

CONFERENCIAS DE POSGRADO – UNIVERSIDAD COMPLUTENSE DE MADRID

Abordando el desbalance de datos y la equidad en modelos de clasificación de Machine Learning para COVID-19

Andreea M. Oprescu (DTE-US)

aoprescu@us.es

11 Junio 2025



Trayectoria

Académica

- 2017 - Graduada en Ingeniería Informática – Tecnologías Informáticas (Universidad de Sevilla)
- 2020 - Máster en Ingeniería Informática (Universidad de Sevilla)
- 2023 - Doctora en Ingeniería Informática (Universidad de Sevilla)

Profesional

Técnica de apoyo a la investigación en el Vicerrectorado de Investigación de la Universidad de Sevilla (Unidad de Bibliometría), programa Empleo Joven
Desarrolladora DevOps en Red Bee Media (Ericsson)
Profesora Sustituta Interina adscrita al departamento de Tecnología Electrónica, Universidad de Sevilla

[Google Scholar](#) [Perfil de Prisma](#)

Los modelos deben ser “buenos” para todos

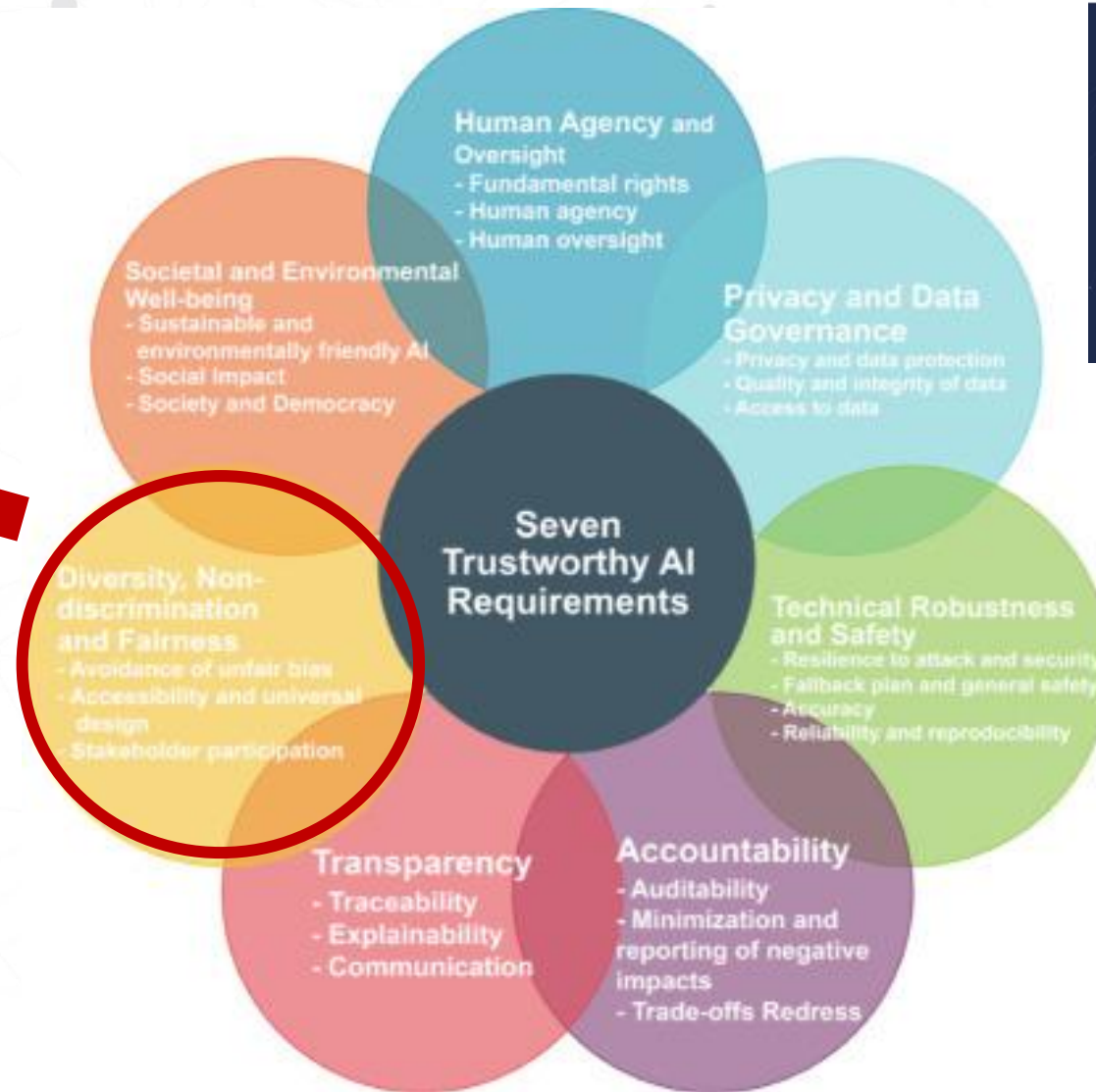
Sesgos en los datos y en los algoritmos

En los datos:

- Análisis de representatividad
- Análisis de sesgos históricos

En los modelos:

- Evaluarlos con métricas adecuadas
- Asegurar que las tasas de error del modelo sean similares entre los diferentes grupos
 - Géneros
 - Etnias
 - Edad
 - Estado civil



Definición del problema



Database Contributor Review

COVID Data for Shared Learning (CDSL): A comprehensive, multimodal COVID-19 dataset from HM Hospitals

Álvaro Ritoré , Andreea M Oprescu , Alberto Estirado Bronchalo , Miguel Ángel Armengol de la Hoz 

Published: Oct. 25, 2024. Version: 1.0.0

<https://physionet.org/>

- Datos clínicos (~1500 muestras)
- Variables categóricas y numéricas
- Variable objetivo mortalidad – problema de clasificación binaria
- Aprendizaje supervisado
- Tecnologías: Python, modelos de ML de Scikit-learn
- Datos desbalanceados (*imbalanced dataset*)

- Cuando los datos no hacen más que darte dolores de cabeza...



● Cuando los datos no hacen más que darte dolores de cabeza...

~~¿caso de
éxito?~~

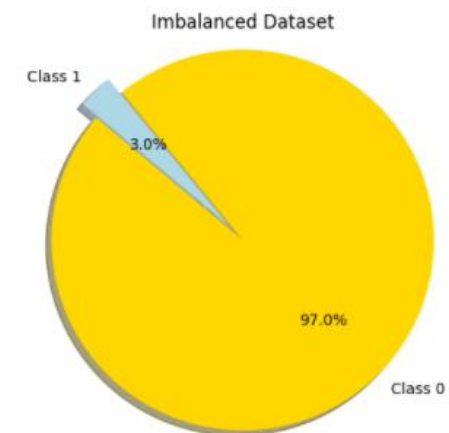
**Caso de
APRENDIZAJE**

Desbalanceo de datos

Definición

Distribución desequilibrada de los valores de la variable objetivo.

En clasificación binaria: clase mayoritaria vs clase minoritaria



Desbalanceo de datos

Definición

¿Cómo medir el desbalanceo?

Porcentaje de datos que pertenecen a la clase minoritaria	Grado de desequilibrio
Entre el 20 y el 40% del conjunto de datos	Leve
Entre el 1 y el 20% del conjunto de datos	Moderado
Menos del 1% del conjunto de datos	Extremo

Fuente: Google, Curso intensivo de aprendizaje automático (2025).

