

# Informe técnico: Medicine and Environment Impact on Migraine

## Resumen ejecutivo

Este estudio evalúa la capacidad de predecir la aparición de un ataque de migraña utilizando el dataset Medicine and Environment Impact on Migraine (Kaggle), que incluye **4.152 registros de 133 pacientes** con variables demográficas, clínicas, ambientales y de tratamiento. La pregunta central - **¿se puede predecir un ataque de migraña?** - se aborda mediante modelado predictivo (Python) e inferencia estadística multivariada (R).

El modelo ganador, una **regresión logística calibrada isotónicamente**, logra una discriminación limitada, reflejando la naturaleza compleja y heterogénea de la migraña. Los análisis confirman que la medicación continua reduce significativamente el riesgo de ataque ( $p < 0.05$  en ANOVA y OR ajustado), siendo el factor más protector identificado. Por otro lado, ni el subtipo de migraña (con/sin aura), ni el sexo ni la edad muestran efectos independientes robustos, aunque se observan tendencias no significativas (ej. mayor riesgo en hombres y en otoño).

La calidad del aire (AQI) y la estacionalidad —especialmente otoño/invierno— aparecen como moduladores contextuales del riesgo, con efectos no lineales y variables según la fase de tratamiento, según los análisis ALE e ICE. Estos últimos revelan una alta heterogeneidad individual: no existe un patrón universal de respuesta, sino interacciones únicas entre historial clínico, tratamiento y entorno.

El modelo presenta sensibilidad elevada ( $>0.80$ ), pero su especificidad no puede evaluarse de forma robusta en todos los subgrupos debido al marcado desbalance del target (64 % de ataques) y la escasez de casos negativos en categorías minoritarias (hombres, subtipo mixed, etc.), lo que también limita el análisis de equidad.

En conclusión, aunque la predicción exacta de un ataque de migraña sigue siendo un reto, los resultados respaldan un enfoque personalizado en la prevención, centrado en la adherencia al tratamiento y la monitorización del contexto ambiental. Se recomienda integrar datos longitudinales en tiempo real y ampliar la cohorte con mayor diversidad para mejorar la validez clínica de futuros modelos predictivos.

## Introducción y contexto

La **migraña** es un **trastorno neurológico crónico de origen genético**, caracterizado por episodios recurrentes de dolor de cabeza moderado a grave, acompañados de síntomas como náuseas, vómitos, fotofobia y fonofobia. En un **20–30 % de los casos se presenta aura**. Su origen combina factores genéticos y ambientales, con desencadenantes como estrés, cambios hormonales, alteraciones del sueño y variaciones climáticas.

Representa una carga significativa de salud pública, siendo una de las principales causas de discapacidad neurológica a nivel mundial y la **tercera enfermedad neurológica más frecuente**.

Este proyecto utiliza el dataset público *Medicine and Environment Impact on Migraine* (*Kaggle*) para desarrollar y validar un modelo predictivo que estime la probabilidad de aparición de un ataque de migraña, considerando variables demográficas, clínicas, ambientales y de tratamiento. De deberá contestas a las siguientes preguntas:

**Pregunta principal:** ¿Se puede predecir la aparición de un ataque de migraña?

**Preguntas secundarias:**

- Tratamiento: ¿La medicación reduce la probabilidad de tener migraña?
- Subtipo clínico: ¿El tipo de migraña influye en la aparición de ataques?
- Ambiente: ¿La calidad del aire se asocia con mayor riesgo de ataque de migraña?
- Factores demográficos: ¿La edad y el sexo modifican la probabilidad de ataque?

**Nota:** se han utilizado dos lenguajes de programación en este estudio (Python y R).

## Datos

Datos del datset:

Variable	Tipo	Descripción	Rango	Unidades	Observaciones
<b>rownames</b>	Entera	Índice del registro	1 – 4152	-	Único para cada registro
<b>id</b>	Entera	Identificador único del paciente	1 – 133	-	Cada paciente tiene múltiples registros
<b>time</b>	Entera	Índice temporal relativo al tratamiento	0 – 23	días	Días antes del inicio, inicio y días posteriores de tratamiento
<b>dos</b>	Entera	Días de estudio	98 – 1239	días	Número de días transcurridos desde el comienzo del estudio
<b>hatype</b>	Categórica	Tipo de migraña	No Aura, Aura, Mixed	-	Variable nominal con 3 categorías
<b>age</b>	Entera	Edad del paciente	18 – 66	años	Distribución centrada en adultos
<b>airq</b>	Numérica	Calidad del aire	3 – 73	AQI	Índice de calidad del aire

