www.aileaders-project.eu

DEMO

Herramienta para auditar variables predictoras en Aprendizaje Automático

CC BY SA

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

DEMO - Herramienta para auditar variables predictoras en Aprendizaje Automático

01  Resumen 3

02  Introducción 4

03  Presentación de herramientas 5

04  Ejecución de la simulación 6

05  Conclusiones 7

06  Referencias 7

01 Resumen

Tipo de REA

Demostración del uso de AEQUITAS en un entorno Google Colab.

Objetivo/Finalidad

Proporcionar al alumnado una herramienta para auditar las variables predictoras de modelos de aprendizaje automático en lo que respecta al sesgo y la equidad. Al centrarse en un escenario de detección de fraudes, el alumnado aprende a evaluar críticamente el rendimiento algorítmico no solo en términos de precisión, sino también en términos de trato equitativo entre los distintos grupos demográficos.

Resultados de aprendizaje esperados

Al final de la demo, el alumnado será capaz de:

01 Aplicar Aequitas para auditar modelos de clasificación (árbol de decisión, *Random Forest*, *FairGBM*) en busca de equidad.

02 Interpretar métricas clave de equidad, como la tasa de falsos positivos (*FPR*), la tasa de falsos descubrimientos (*FDR*) y la paridad estadística.

03 Identificar y explicar las disparidades a nivel de grupo en los resultados de las predicciones.

04 Reflexionar sobre el papel de la auditoría de sesgos y las técnicas de aprendizaje automático justas en contextos de toma de decisiones de alto riesgo.

Palabras clave

- Aprendizaje automático
- Aequitas
- Auditoría
- Sesgos
- Equidad

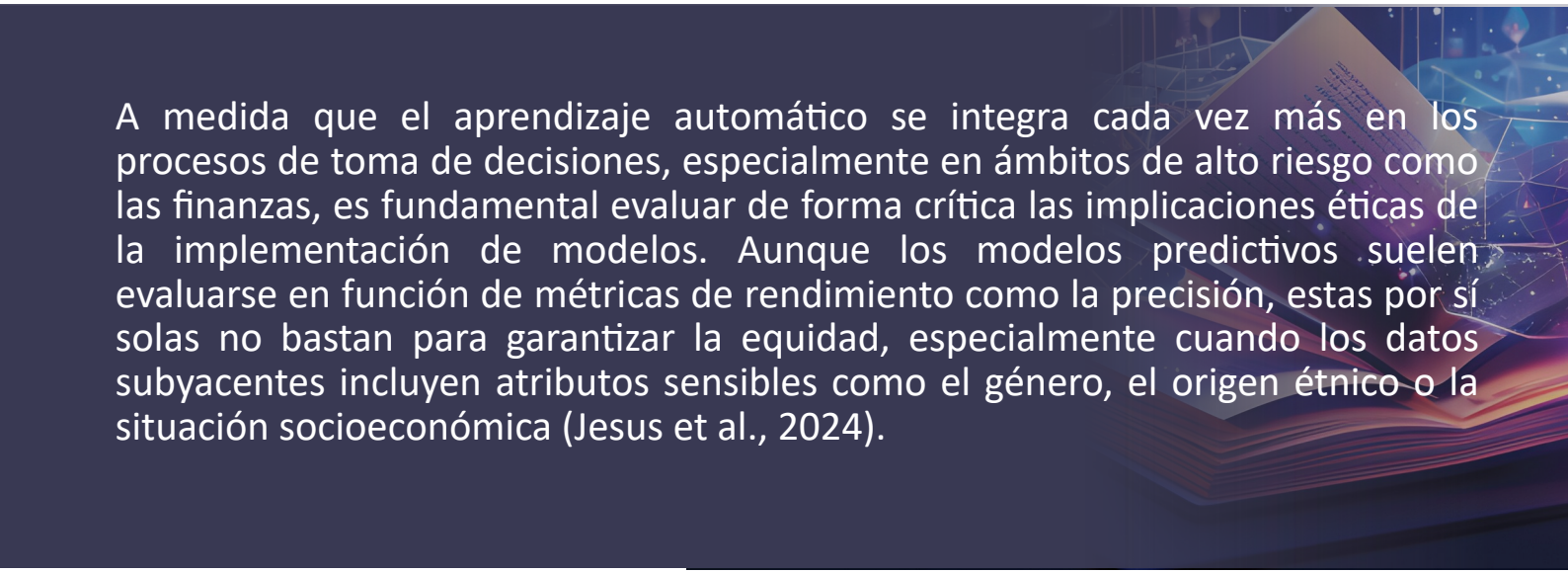
Enfoque metodológico sugerido

Aprendizaje basado en problemas

NOTA

Se requieren conocimientos intermedios de **programación en Python** para comprender y trabajar con el contenido de este REA.

02 Introducción



A medida que el aprendizaje automático se integra cada vez más en los procesos de toma de decisiones, especialmente en ámbitos de alto riesgo como las finanzas, es fundamental evaluar de forma crítica las implicaciones éticas de la implementación de modelos. Aunque los modelos predictivos suelen evaluarse en función de métricas de rendimiento como la precisión, estas por sí solas no bastan para garantizar la equidad, especialmente cuando los datos subyacentes incluyen atributos sensibles como el género, el origen étnico o la situación socioeconómica (Jesus et al., 2024).

Esta actividad basada en una demo invita al alumnado a familiarizarse con Aequitas, un conjunto de herramientas de auditoría de código abierto, a través de la exploración práctica de un escenario realista de detección de fraudes. El objetivo es doble: profundizar en la comprensión del alumnado sobre el sesgo algorítmico y familiarizarlo con herramientas y técnicas prácticas que apoyan el desarrollo de sistemas de aprendizaje automático más transparentes y responsables.

