THE BRIDGE DIGITAL TALENT ACCELERATOR

MODELO PREDICTIVO PARA LA TOXICIDAD DE SUSTANCIAS QUÍMICAS EN PECES

HACIA UNA EVALUACIÓN MÁS ÉTICA Y EFICIENTE

Objetivo

Responder a la necesidad de métodos alternativos más rápidos, éticos y explicables.

Reducción de ensayos *in vivo*

Predicciones computacionales más éticas.

Ahorro de recursos

Disminuir costes económicos y tiempos asociados a ensayos tradicionales.

Aceleración de toma de decisiones

Facilitar decisiones regulatorias más rápidas con modelos reproducibles.

Enfoque supervisado

Entrenamiento del modelo a partir de datos etiquetados y verificables.

Clasificación binaria

Predicción de toxicidad como clase dicotómica: alta o baja.

Dataset y Diseño experimental

ADORE Benchmark (A Data-driven benchmark fOR Ecotoxicology)

Schür *et al.* (2023)

t-F2F_mortality

Fish 2 Fish ⇧ 1200

Basado en ECOTOX (EPA) + ADORE

result_conc1_mean_binary

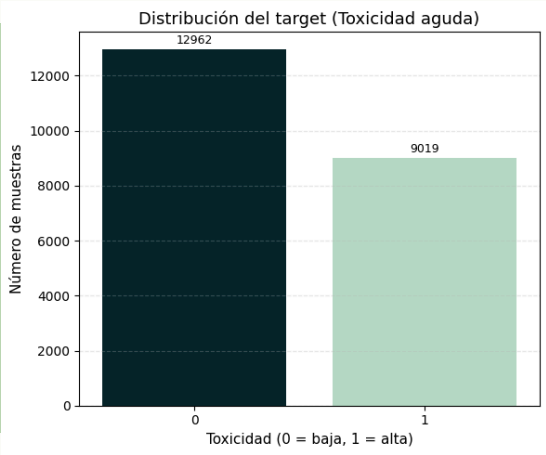
Target ⇧ 0

Etiqueta Binaria

split_occurrence

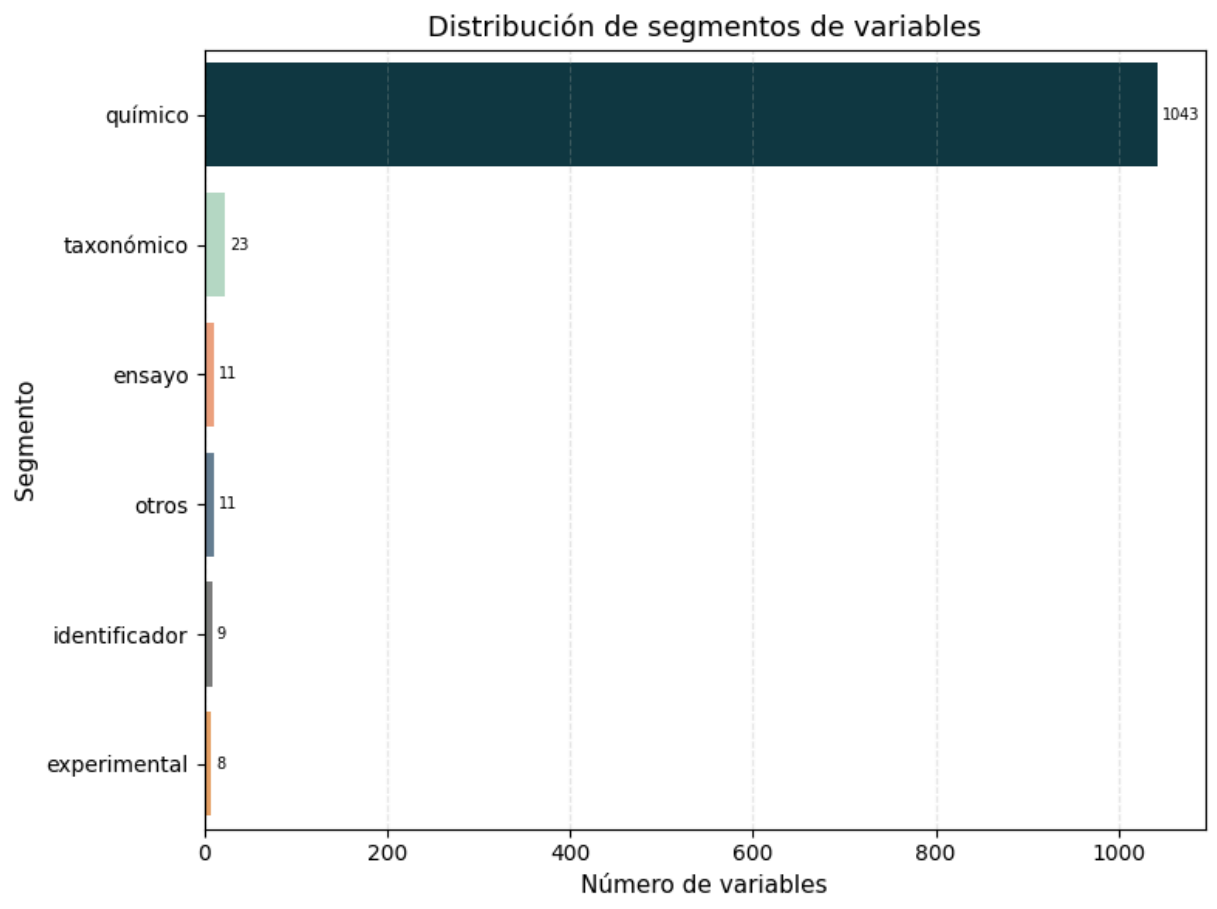
Train/Test

Basada en la ocurrencia química



Reducción de Features

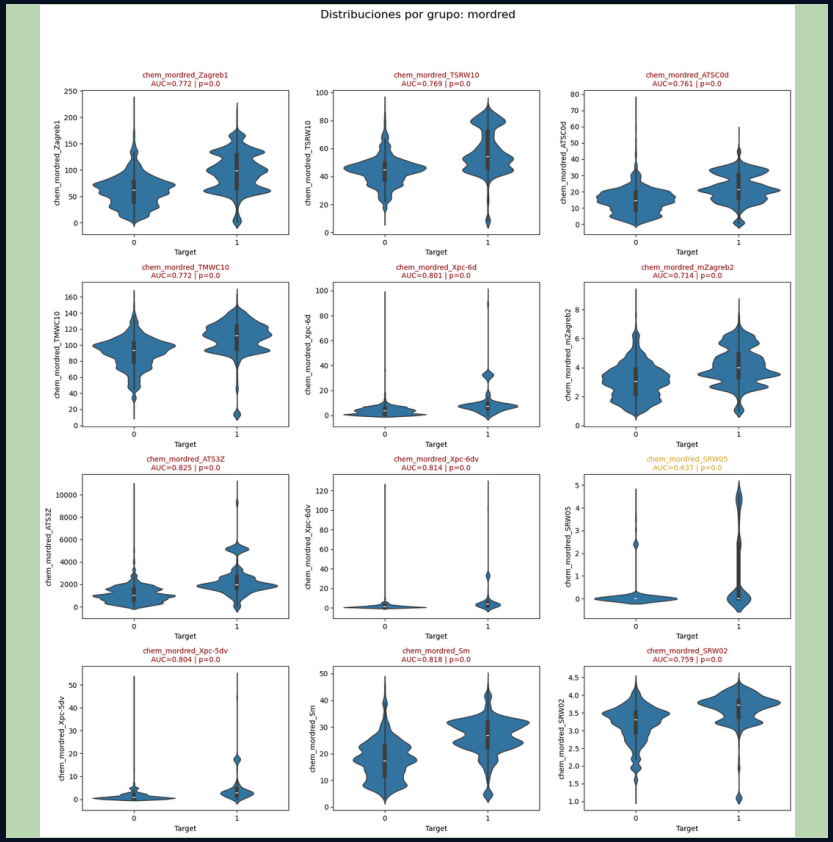
95 % de las más de mil variables corresponden a descriptores químicos



1. Reducción estructural
2. **Correlación** lineal con el target
3. Importancia mediante **Random Forest**
4. Test estadístico **SelectKBest**