<b>medication</b>	Categórica	Tipo de medicación	continuing, reduced, none	-	Régimen de tratamiento
<b>headache</b>	Categórica	Presencia de dolor de cabeza en el registro	yes, no	-	Variable binaria (target)
<b>sex</b>	Categórica	Sexo del paciente	female, male	-	Variable binaria

A tener en cuenta:

- **Múltiples entrada por id:** hay 133 pacientes distribuidos en 4152 entradas.
- **Target (headache):** desbalanceado (64,21% yes / 35,79% no).
- **Sex:** desbalanceado (85,38% female / 14,62% male), cumple con la realidad clínica.
- **Hatype (tipo de migraña):** también desbalanceado (47,81% no aura / 41,18% aura / 11,01% mixed), cumple con la realidad clínica.
- **Medication (régimen de medicación):** también desbalanceado (57,47% continuing / 23,63% reduced / 18,91% none), cumple con la realidad clínica.
- **Time y dos:** aunque son variables temporales no se pueden utilizar como tal debido a que no separan ataques de migraña de no ataques; time trata de cuando empieza el tratamiento y dos del día del estudio que se produjo el ataque o no.
- No hay valores nulos y solo 1 duplicado.
- **Outliers** en *time* y *airq* que se mantienen al no ser considerados errores de captura, ya que aparecen en múltiples pacientes y son consistentes.

## Nuevas variables

Variable	Tipo	Descripción	Rango	Unidades	Observaciones
<b>fase</b>	Categórica	Clasifica registros según si hay o no tratamiento	Sin tratamiento / Con tratamiento	-	Derivada de 'time'
<b>month_study</b>	Numérica	Mes calendario	1 – 12	Mes	Calculado a partir de 'dos'
<b>season_study</b>	Categórica	Estación del año según mes de visita	winter, spring, summer, autumn	None	Mapeo de meses a estaciones
<b>days_since_first_visit</b>	Numérica	Días desde la primera visita del paciente	0 - 128	Días	Diferencia entre cada visita y la primera visita
<b>age_band</b>	Categórica	Bandas de edad	18–30, 31–40, 41–50, 51–66	Años	Agrupación de la variable 'age' en intervalos
<b>n_visitas</b>	Numérica	Número acumulado de visitas por paciente	1 – 121	Visitas	Contador longitudinal por paciente
<b>target_num</b>	Numérica	Variable objetivo binaria	0, 1	-	Ataque=1, no ataque=0
<b>airq_prev_mean</b>	Numérica	Promedio acumulado de calidad del aire	8 – 56	Índice AQI	Calculado por paciente, excluye la visita actual

<b>prev_attacks</b>	Numérica	Número acumulado de ataques previos	0 - 77	Ataques	Cálculo longitudinal, excluye el ataque actual
---------------------	----------	-------------------------------------	--------	---------	--

## Metodología (programación: Python)

- **Split:** 60/20/20 con SimpleImputer y StandardScaler para numéricas y OneHotEncoder para categóricas.
- **Variables numéricas:** prev\_attacks, n\_visitas ,days\_since\_first\_visit, airq y airq\_prev\_mean.
- **Variables categóricas:** hatype, medication, sex, season\_study, fase y age\_band
- **Drops:** rownames, id, sample\_weight, target\_num, time, dos, target, headache, age y month\_study.
- **Target:** target\_num.
- **Modelos entrenados:** Logistic Regression, Gradient Boosting y XGBoost.
- **Calibración:** Isotónica con GroupKFold(n\_splits=5).
- **Tunning:** GridSearchCV por scoring='roc\_auc'.
- **Intentos de mejora calibración:** se realiza diagnóstico para ver variables más predictoras, se crean flags y se re-entrenan los modelos → no hay mejora importante → Se decide mantener calibración sin flags.
- **Modelo ganador:** Logistic Regression Tuned (Isotonic).

