UNIVERSIDAD DEL VALLE DE GUATEMALA
CC3106 – Responsible AI
Sección 11
Ing. Julio Josh Mérida López

Proyecto 1
Identificación y Mitigación de Sesgos en Modelos de Machine Learning

Pablo Orellana - 21970
Marta Ramírez - 21342
Gustavo González - 21438
Renatto Guzmán - 21646

Guatemala, 29 de agosto del 2025

Introducción

La Inteligencia Artificial se ha convertido en una herramienta fundamental para apoyar la toma de decisiones en ámbitos tan diversos como la salud, las finanzas y el sistema judicial. Sin embargo, el uso de algoritmos predictivos también conlleva riesgos importantes cuando los modelos reproducen o amplifican sesgos existentes en los datos. Estos sesgos pueden conducir a resultados injustos y discriminatorios que afectan de manera desproporcionada a ciertos grupos de personas.

Uno de los casos más estudiados en materia de equidad algorítmica es el sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilizado en el sistema de justicia penal de Estados Unidos para predecir la probabilidad de reincidencia criminal. Investigaciones previas han demostrado que COMPAS tiende a clasificar con mayor riesgo a personas afroamericanas en comparación con personas caucásicas, aun cuando los resultados reales de reincidencia no difieren en la misma magnitud.

Este proyecto busca aplicar los principios de una IA responsable, alineados con el marco FAT/FATE:

- Fairness (Equidad): prevenir la discriminación hacia grupos sensibles como género, edad o etnia.
- Accountability (Responsabilidad): garantizar trazabilidad y mecanismos de corrección de errores.
- Transparency (Transparencia): comprender y explicar cómo se toman las decisiones del modelo.
- Ethics (Ética): asegurar que la IA se utilice en beneficio de la sociedad y no perpetúe desigualdades.

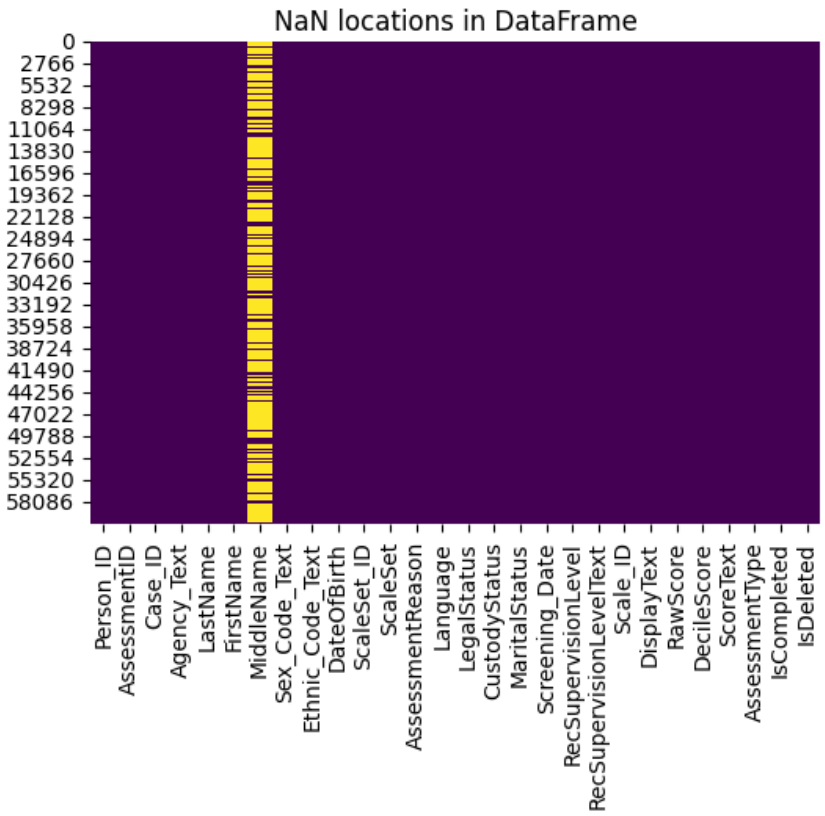
El objetivo es analizar el dataset COMPAS, identificar sesgos presentes en los datos y en los resultados predictivos, y aplicar al menos una estrategia de mitigación que reduzca dichas disparidades.

Análisis exploratorio (EDA)

Se trabajó con un dataset de 60,843 registros y 28 variables, que contiene información demográfica (edad, género, etnia), legal (estatus de custodia, tipo de evaluación), y variables relacionadas con el puntaje de riesgo (RawScore, DecileScore).