$\text{Imbalanced Ratio} = \text{N}^\circ \text{ de muestras de la clase mayoritaria} / \text{N}^\circ \text{ de muestras de la clase minoritaria}$

Auditoría de equidad con Aequitas

Las métricas globales de rendimiento pueden ocultar sesgos aprendidos por los modelos



[Esta foto](#) de Autor desconocido está bajo licencia [CC BY-NC-ND](#)

Buen valor de F1 pero sistemáticamente injusto con un subgrupo demográfico específico

Auditoría de equidad con Aequitas

¿Cómo utilizar la herramienta?

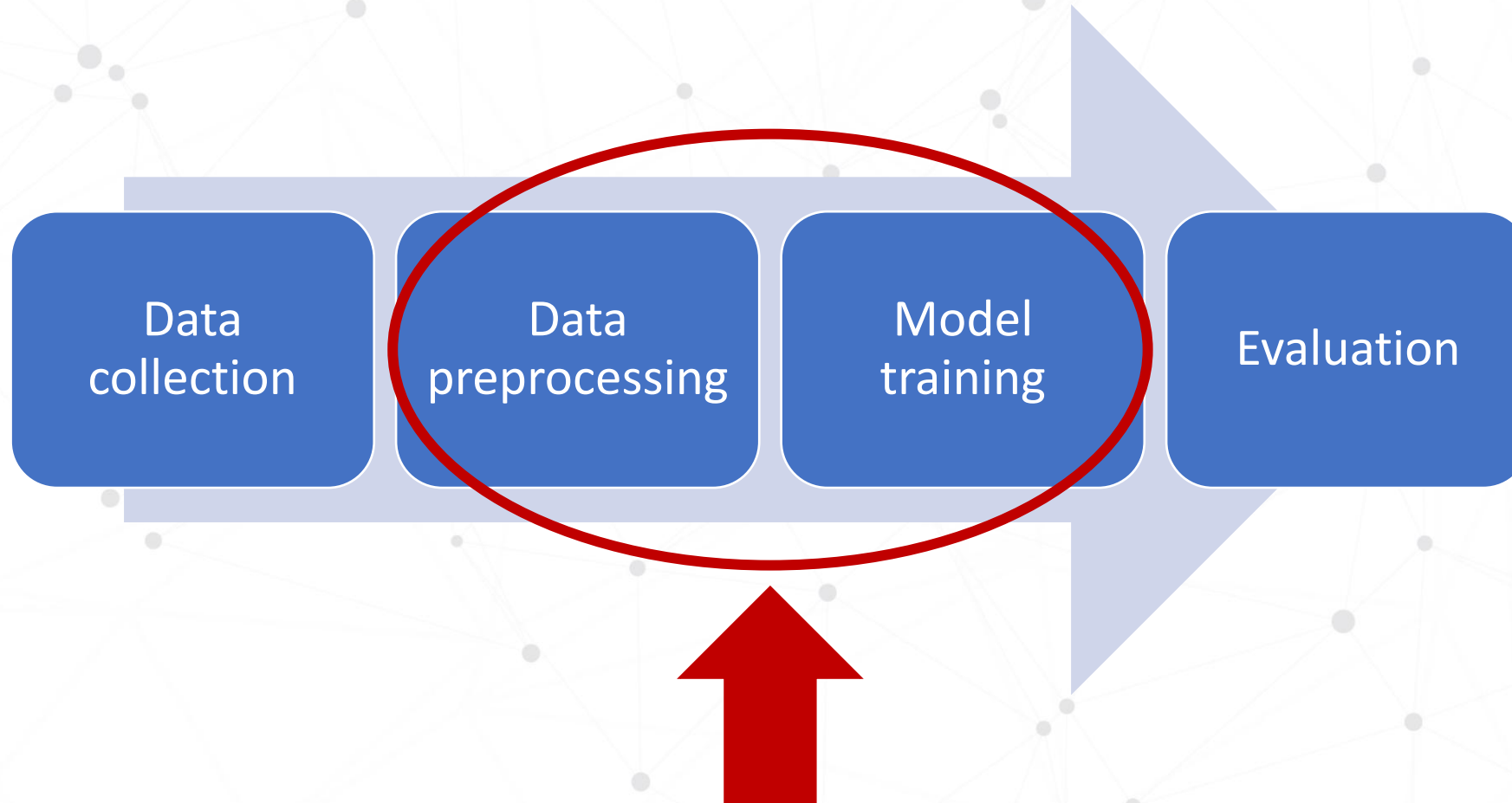
- Identificar los grupos: Separar las predicciones del modelo según los atributos sensibles definidos
- Calcular las métricas de aequitas para cada subgrupo
- Medir la disparidad, comparando las métricas entre grupos

aequitas_df

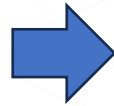
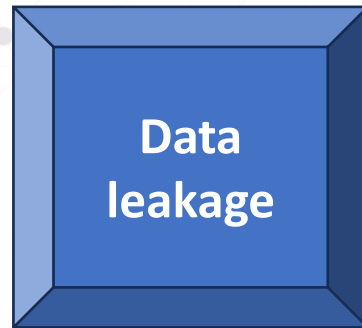
✓ 0.0s  Open 'aequitas_df' in Data Wrangler

#	score	label_value	sex	age_group
1770	0	0	Mujer	80+
1635	0	0	Hombre	65-79
124	0	0	Hombre	18-49
927	0	0	Mujer	65-79
1910	0	0	Hombre	50-64
948	0	0	Mujer	50-64
2110	0	0	Hombre	18-49
1335	0	0	Mujer	65-79
1679	0	0	Hombre	50-64
519	0	0	Hombre	65-79

Back to the basics



Back to the basics



Filtración involuntaria de información
del conjunto de prueba al conjunto de entrenamiento

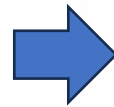
¿Cuándo ocurre?



Preprocesamiento de los datos PREVIO
a la división de los conjuntos

Back to the basics

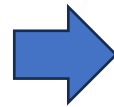
¿Cómo voy a
evaluar mi
modelo?



	Alta Precisión (Pocas Falsas Alarmas)	Baja Precisión (Muchas Falsas Alarmas)
Alta Sensibilidad (Detecta casi todos los casos reales)		
Baja Sensibilidad (Pasa por alto muchos casos reales)		

Back to the basics

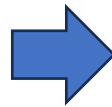
¿Cómo voy a
evaluar mi
modelo?



	Alta Precisión (Pocas Falsas Alarmas)	Baja Precisión (Muchas Falsas Alarmas)
Alta Sensibilidad (Detecta casi todos los casos reales)		Modelo Alarmista Bueno para no pasar por alto a nadie, pero genera mucho ruido con falsas alarmas.
Baja Sensibilidad (Pasa por alto muchos casos reales)		

Back to the basics

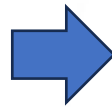
¿Cómo voy a
evaluar mi
modelo?



	Alta Precisión (Pocas Falsas Alarmas)	Baja Precisión (Muchas Falsas Alarmas)
Alta Sensibilidad (Detecta casi todos los casos reales)		Modelo Alarmista Bueno para no pasar por alto a nadie, pero genera mucho ruido con falsas alarmas.
Baja Sensibilidad (Pasa por alto muchos casos reales)	Modelo Cauteloso Es muy seguro cuando predice un caso, pero se le escapan muchos.	

Back to the basics

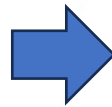
¿Cómo voy a
evaluar mi
modelo?



	Alta Precisión (Pocas Falsas Alarmas)	Baja Precisión (Muchas Falsas Alarmas)
Alta Sensibilidad (Detecta casi todos los casos reales)		Modelo Alarmista Bueno para no pasar por alto a nadie, pero genera mucho ruido con falsas alarmas.
Baja Sensibilidad (Pasa por alto muchos casos reales)	Modelo Cauteloso Es muy seguro cuando predice un caso, pero se le escapan muchos.	Modelo Inútil Ni detecta bien los casos ni es fiable cuando lo hace.