	ROC-AUC	PR-AUC	Brier	ECE	TP	TN	FP	FN	Precision	Sensitivity
<b>Logistic Regression Test</b>	0.706945	0.748382	0.215519	0.078108	434.0	158.0	197.0	90.0	0.687797	0.828244
<b>Gradient Boosting Test</b>	0.661281	0.718108	0.235093	0.100923	404.0	155.0	200.0	120.0	0.668874	0.770992
<b>XGBoost Test</b>	0.649806	0.701571	0.243897	0.132894	421.0	148.0	207.0	103.0	0.670382	0.803435
<b>Logistic Regression (Isotonic)</b>	0.706997	0.743733	0.211640	0.055063	422.0	172.0	183.0	102.0	0.697521	0.805344
<b>Gradient Boosting (Isotonic)</b>	0.658158	0.706390	0.227895	0.052225	435.0	127.0	228.0	89.0	0.656109	0.830153
<b>XGBoost (Isotonic)</b>	0.647906	0.692573	0.229698	0.054663	401.0	156.0	199.0	123.0	0.668333	0.765267
<b>Logistic Regression Tuned (Isotonic)</b>	0.714372	0.751580	0.209624	0.044401	431.0	167.0	188.0	93.0	0.696284	0.822519
<b>Gradient Boosting Tuned (Isotonic)</b>	0.658158	0.706390	0.227895	0.052225	435.0	127.0	228.0	89.0	0.656109	0.830153
<b>XGBoost Tuned (Isotonic)</b>	0.683881	0.723920	0.219329	0.048293	378.0	199.0	156.0	146.0	0.707865	0.721374

- **Métricas del modelo ganador:** (gráficas en anexo A1)

Métrica	ROC-AUC	PR-AUC	Brier	ECE	TP	TN	FP	FN	Precision	Sensitivity
<b>Media</b>	0.714	0.751	0.210	0.053	430	167	189	93	0.695	0.823
<b>Std</b>	0.018	0.021	0.008	0.013	15	12	12	9	0.019	0.017
<b>IC95_min</b>	0.677	0.710	0.195	0.029	401	144	166	76	0.659	0.789
<b>IC95_max</b>	0.749	0.790	0.225	0.079	460	191	212	111	0.730	0.853

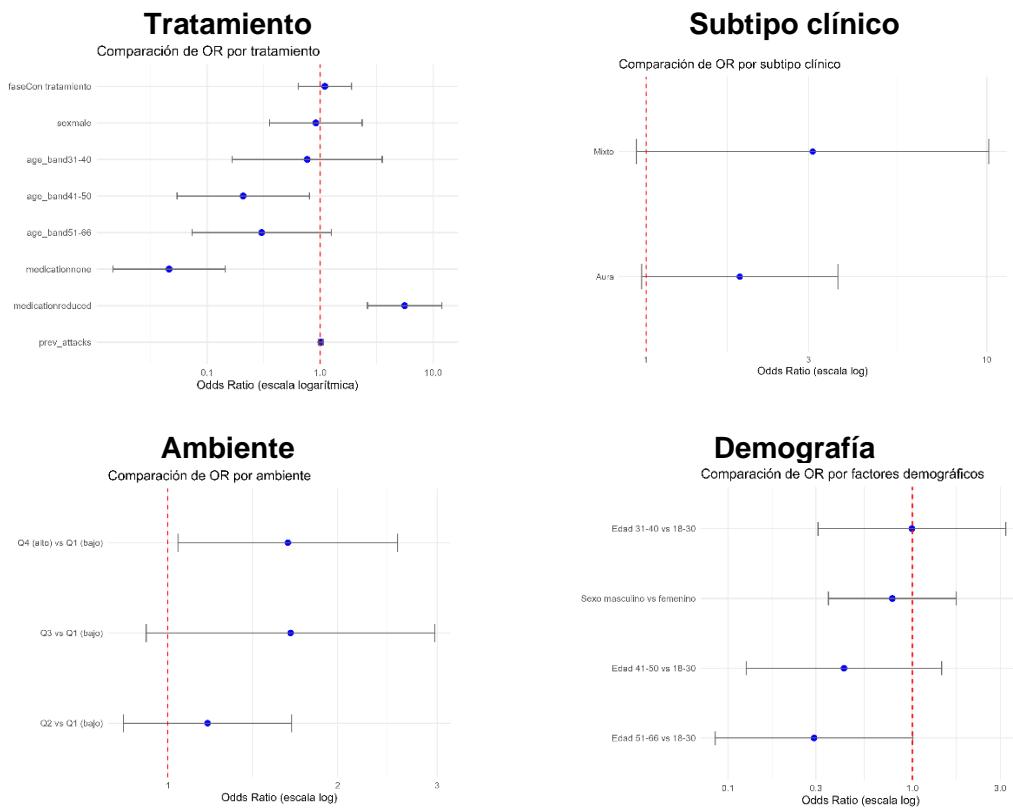
- **Umbral seleccionado:** debido al gran desbalance que tiene este dataset en el target, se busca la optimización del ratio C\_FP/C\_FN como criterio con una sensibilidad mínima del 0.80.