Aequitas proporciona un marco integral para evaluar la equidad en los modelos de clasificación mediante el examen de la distribución de los resultados predictivos entre los grupos demográficos. Ofrece una serie de métricas de equidad, entre las que se incluyen la tasa de falsos positivos (FPR), la tasa de falsos descubrimientos (FDR) y la paridad estadística, que ayudan a descubrir disparidades en el rendimiento de los modelos que pueden no ser visibles mediante los métodos de evaluación convencionales. Incluso cuando se entrenan con conjuntos de datos desequilibrados o sesgados, los modelos pueden ser auditados para evaluar si tratan a los diferentes subgrupos de manera equitativa.

Al interactuar directamente con Aequitas en este contexto aplicado, el alumnado adquirirá tanto habilidades técnicas en la auditoría de equidad como una conciencia ética más amplia de los retos inherentes al despliegue del aprendizaje automático en ámbitos en los que las consecuencias del sesgo pueden ser especialmente graves.

03 Presentación de herramientas

El escenario consiste en auditar un **modelo de clasificación binaria** entrenado para detectar **fraudes en cuentas bancarias.**

Los sistemas de detección de fraudes suelen presentar:

- **Desequilibrio grave entre clases** (los casos de fraude son poco frecuentes);
- **Altos riesgos** (una clasificación errónea puede perjudicar a personas o instituciones);
- **Sesgos ocultos** (las características sensibles pueden correlacionarse con los resultados de las etiquetas).

*La simulación utiliza el **conjunto de datos Bank Account Fraud (BAF)**, un conjunto de datos sintéticos a gran escala que preserva la privacidad y reproduce los patrones reales del fraude bancario.*

Las características principales incluyen:

01 Variantes que simulan sesgos de muestreo, deriva temporal y desequilibrio de características

02 Atributos demográficos que permiten el análisis de la equidad.

03 Fuerte desequilibrio de clases para mayor realismo.

Estas condiciones permiten realizar experimentos significativos sobre cómo se comportan las métricas de equidad bajo diferentes supuestos de modelos y configuraciones de datos.

04 Ejecución de la simulación

01 Acceda al cuaderno de simulación

Vaya a https://tinyurl.com/4b9u9hun

02 Ejecutar todo el código

Ejecute el cuaderno por completo para cargar el conjunto de datos, entrenar un modelo de clasificación binaria y generar predicciones. Para ello, haga clic en el botón de reproducción situado en la parte superior izquierda de cada celda.

03 Realice una auditoría de equidad con Aequitas

- Utilice Aequitas para generar un informe de equidad
- Céntrese en métricas como FPR, FDR y paridad estadística
- Identifique qué grupos reciben un trato injusto en las predicciones del modelo

04 Compare los resultados e interprete las métricas

Revise las disparidades de equidad entre los grupos. Compare las métricas de rendimiento (por ejemplo, la precisión) con las métricas de equidad para evaluar las compensaciones.

05 Reflexione y debata

- ¿Qué patrones de injusticia se observaron?
- ¿Cómo podrían afectar estos resultados a personas reales?
- ¿Cómo se pueden incorporar estas herramientas en el proceso de aprendizaje automático para mejorar los resultados éticos?

Detalles del proceso

Este proceso consta de varios pasos secuenciales diseñados para preparar los datos, crear modelos predictivos y evaluarlos en términos de rendimiento y equidad.
Las etapas clave son las siguientes:

01 **Carga de datos**: primero se carga el conjunto de datos desde una fuente específica (por ejemplo, un archivo CSV o una base de datos). Se espera que contenga tanto variables de características como uno o más atributos sensibles (por ejemplo, género, raza) necesarios para la evaluación de la equidad.

02 **Preprocesamiento**: los datos se someten a varios pasos de preprocesamiento para garantizar su calidad y coherencia (imputación y normalización).

03 **Modelización:** se entrenan y evalúan múltiples modelos de aprendizaje automático. Estos modelos pueden incluir tanto algoritmos que no tienen en cuenta la equidad (estándar) como enfoques que sí la tienen en cuenta y que incorporan técnicas de mitigación del sesgo durante el entrenamiento o el posprocesamiento.

04 **Evaluación de la equidad con Aequitas**: los modelos entrenados se evalúan no solo en términos de precisión y otras métricas de rendimiento, sino también en cuanto a equidad utilizando el kit de herramientas Aequitas.

05 Conclusión

Esta demo muestra que incluso los modelos precisos pueden dar lugar a **resultados injustos** cuando se implementan sin una auditoría de equidad.

Con Aequitas, el alumnado descubre cómo **el sesgo puede persistir** en entornos de clasificación binaria y cómo diferentes grupos demográficos pueden experimentar diferentes índices de clasificación errónea. Al trabajar directamente con métricas de equidad y realizar auditorías reales, los estudiantes desarrollan competencias técnicas y conciencia ética. Y lo que es más importante, aprenden **que la detección de sesgos no es un complemento,** sino una parte esencial de la creación **de sistemas de IA fiables**, especialmente en ámbitos como las finanzas, donde las predicciones de los modelos tienen graves consecuencias.

06 Referencias

- NJesus, S., Saleiro, P., e Silva, I. O., Jorge, B. M., Ribeiro, R. P., Gama, J., ... & Ghani, R. (2024). Aequitas flow: Streamlining fair ml experimentation. Journal of Machine Learning Research, 25(354), 1-7.

www.aileaders-project.eu

CC BY SA

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.