Back to the basics

¿Cómo voy a
evaluar mi
modelo?



	Alta Precisión (Pocas Falsas Alarmas)	Baja Precisión (Muchas Falsas Alarmas)
Alta Sensibilidad (Detecta casi todos los casos reales)	Modelo Ideal Fiable y exhaustivo. Encuentra a casi todos los que debe y rara vez se equivoca al hacerlo.	Modelo Alarmista Bueno para no pasar por alto a nadie, pero genera mucho ruido con falsas alarmas.
Baja Sensibilidad (Pasa por alto muchos casos reales)	Modelo Cauteloso Es muy seguro cuando predice un caso, pero se le escapan muchos.	Modelo Inútil Ni detecta bien los casos ni es fiable cuando lo hace.

Desbalanceo de datos

Posibles soluciones “más conocidas”

Oversampling

- Aumentar artificialmente el número de muestras de la clase minoritaria hasta que esté balanceada con la mayoritaria.

Undersampling

- Reducir aleatoriamente el número de muestras de la clase mayoritaria hasta que esté balanceada con la minoritaria.

Métodos para abordar el desbalanceo

Métricas adecuadas al problema



Optimización del modelo



Optimización del umbral



Medición de cuánto de bueno es el modelo

F1 score

Métodos para abordar el desbalanceo

Muestreo estratificado (*Stratified sampling*)



Estratificar el conjunto de datos

Al hacer la división entre conjuntos de entrenamiento y test

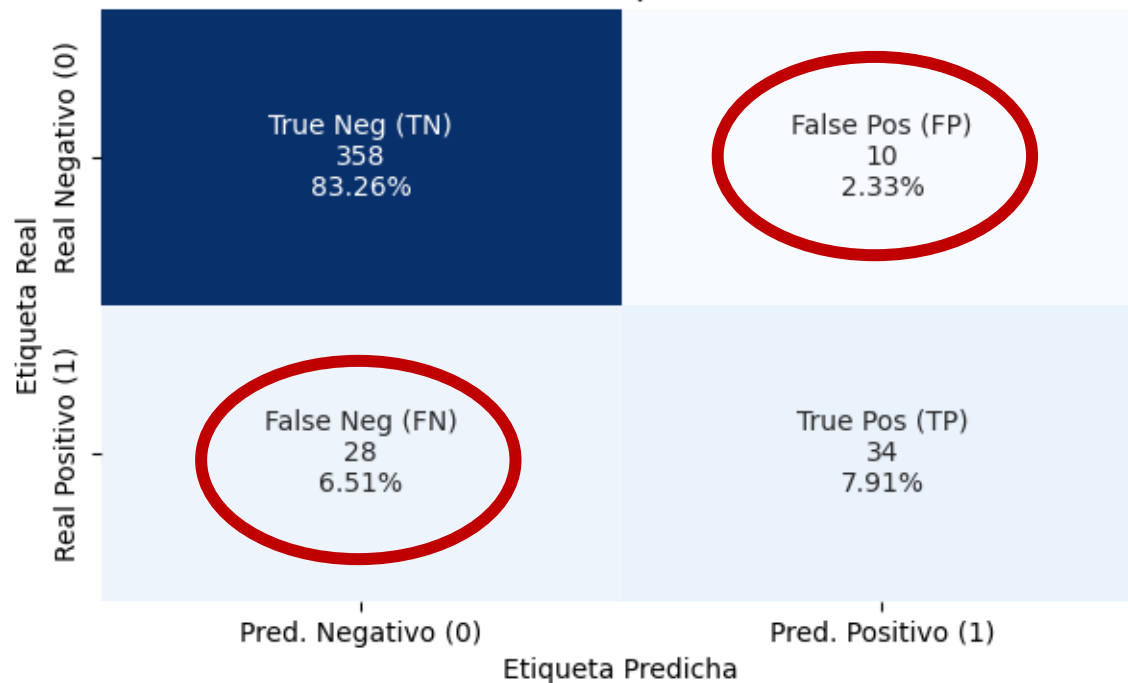
Al hacer la validación cruzada, para mantener en cada fold la proporción de clases del conjunto de prueba

Métodos para abordar el desbalanceo

Muestreo estratificado (*Stratified sampling*)



Matriz de Confusión para XGBClassifier

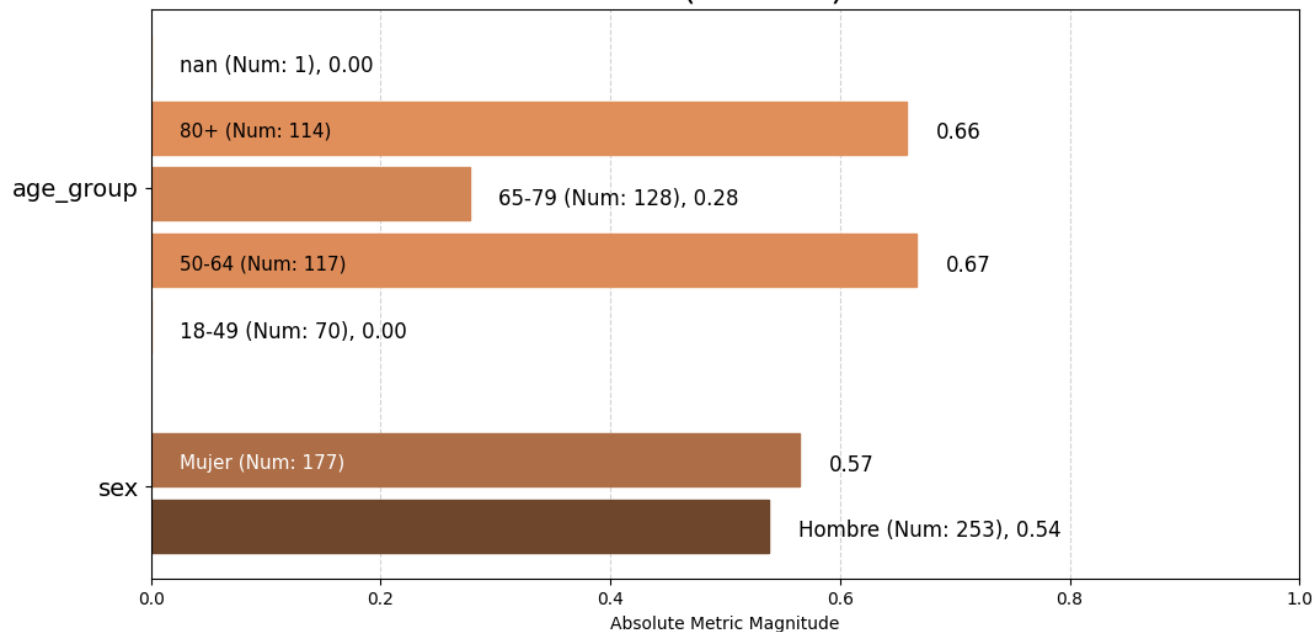


	Precisión	Recall	F1-Score
Vivos (Clase 0)	0.93	0.97	0.95
Fallecidos (Clase 1)	0.77	0.55	0.64