Exploratory Data Analysis

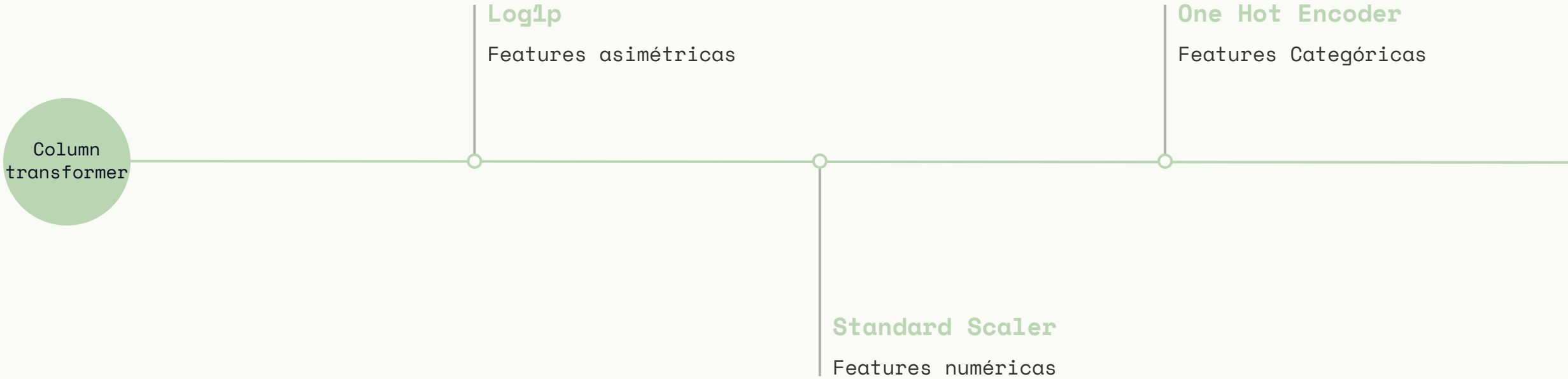
Subconjunto de variables numéricas y categóricas, seleccionadas por importancia estadística y conocimiento experto.



- ## Análisis Univariante
  - **Distribuciones asimétricas** y la presencia de outliers, sobre todo en **variables moleculares**.
- ## Análisis Bivariante
  - Poder predictivo frente al target mediante visualizaciones, **AUC y tests estadísticos**.
- ## 26 variables finales
  - **21 predictivas**: descriptores moleculares, condiciones del test y taxonomía
  - **5 contextuales**: pH, temperatura y rasgos biológicos del organismo.

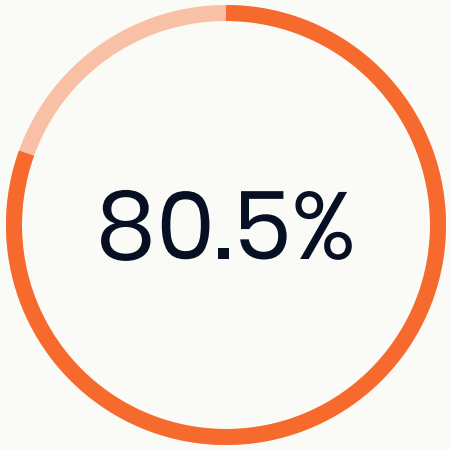
Pipeline

Preprocesador completo y una **regresión logística** como modelo **baseline**.



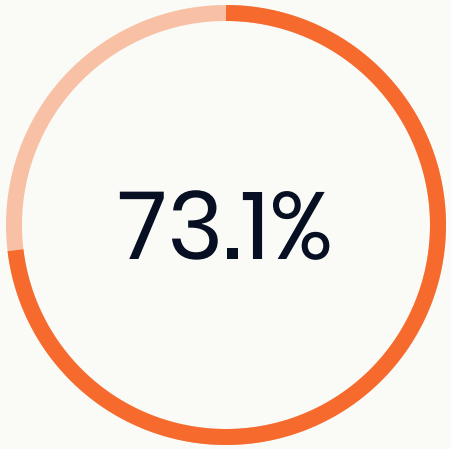
Validación del baseline

Validación cruzada estructurada siguiendo los **splits propuestos por ADORE**, lo que evita fugas químicas o taxonómicas entre entrenamiento y validación.