C_FP	C_FN	threshold	Sens	Spec	F1	Youden	TP	FP	FN	TN	Coste	Media_alertas
1.0	1.0	0.534	0.695	0.662	0.722	0.357	364.0	120.0	160.0	235.0	280.0	0.551

## Resultados (programación: R, excepto SHAP)

Se ha hecho un análisis multivariante (OR ajustados con SE robusto) para diferentes variables que pueden ayudar a predecir un ataque de migraña. Los resultados obtenidos son:

- **Tratamiento:** la medicación emerge como un factor en el riesgo de migraña según los OR ajustados, y el análisis secundario (ANOVA) confirma diferencias significativas en la proporción de ataques entre tipos de medicación (ver análisis ANOVA en anexos A2).
- **Subtipo clínico:** no se demuestra un efecto significativo del subtipo clínico en la aparición de migrañas, aunque el subtipo mixto muestra una tendencia no confirmada como independiente.
- **Ambiente:** a medida que empeora la calidad del aire (cuartiles altos) se perfila una tendencia a mayor riesgo de migraña, pero sin significación estadística robusta.
- **Demografía:** se observa un mayor riesgo de migraña en hombres y en el grupo de 31–40 años, pero sin significación estadística robusta.



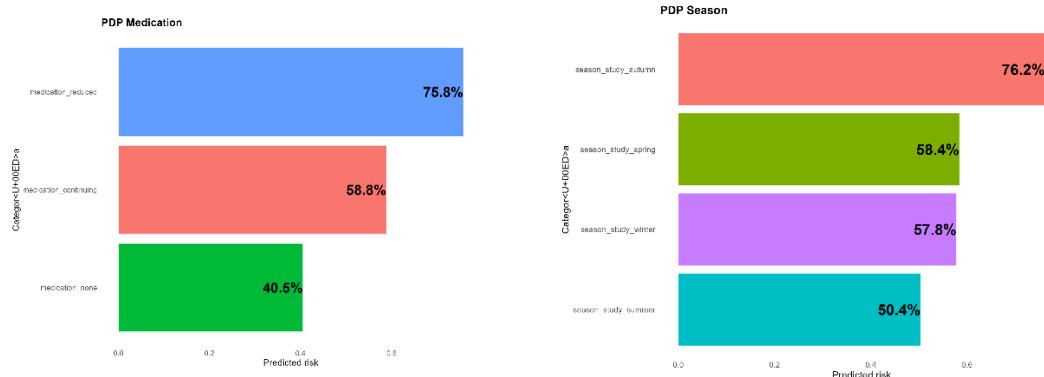
## Equidad (fairness)

El modelo muestra sensibilidad perfecta en todos los subgrupos, pero la falta de valores negativos en ciertos subgrupos impide evaluar la especificidad y el AUC de forma fiable y por tanto, para un análisis de fairness robusto, se necesitarían más pacientes negativos en cada categoría.

**La Importancia de permutaciones** muestra que ninguna variable aislada tiene un impacto fuerte en el PR-AUC (están en el rango de 1e-4 a 1e-2), solo la calidad del aire (*airq*) aporta algo de señal, pero muy débil (ver gráfica en anexo A3)

### PDP para medication, fase, hatype y season

Este análisis muestra que la continuidad del tratamiento protege, el aura aumenta el riesgo y el otoño es la estación más vulnerable (gráficas más representativas).