F1 = 0.641

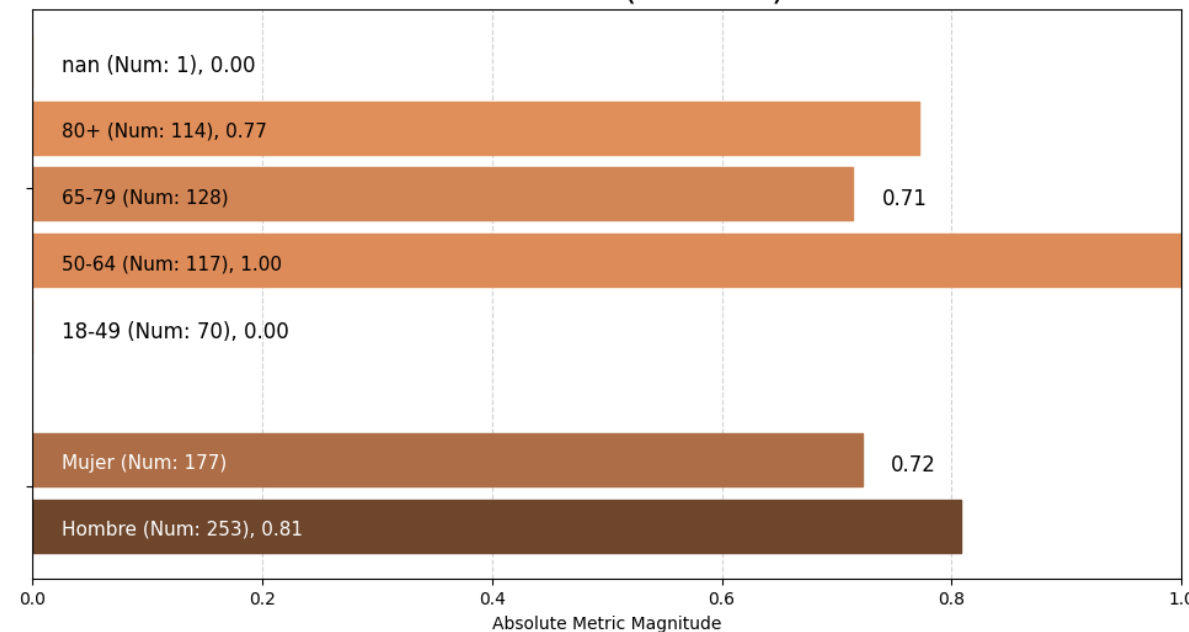
Análisis con Aequitas

TPR (Model 0)



De los que **realmente fallecieron**,
¿qué porcentaje detectó el
modelo?

PRECISION (Model 0)



¿Qué porcentaje de las alertas de
muertes emitidas por el modelo
fue correcto?

Métodos para abordar el desbalanceo

Aprendizaje sensible al coste (*cost-sensitive learning*)



Opción 1

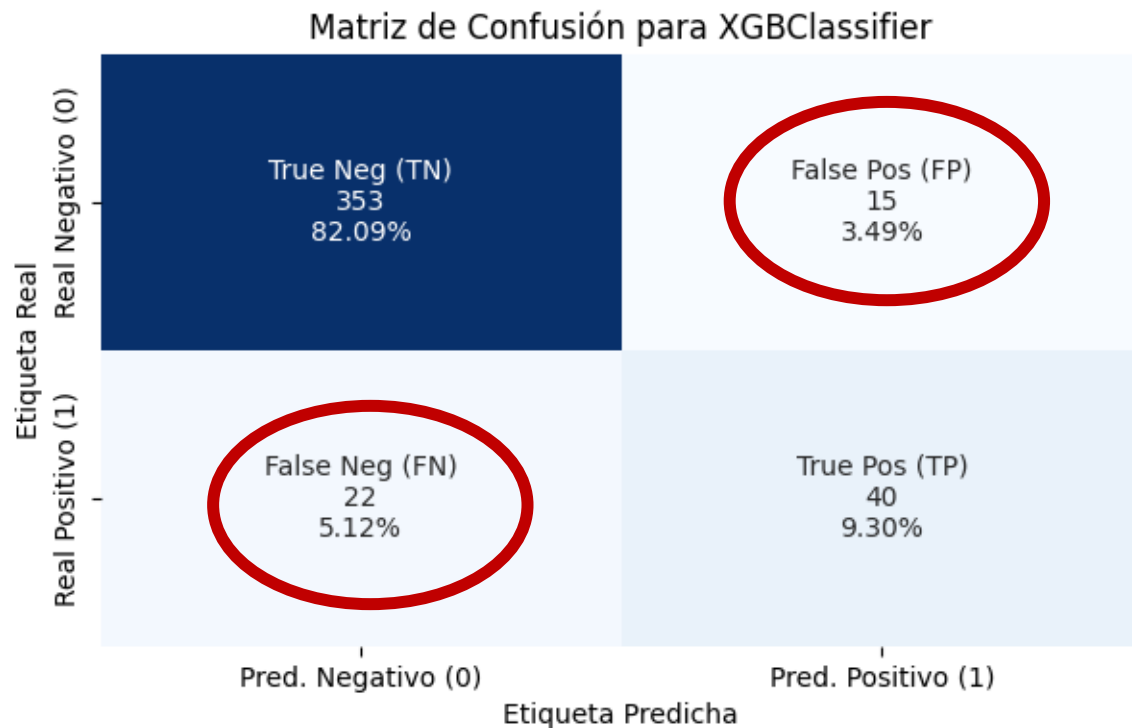
- No cambia la distribución de los datos
- Penaliza más los errores del modelo sobre la clase minoritaria
- Da más importancia a toda la clase minoritaria
- Argumento "class_weight" en clasificadores de sklearn

Opción 2

- Dar más importancia a algunas filas en particular, normalmente, aquellas que el modelo no termina de aprender
- Se implementa con "sample weight", disponible en Sklearn

Métodos para abordar el desbalanceo

Aprendizaje sensible al coste (*cost-sensitive learning*)



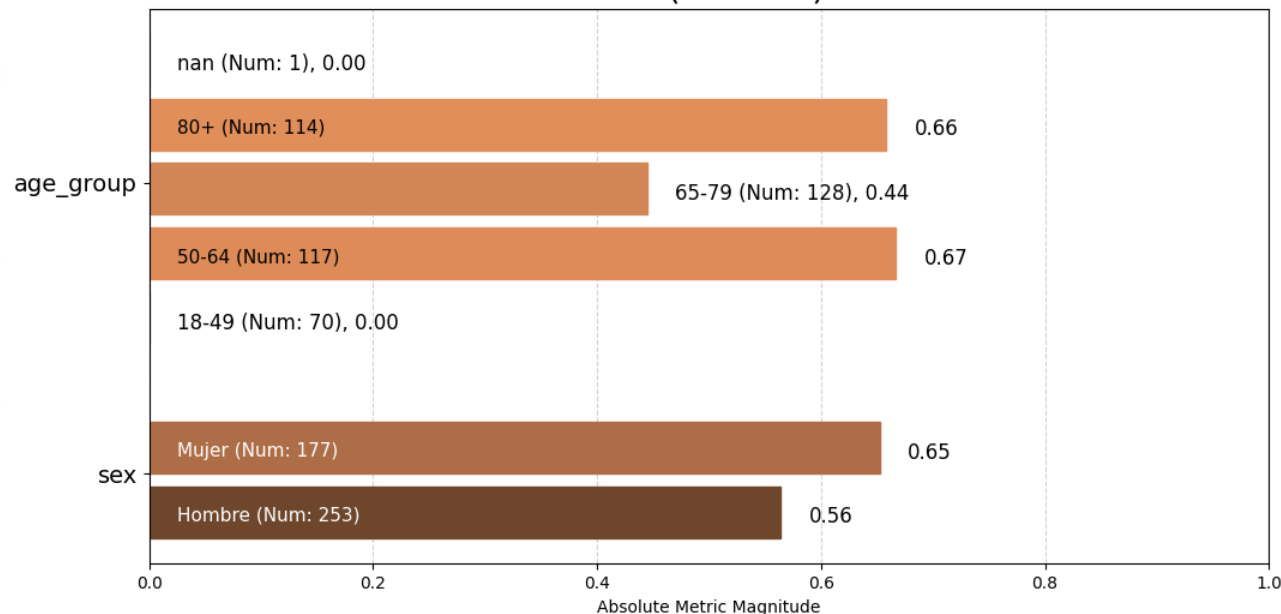
	Precisión	Recall	F1-Score
Vivos (Clase 0)	0.94	0.95	0.95
Fallecidos (Clase 1)	0.72	0.64	0.68

