurdi, Lucía Fernández, Marie Lallier, Manuel Carreiras, and Juan C Valle-Lisboa. Lexiland: A tablet-based universal screener for reading difficulties in the school context. *Journal of Educational Computing Research*, 60(7):1688–1715, 2022.

[63] Camila Zugarramurdi, Lucía Fernández, Marie Lallier, Juan C Valle-Lisboa, and Manuel Carreiras. Mind the orthography: Revisiting the contribution of prereading phonological awareness to reading acquisition. *Developmental psychology*, 58(6):1003–1016, 2022.

Índice de tablas

Esta página ha sido intencionalmente dejada en blanco.

Índice de figuras

Esta es la última página.
Compilado el lunes 2 diciembre, 2024.
http://iie.fing.edu.uy/