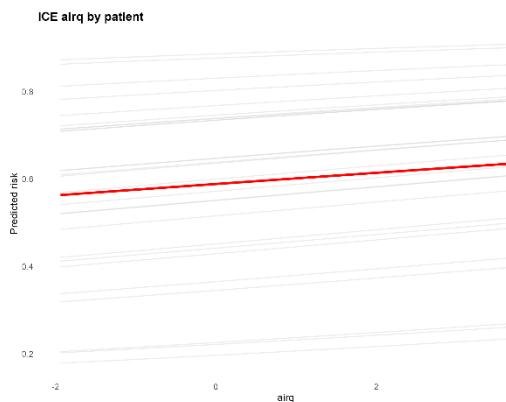
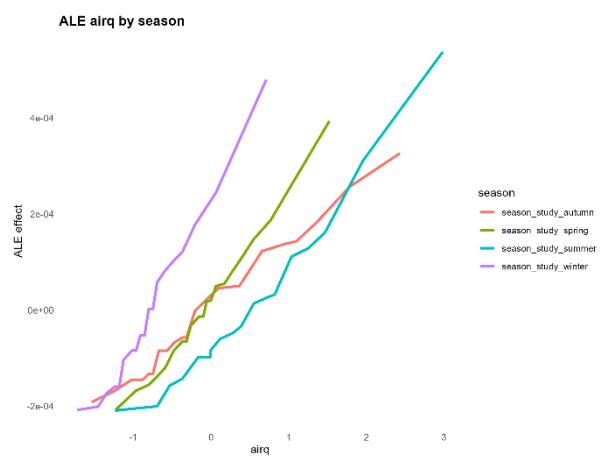


Por los resultados de las permutaciones, se hace **ALE** para ver si la calidad del aire está correlacionada con la estacionalidad y el régimen de tratamiento.

Una peor calidad del aire (mayor *airq*) se le asocia un mayor riesgo de migraña (ver gráfica en anexo A4).

Interacciones no lineales con efecto no uniforme pero sí relevante en la estacionalidad (mayor riesgo en otoño/invierno).

Interacción clínica relevante en pacientes en tratamiento activo migraña (ver gráfica en anexo A4).

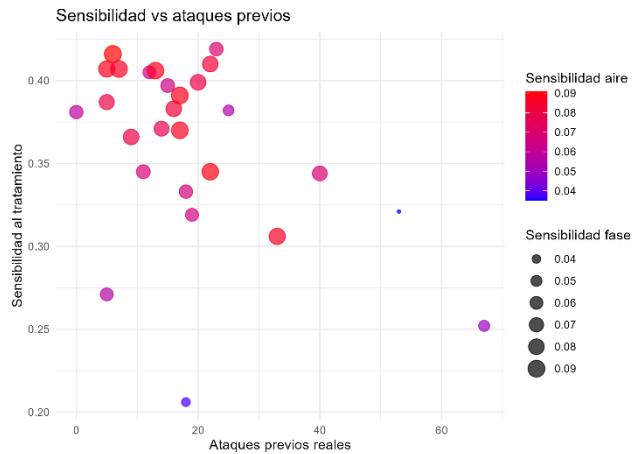


También se hizo un **ICE** que reveló la heterogeneidad individual que los promedios ocultan. El modelo capta efectos acumulativos y contextuales en aire, medicación y fase, mostrando que no hay un patrón universal, sino respuestas diferenciadas por paciente (ver ICE por medicación y fase de tratamiento en anexo A5).

## Índice de sensibilidad

El modelo capta que la sensibilidad al tratamiento no depende linealmente del número de ataques previos, sino que está modulada por la sensibilidad ambiental y física.

Parece que la respuesta al tratamiento es individual y multifactorial, y que pacientes con alta sensibilidad al aire o al tratamiento pueden requerir ajustes personalizados en el enfoque terapéutico, independientemente de su historial de ataques.



## SHAP

Se hace un individual y uno global, y el modelo indica que el control clínico —número de visitas, historial de ataques y medicación activa— es el principal determinante del riesgo. Edad, tipo de migraña y factores ambientales/estacionales influyen, pero solo como moduladores secundarios (ver gráfica completa en anexo A6).

## Discusión, implicaciones prácticas y recomendaciones

Los resultados del modelo predictivo sugieren que, si bien es posible estimar la probabilidad de un ataque de migraña con cierto grado de precisión, la señal predictiva es débil y altamente heterogénea entre pacientes. La medicación continua se asoció claramente con una reducción del riesgo de ataque, lo que refuerza la importancia del cumplimiento terapéutico en la prevención. Además, el análisis mediante PDP y ALE mostró que la estacionalidad —especialmente el otoño— y la calidad del aire actúan como moduladores del riesgo, aunque su efecto no alcanza significación estadística robusta en el análisis multivariado.