FN ha bajado en 6
FP ha subido en 5

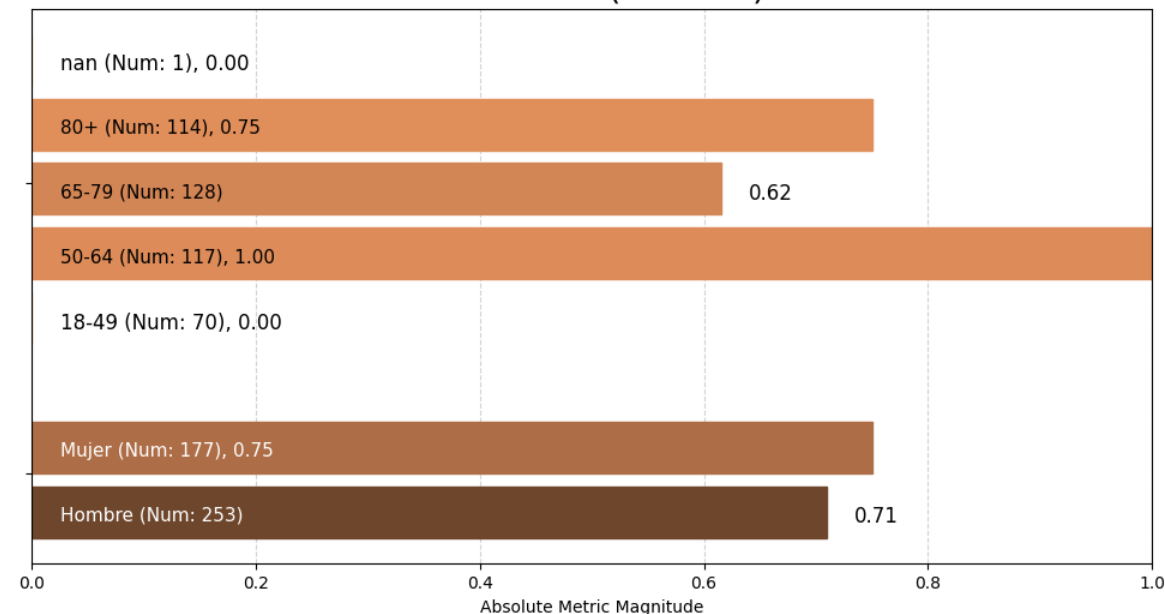
F1 = 0.68

Análisis con Aequitas

TPR (Model 0)



PRECISION (Model 0)



De los que **realmente fallecieron**,
¿qué porcentaje detectó el
modelo?

¿Qué porcentaje de las alertas de
muertes emitidas por el modelo
fue correcto?

Métodos para abordar el desbalanceo

Optimización del umbral

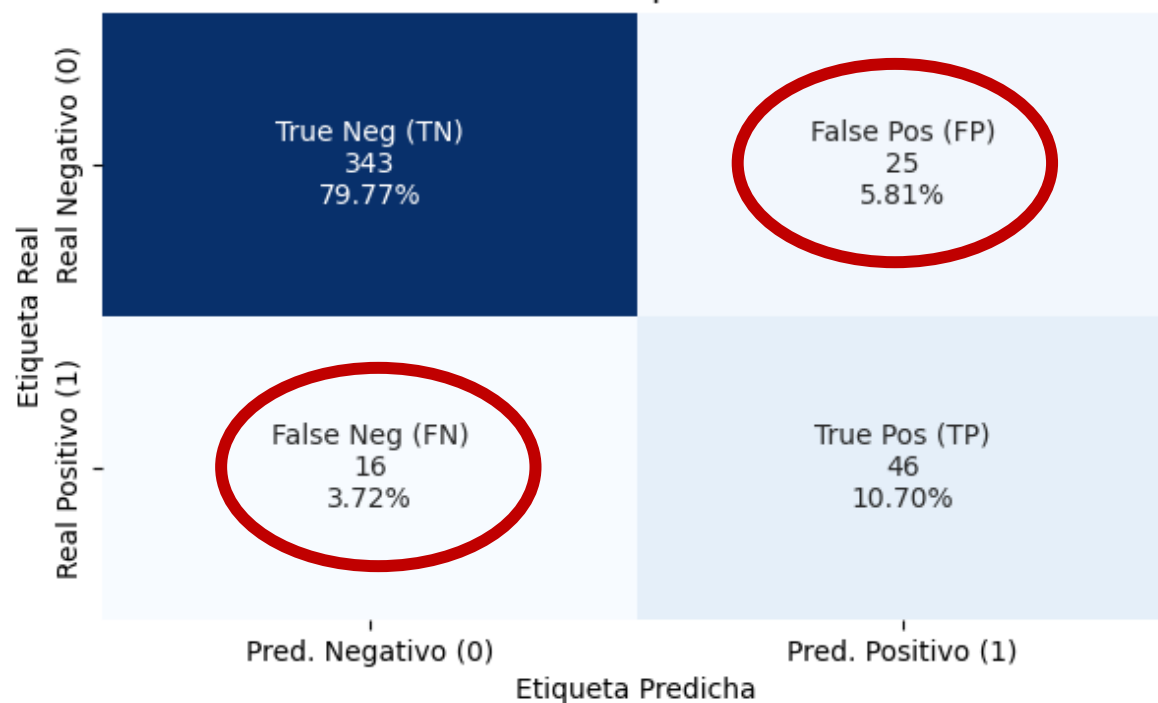
- No considerar el umbral por defecto
- Buscar el mejor umbral posible para optimizar la métrica deseada
- TunedThreshholdClassifier CV funciona como wrapper para clasificadores de Sklearn
- Usa validación cruzada para elegir el mejor umbral sobre varios folds
- No realiza cambios sobre el modelo, si no sobre el umbral