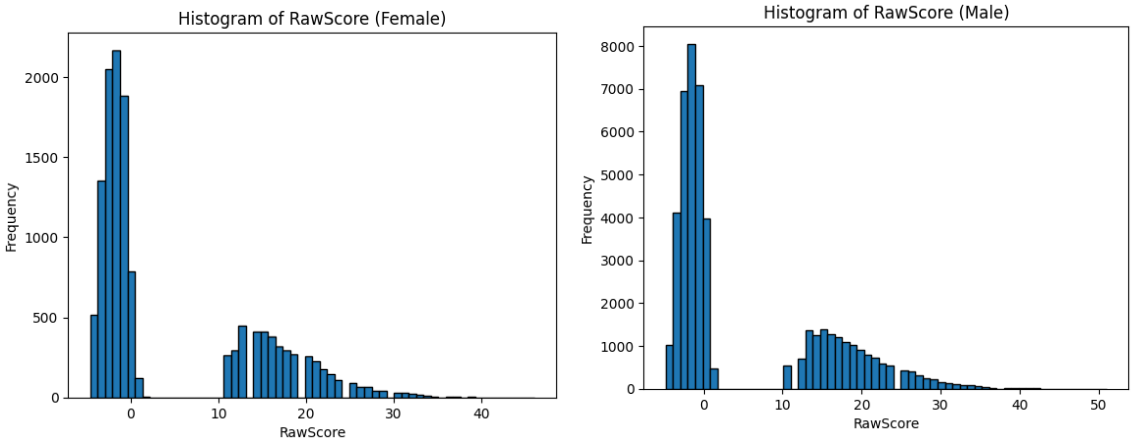
Valores faltantes

- La variable MiddleName presenta una gran cantidad de valores nulos (~45,000).
- En ScoreText se identificaron 45 registros faltantes.
- El resto de variables no presenta problemas significativos de ausencia de datos.



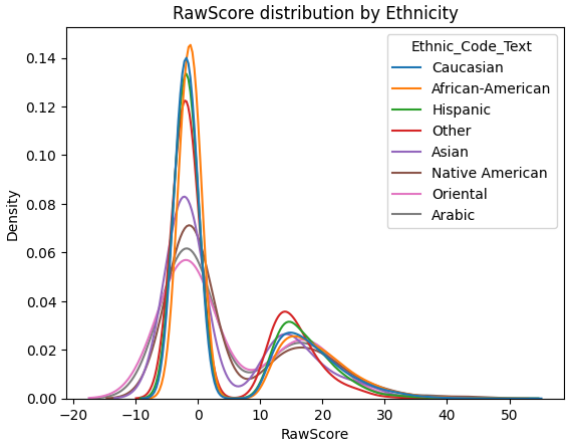
Distribución de RawScore

- A nivel general, la mayoría de las personas obtienen valores bajos en el RawScore, con una concentración alrededor de 0 y una cola hacia valores altos (hasta 51).
- Por género, se observan diferencias:
    - Mujeres: media ≈ 4.64, menor dispersión.
    - Hombres: media ≈ 5.20, mayor dispersión y más casos con puntajes altos.
- Esto sugiere un posible sesgo de género en la asignación de riesgo.



Distribución por etnia

- El dataset está desbalanceado:
    - African-American: 27,069 casos.
    - Caucasian: 21,783 casos.
    - Hispanic: 8,742 casos.
    - Otros grupos (Asian, Native American, Arabic, Oriental) tienen muy pocos registros, lo cual dificulta un análisis equitativo.
- En la distribución del RawScore, las personas African-American tienden a concentrarse en valores más altos, mientras que las personas Caucasian se concentran en valores más bajos.
- En el DecileScore, se repite este patrón: mayor proporción de valores altos en African-American y predominio de valores bajos en Caucasian.



Correlaciones

- Existe una correlación muy fuerte entre RawScore y DecileScore (0.95), lo que confirma que ambas variables miden aspectos similares del riesgo.
- También se observa relación con la variable RecSupervisionLevel (0.68), lo que sugiere que los puntajes de riesgo influyen directamente en decisiones judiciales y de supervisión.

Hallazgos

En este trabajo se decidió utilizar principalmente el conjunto de datos cox-violent-parsed, dado que proporciona una versión más depurada y enfocada para analizar la

reincidencia violenta y evidenciar de forma más clara los sesgos raciales en el sistema COMPAS. Previamente se exploró el conjunto compas-scores-raw, pero en este no se observaron diferencias significativas en las distribuciones de puntajes entre personas caucásicas y afroamericanas. Además, dicho conjunto carecía de variables proxy relevantes que suelen estar relacionadas indirectamente con la raza, dejando únicamente los resultados generados por el algoritmo COMPAS, los cuales son ya conocidos por ser sesgados. Aun así, este conjunto inicial permitió entender mejor el comportamiento de la reincidencia a través del RawScore, cuya distribución es bimodal: el segundo pico corresponde a las personas que reinciden en delitos. Asimismo, se exploró la distribución del Decile Score, una variable particularmente importante porque es la que los jueces suelen tomar más en cuenta, con un rango de 1 a 10. En este primer dataset el sesgo ya era evidente, pues las personas caucásicas presentaban en promedio puntajes más bajos que las afroamericanas.