Desde una perspectiva clínica, estos hallazgos respaldan un enfoque personalizado en el manejo de la migraña. La heterogeneidad revelada por los perfiles ICE indica que la respuesta al tratamiento y a factores ambientales varía significativamente entre individuos, incluso con características demográficas o clínicas similares. Por tanto, se recomienda integrar variables contextuales (como la fase del tratamiento y la exposición ambiental reciente) en la toma de decisiones clínicas, y considerar ajustes dinámicos del régimen farmacológico según la evolución del paciente.

**Recomendación práctica:** reforzar el seguimiento longitudinal, asegurar la continuidad de la medicación, monitorizar las migrañas y considerar factores ambientales/estacionales en la planificación clínica.

Además, los resultados subrayan la necesidad de herramientas digitales que incorporen datos longitudinales (historial de ataques, número de visitas, tendencias ambientales) para apoyar la monitorización proactiva y la intervención temprana.

## **Limitaciones del estudio y líneas futuras**

Este estudio presenta varias limitaciones. En primer lugar, el desbalance pronunciado en la variable objetivo (64 % de ataques frente a 36 % de no ataques) y en variables como sexo (85 % mujeres) limita la generalización de los hallazgos, especialmente en subgrupos minoritarios. En segundo lugar, la ausencia de registros negativos en ciertas categorías impidió una evaluación robusta de la equidad (fairness) del modelo, particularmente en especificidad y AUC por subgrupos.

Además, aunque se generaron variables derivadas temporales (como `days_since_first_visit` o `prev_attacks`), no fue posible modelar la temporalidad de los ataques con resolución suficiente debido a la estructura del dataset original. Tampoco se dispuso de información sobre la intensidad del dolor, otros desencadenantes individuales (alimentos, estrés percibido) ni sobre comorbilidades relevantes.

Líneas futuras, se sugiere:

- Ampliar la cohorte con más pacientes varones y casos sin ataque, para mejorar la representatividad y la capacidad de análisis de equidad.
- Incorporar sensores ambientales en tiempo real (calidad del aire, temperatura, humedad) y registros subjetivos de estilo de vida mediante aplicaciones móviles.
- Explorar arquitecturas de modelado secuencial (como RNN o Transformers) que capturen mejor la dinámica temporal de los ataques.
- Validar prospectivamente el modelo en entornos clínicos reales como apoyo a la decisión terapéutica.

## **Conclusiones**

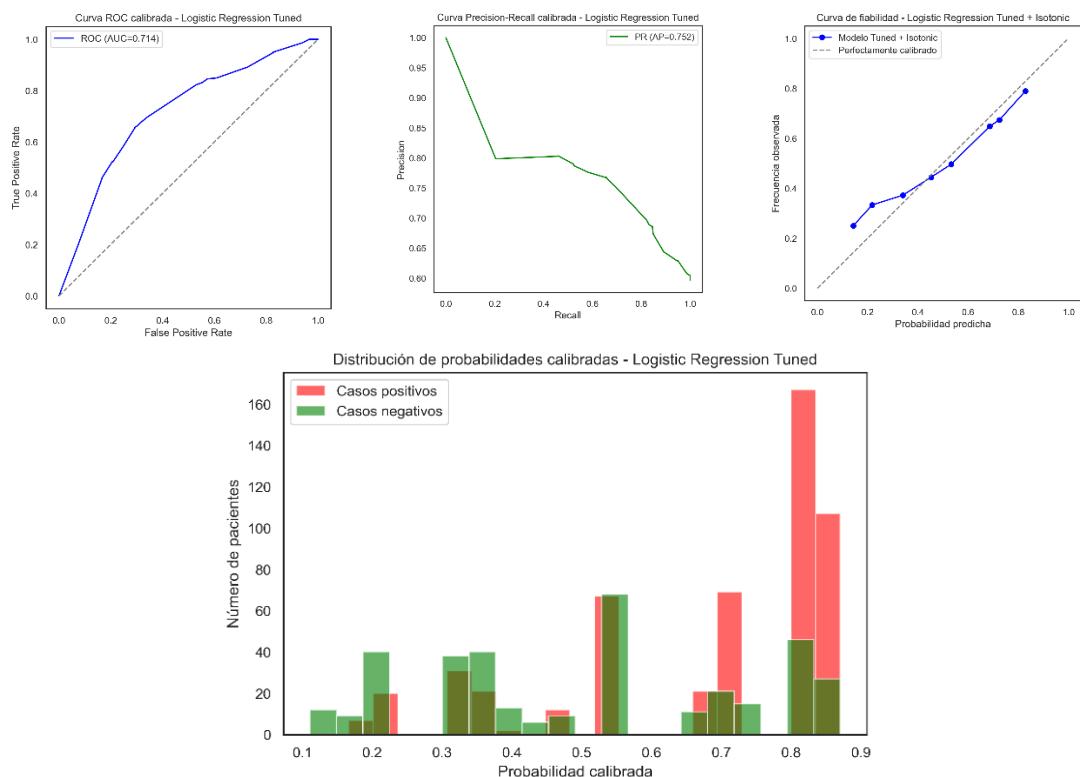
El estudio demuestra que, si bien es técnicamente posible predecir la aparición de un ataque de migraña, la precisión clínica del modelo es limitada por la complejidad multifactorial y la alta variabilidad interindividual del fenómeno. El tratamiento continuo emerge como el factor más protector, mientras que la estacionalidad y la calidad del aire actúan como moduladores contextuales del riesgo.