Métodos para abordar el desbalanceo

Optimización del umbral

Matriz de Confusión para XGBClassifier



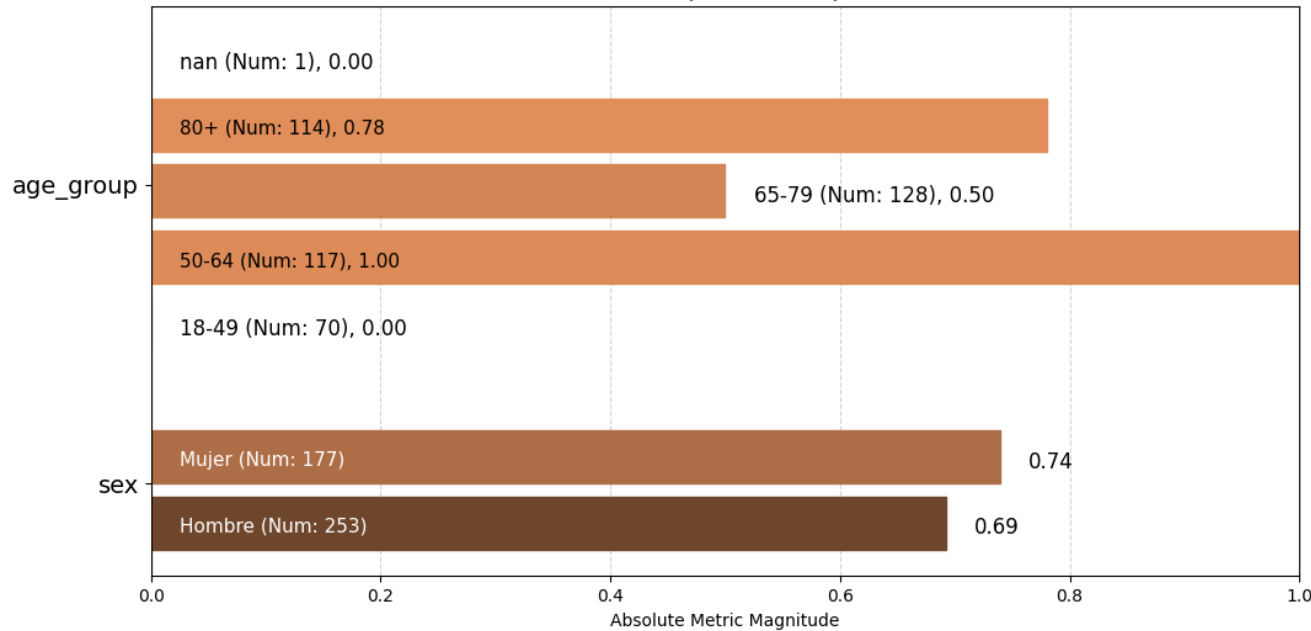
	Precisión	Recall	F1-Score
Vivos (Clase 0)	0.96	0.93	0.94
Fallecidos (Clase 1)	0.65	0.74	0.69

FN ha bajado en 6
FP ha subido en 10

F1 = 0.69

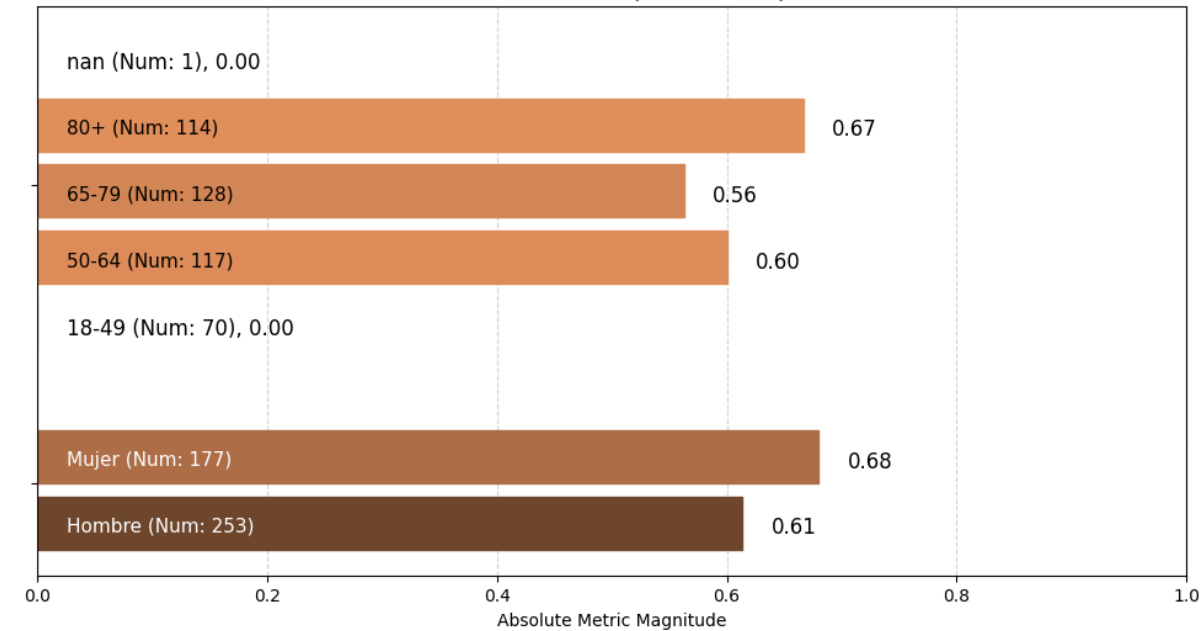
Análisis con Aequitas

TPR (Model 0)



De los que **realmente fallecieron**,
¿qué porcentaje detectó el
modelo?

PRECISION (Model 0)



¿Qué porcentaje de las alertas de
muertes emitidas por el modelo
fue correcto?

Métodos para abordar el desbalanceo

Comparación

Sin aprendizaje sensible al coste
Umbral por defecto

Matriz de Confusión para XGBClassifier

Etiqueta Real	Etiqueta Predicha	
	Pred. Negativo (0)	Pred. Positivo (1)
Real Negativo (0)	True Neg (TN) 358 83.26%	False Pos (FP) 10 2.33%
Real Positivo (1)	False Neg (FN) 28 6.51%	True Pos (TP) 34 7.91%

Aprendizaje sensible al coste
Umbral por defecto

Matriz de Confusión para XGBClassifier

Etiqueta Real	Etiqueta Predicha	
	Pred. Negativo (0)	Pred. Positivo (1)
Real Negativo (0)	True Neg (TN) 353 82.09%	False Pos (FP) 15 3.49%
Real Positivo (1)	False Neg (FN) 22 5.12%	True Pos (TP) 40 9.30%

Aprendizaje sensible al coste
Umbral optimizado

Matriz de Confusión para XGBClassifier

Etiqueta Real	Etiqueta Predicha	
	Pred. Negativo (0)	Pred. Positivo (1)
Real Negativo (0)	True Neg (TN) 343 79.77%	False Pos (FP) 25 5.81%
Real Positivo (1)	False Neg (FN) 16 3.72%	True Pos (TP) 46 10.70%

Bajan los FN a costa de los FP

Bajan los FN a costa de los FP

Aequitas

Comparación

Sin aprendizaje sensible al coste Umbral por defecto

Grupo 18-49 años

- No identifica a ningún paciente que fallece.

Grupo de 50-64 años

- 100% de precisión pero pasa por alto 1 de cada 3 pacientes que fallece

Grupo de 65-79 años

- El modelo solo identifica al 50% de los pacientes que realmente fallecen en este grupo, y cuando lo hace, solo la mitad de las veces acierta

Grupo de 80+

- Modelo equilibrado y medianamente competente. Capacidad de detección del 78% y precisión del 67%

Sexo

- Ligeramente mejor funcionamiento para mujeres que para hombres

Aprendizaje sensible al coste Umbral por defecto

Grupo 18-49 años

- Igual que el anterior

Grupo de 50-64 años

- Igual que el anterior

Grupo de 65-79 años

- El modelo solo identifica al 30% de los pacientes que realmente fallecen en este grupo, y cuando lo hace, acierta el 70% de las veces

Grupo de 80+

- Se mantiene, de manera aproximada, el resultado anterior

Sexo

- Ligeramente mejor funcionamiento para mujeres que para hombres

Aprendizaje sensible al coste Umbral optimizado

Grupo 18-49 años

- Igual que el anterior.