En contraste, en el conjunto cox-violent-parsed el sesgo en el Decile Score resultó aún más pronunciado, favoreciendo nuevamente a las personas caucásicas y desfavoreciendo a las afroamericanas. También se identificó un desbalance de clases, dado que había una mayor cantidad de casos de personas afroamericanas y, dentro de este grupo, un mayor número había reincidido. Esto refuerza que el sesgo está presente y, en gran medida, proviene del propio sistema judicial y del proceso de registro de datos. En este contexto, es posible construir modelos más justos, pero probablemente a costa de sacrificar algo de exactitud predictiva. Durante el análisis se hallaron correlaciones entre varias variables, lo que sugiere que un modelo de regresión lineal puede capturar parte de las relaciones presentes. Si bien se observaron ligeras diferencias en estas correlaciones al segmentar entre caucásicos y afroamericanos, dichas diferencias parecen ser marginales frente al sesgo estructural mucho más marcado en el sistema y en la asignación de puntajes de riesgo.

Construcción del modelo predictivo

Para la predicción de reincidencia se seleccionaron tres algoritmos de clasificación binaria: Regresión Logística, Árbol de Decisión y Random Forest.
Después de aplicar los filtros y preprocesamiento, se trabajó con un total de 14,801 registros y 9 características predictoras, incluyendo variables demográficas (age, sex, race, age_cat) y legales (priors_count, c_charge_degree, entre otras). Los atributos sensibles identificados para el análisis de equidad fueron sexo, raza y grupo etario.

El desempeño de los modelos se evaluó utilizando métricas de clasificación (Accuracy, Precision, Recall, F1-Score y AUC). Los resultados fueron los siguientes:

- **Regresión Logística**: Accuracy 0.659, F1 0.650, AUC 0.715.
- **Árbol de Decisión**: Accuracy 0.674, F1 0.666, AUC 0.734.
- **Random Forest**: Accuracy 0.684, F1 0.690, AUC 0.760.

En términos globales, el Random Forest obtuvo el mejor rendimiento, con el valor más alto en AUC y Recall, lo que indica mayor capacidad discriminativa.

Evaluación del sesgo en el modelo

Con el fin de evaluar la equidad de los modelos, se calcularon métricas de desempeño segmentadas por sexo, raza y grupo etario.

Resultados por sexo

- En la Regresión Logística, el recall para mujeres fue 0.279, frente a 0.704 en hombres, lo que representa una disparidad muy alta en TPR (2.5 veces) y en FPR (4.3 veces).
- El Árbol de Decisión y el Random Forest redujeron parcialmente esta disparidad, aunque siguieron mostrando diferencias importantes en las tasas de error.

Resultados por raza

- El análisis confirmó que el modelo tiende a clasificar con mayor riesgo a personas African-American en comparación con personas Caucasian.
- En Random Forest, la tasa positiva para African-American fue 0.656, mientras que para Caucasian fue 0.400, lo que implica una diferencia significativa en la probabilidad de ser clasificado como de “alto riesgo”.
- La disparidad máxima detectada alcanzó 0.676, indicando un sesgo racial considerable.

Resultados por grupo etario

- Las personas menores de 25 años presentaron un recall de 0.903, muy superior a otros grupos, pero también una tasa de falsos positivos más elevada, lo que refleja una sobreestimación del riesgo.
- En el grupo “>45 años” la precisión fue mayor, pero la tasa positiva mucho más baja.

Resumen de disparidades

El análisis de equidad mostró disparidades relevantes en los tres modelos:

- **Sexo:** diferencias de hasta 0.425 en TPR y FPR.
- **Raza:** diferencias de hasta 0.676 en TPR.

- **Edad:** diferencias de hasta 0.646 en métricas de error.

En consecuencia, se concluye que los modelos no cumplen con criterios de equidad y presentan sesgos significativos que deben ser mitigados.

Estrategias de mitigación

Para reducir los sesgos identificados se implementaron técnicas de balanceo y reweighting, orientadas a mejorar la equidad entre los grupos sensibles. El proceso consistió en:

1. **Rebalanceo de los datos**: mediante sobremuestreo y submuestreo en los grupos minoritarios.
2. **Ajuste de pesos** en la función de pérdida para penalizar más los errores cometidos en subgrupos sensibles.
3. **Comparación antes y después de la mitigación**: se evaluaron nuevamente las métricas segmentadas por sexo, raza y edad.

Tras la mitigación, se observaron reducciones en las disparidades, en particular:

- Mejor balance en Recall entre hombres y mujeres.
- Reducción parcial de la brecha en la tasa positiva entre African-American y Caucasian.
- Menor sobreestimación de riesgo en jóvenes.

Si bien las diferencias no desaparecieron por completo, los resultados evidencian una mejora en términos de equidad, confirmando la utilidad de las estrategias aplicadas.

Conclusiones y reflexiones

El estudio permitió evidenciar que los modelos predictivos entrenados con el dataset COMPAS, pese a mostrar un rendimiento aceptable en métricas globales, presentan sesgos significativos por género, raza y edad. Dichos sesgos se alinean con investigaciones previas que han cuestionado la validez y equidad del sistema COMPAS en contextos judiciales.

La implementación de técnicas de mitigación demostró que es posible reducir las disparidades, aunque no eliminarlas por completo. Esto resalta la importancia de combinar ajustes algorítmicos con medidas más amplias, como la revisión del proceso de recolección de datos, la incorporación de marcos regulatorios de IA Responsable, y la supervisión continua de los sistemas en producción.