Los análisis de interpretabilidad (SHAP, ALE, ICE) confirman que el riesgo no se explica por factores aislados, sino por interacciones dinámicas entre historial clínico, fase del tratamiento y entorno. Esto refuerza la necesidad de un enfoque personalizado y adaptativo en la prevención de la migraña, más allá de las categorías diagnósticas tradicionales.

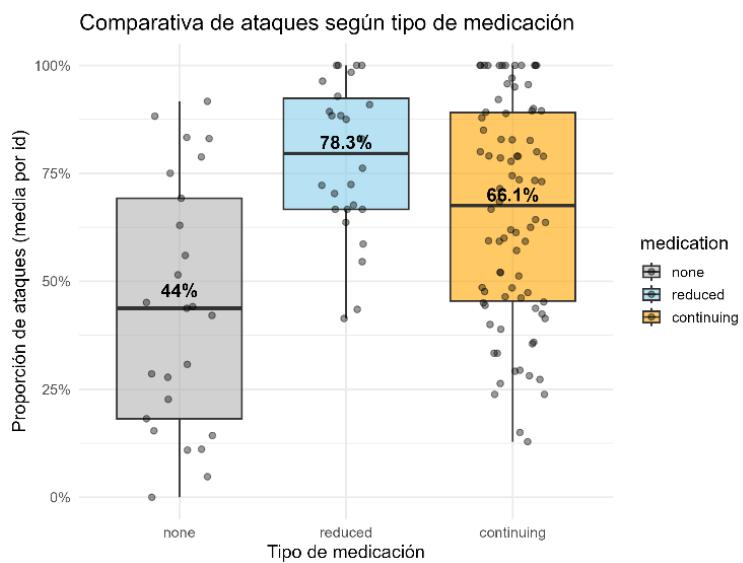
En conjunto, este trabajo sienta las bases para sistemas de apoyo a la decisión clínica basados en datos longitudinales y factores ambientales, aunque se requieren mejoras sustanciales en la calidad y diversidad de los datos para lograr impacto real en la práctica médica.