Grupo de 50-64 años

- Predice correctamente a todos los fallecidos, pero tiene un 40% de falsos positivos

Grupo de 65-79 años

- El modelo solo identifica al 50% de los pacientes que realmente fallecen en este grupo, y cuando lo hace, acierta el 50% de las veces

Grupo de 80+

- Se mantiene, de manera aproximada, el resultado anterior

Sexo

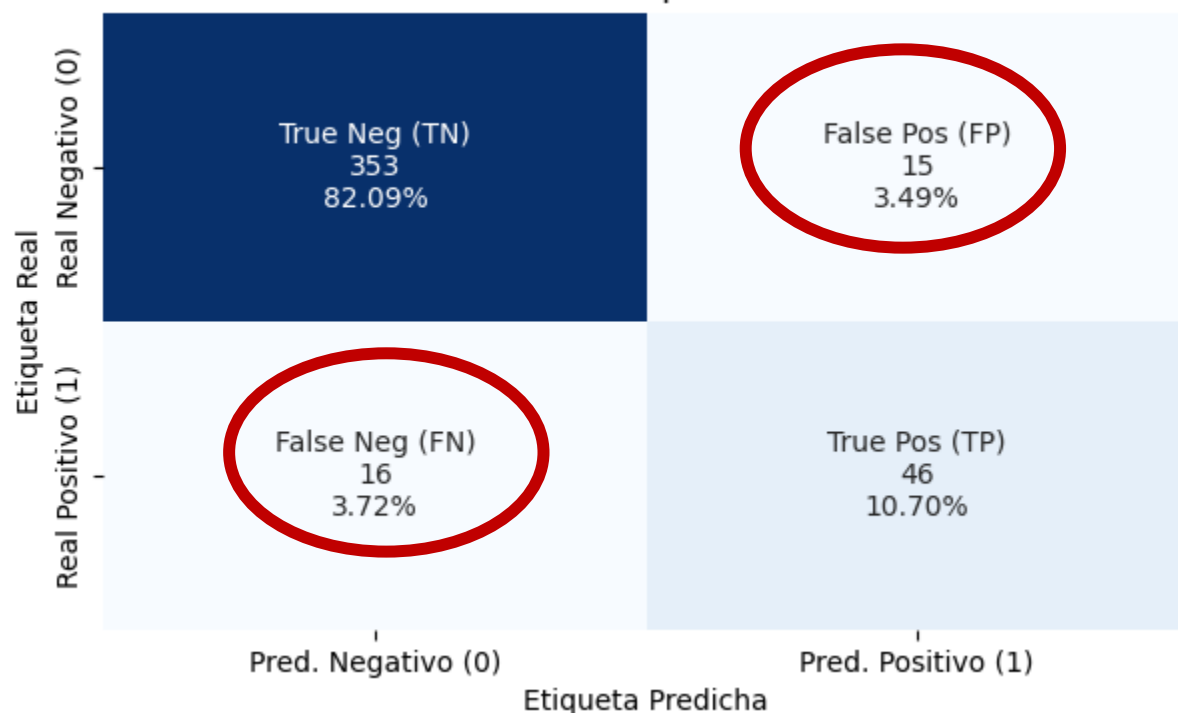
- Ligeramente mejor funcionamiento para mujeres que para hombres

Métodos para abordar el desbalanceo

Otro ejemplo

Aprendizaje sensible al coste, Umbral por defecto
Optimización de hiperparámetros

Matriz de Confusión para XGBClassifier

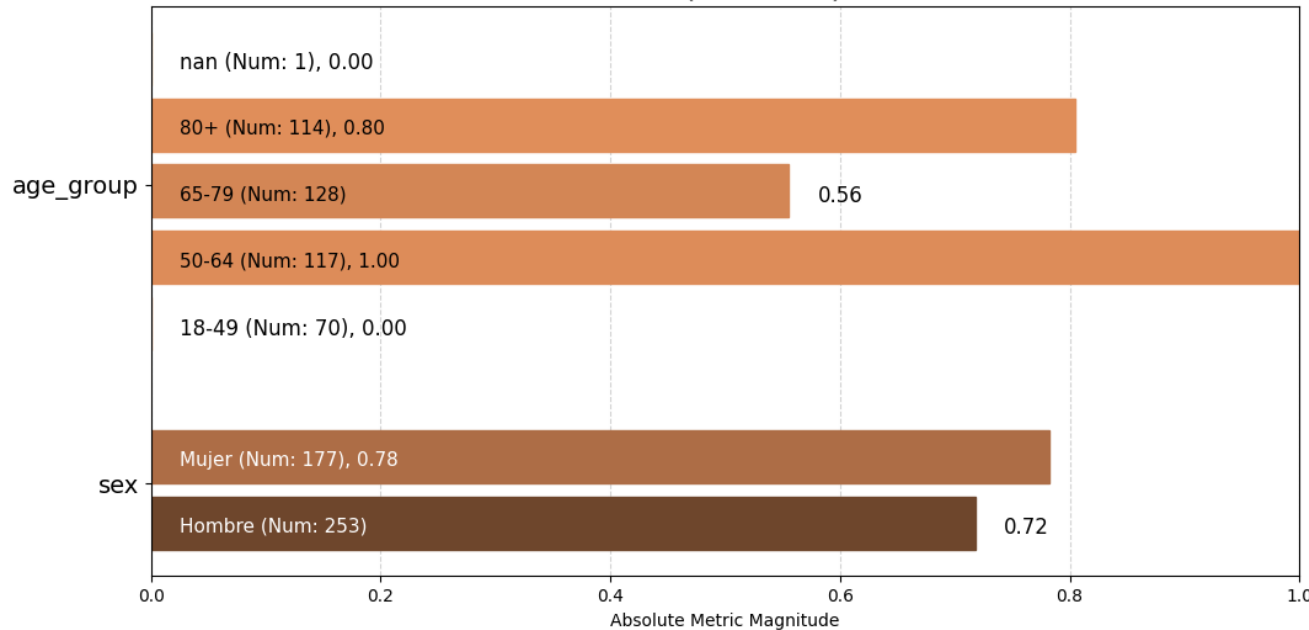


	Precisión	Recall	F1-Score
Vivos (Clase 0)	0.96	0.96	0.94
Fallecidos (Clase 1)	0.75	0.74	0.69

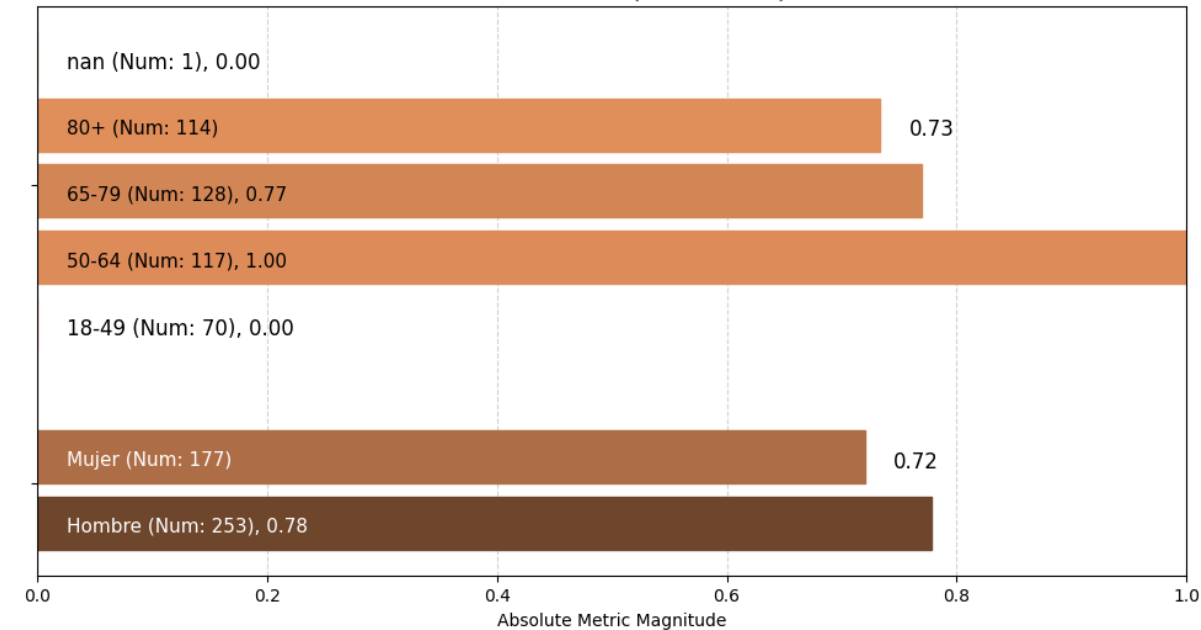
F1 = 0.75

Análisis con Aequitas

TPR (Model 0)

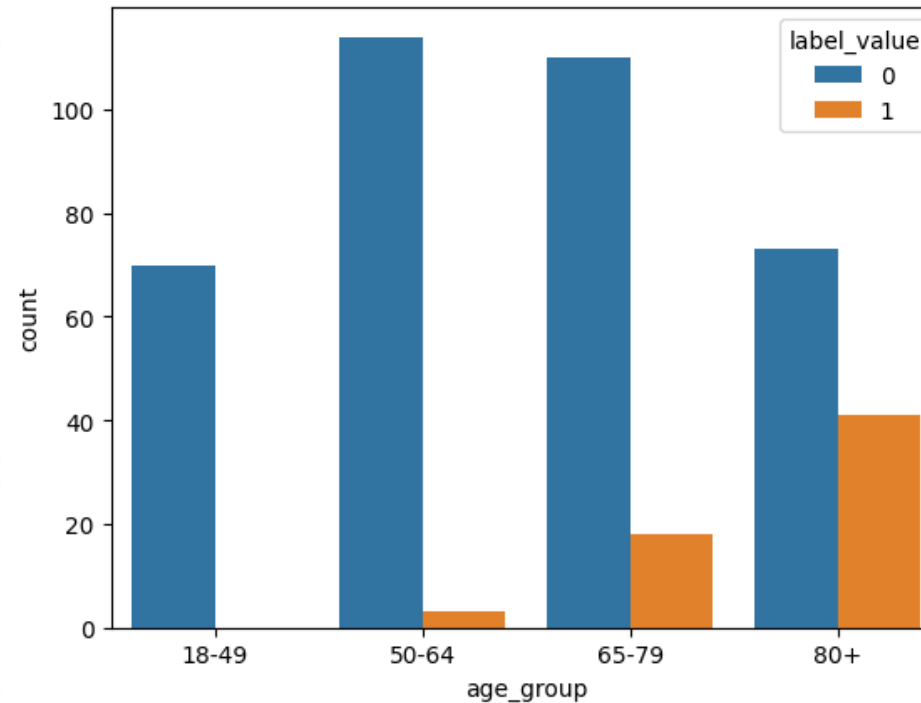


PRECISION (Model 0)



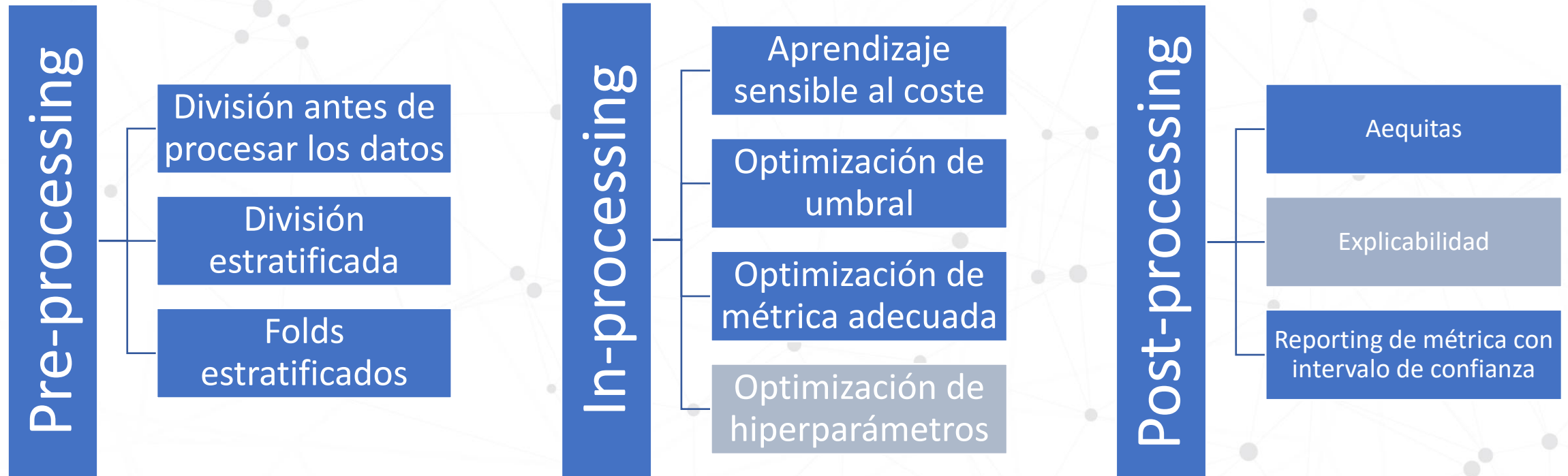
Inútil total para grupo de edad 18-49.
Rendimiento perfecto para grupo de edad 50-64.
Ligera mejora para los grupos de edad 65-79 y 80+.

Análisis con Aequitas



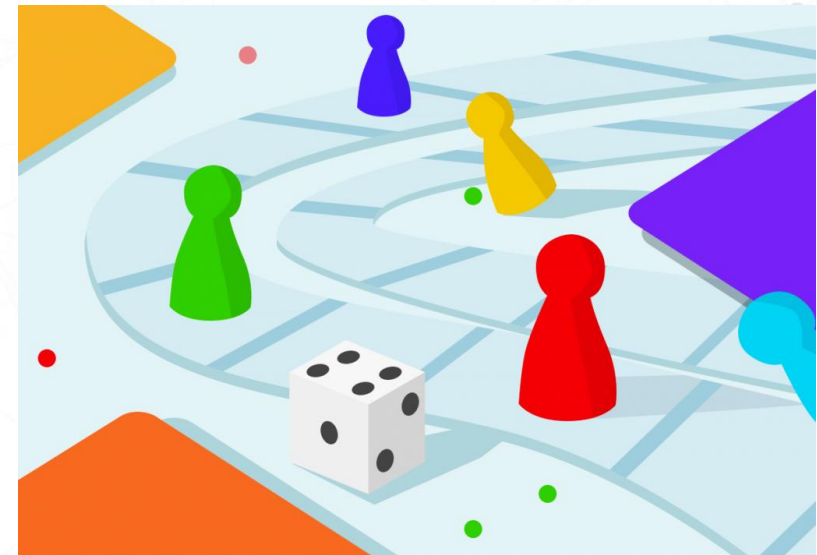
Rendimiento perfecto para grupo de edad 50-64.
Ligera mejora para los grupos de edad 65-79 y 80+, pero funciona mejor
para los grupos de edad más afectados.

Recapitulando



Moving forward

- ¿Qué pasa con el rango de edad 18-49?
- Modelo no recomendable en su estado actual.
- Es muy complejo abordar la problemática en los datos desde el punto de vista técnico.
- Hacen falta equipos multidisciplinarios.



Fuente: <https://desktime.com/blog/es/51-juegos-de-trabajo-en-equipo>

CONFERENCIAS DE POSGRADO – UNIVERSIDAD COMPLUTENSE DE MADRID

Abordando el desbalance de datos y la equidad en modelos de clasificación de Machine Learning para COVID-19

Dra. Andreea M. Oprea

aoprescu@us.es

11 Junio 2025