## Anexos

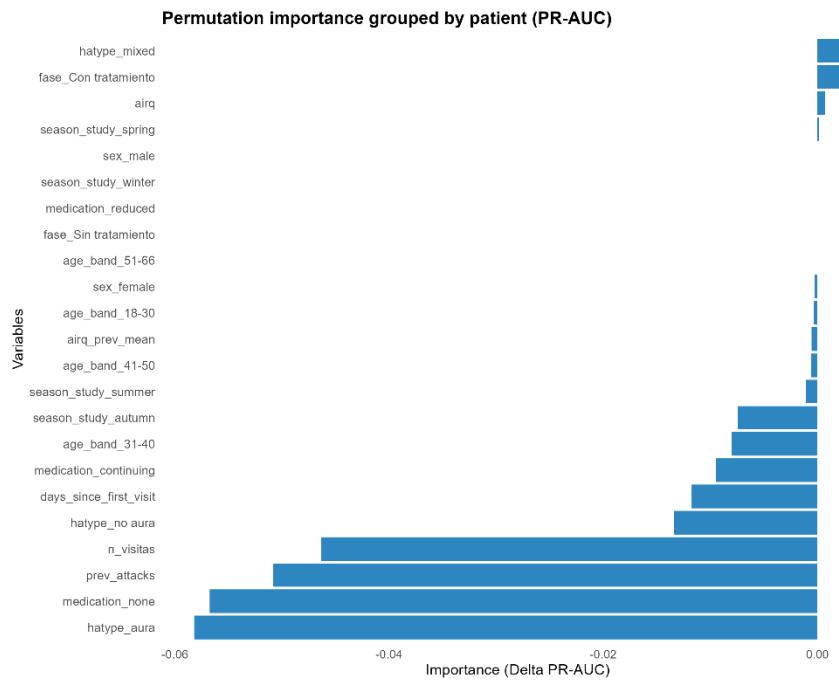
### A1. Curvas calibradas modelo ganador



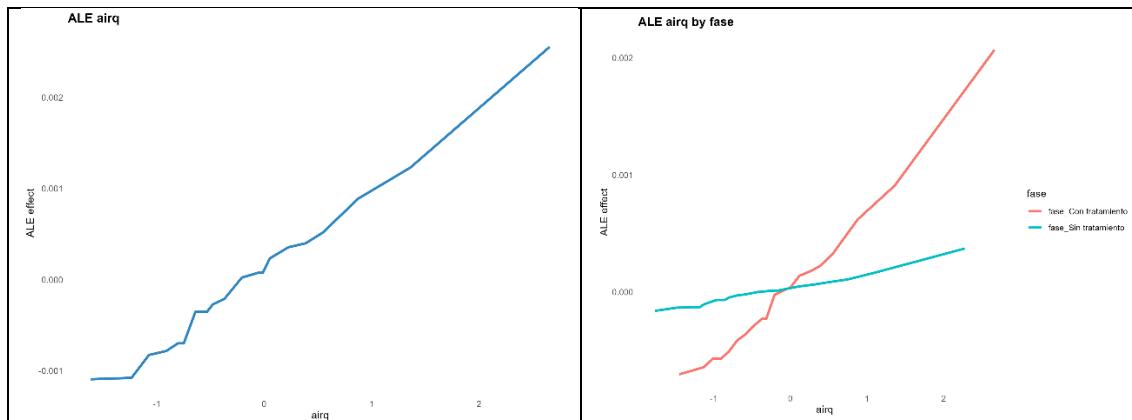
### A2. Análisis ANOVA ponderado para tratamiento



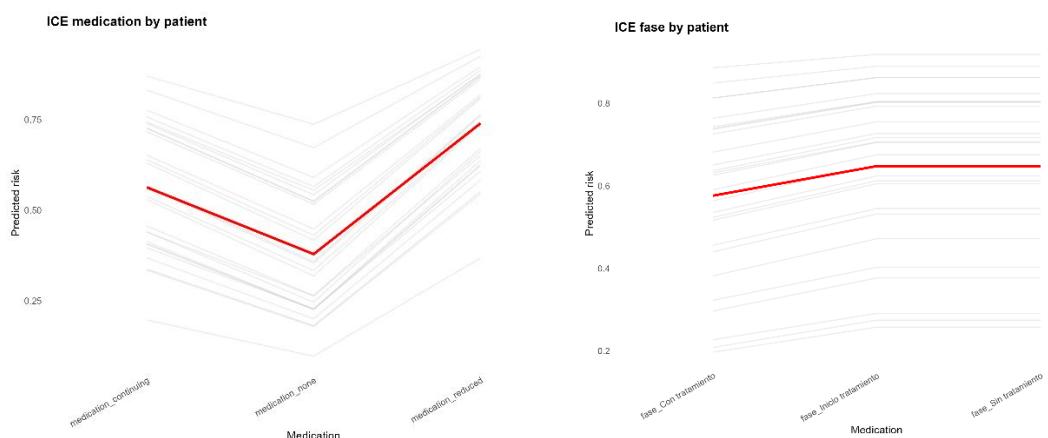
### A3. Importancia de permutaciones



### A4. ALE para airq y airq por fase



### A5. ICE por medication y fase



## A6. SHAP global