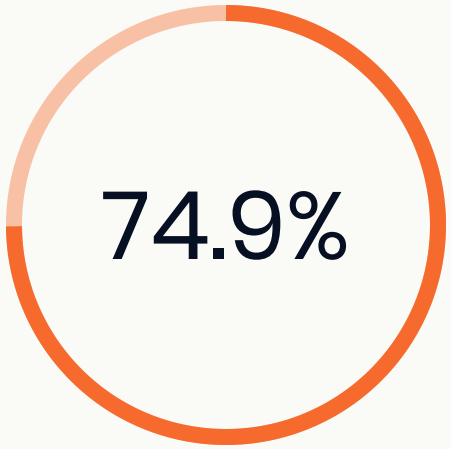
Accuracy promedio

Proporción de predicciones correctas entre todas las muestras.



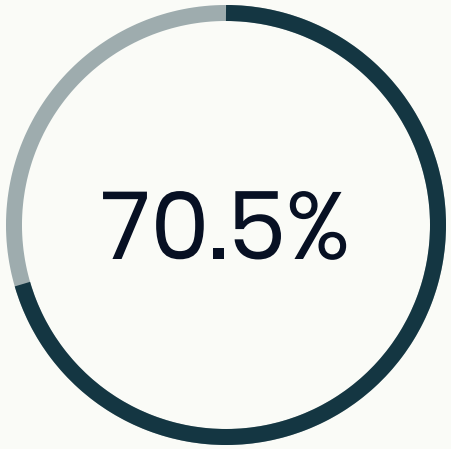
Recall

Capacidad del modelo para detectar sustancias tóxicas (positivos reales).



F1 Score

Equilibrio entre precisión y recall: penaliza los errores en ambas clases.



F1 Score en **test**

Desempeño final del modelo sobre datos completamente nuevos.

Comparativa de métricas

Modelo Baseline: Regresión Logística

Rendimiento estable

Random Forest (tuned):

- F1 (CV ADORE) **0.667**
- Accuracy (CV) –
- Recall (CV) –
- Evaluado en Test ❌ No

XGBoost (sin tuning):

- F1 (CV ADORE) **0.621**
- Accuracy (CV) 0.730
- Recall (CV) 0.542
- Evaluado en Test ❌ No

Logistic Regression:

- F1 (CV ADORE) **0.749**
- Accuracy (CV) 0.805
- Recall (CV) 0.731
- Evaluado en Test ✅ Sí

Modelo principal

Regresión Logística con ajuste de hiperparámetros

Random Forest y XGBoost

Ninguna mejoró consistentemente el rendimiento: presentaban **mayor variabilidad entre folds** y **peor F1 promedio**.

Variables Contextuales

No aportaron **valor predictivo** y se descartaron.

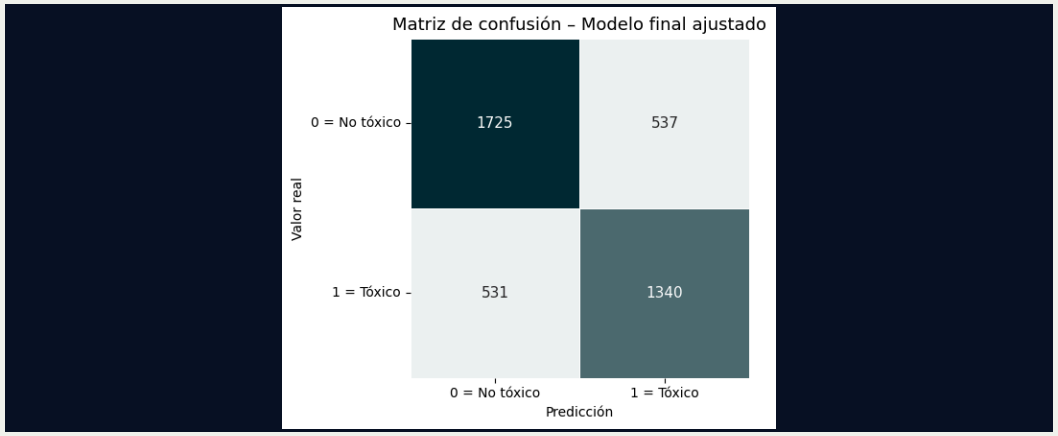
Regresión Logística

Hiperparámetros (*C=0.1*, *penalty='l2'*, *solver* = 'liblinear')

Validación según benchmark ADORE

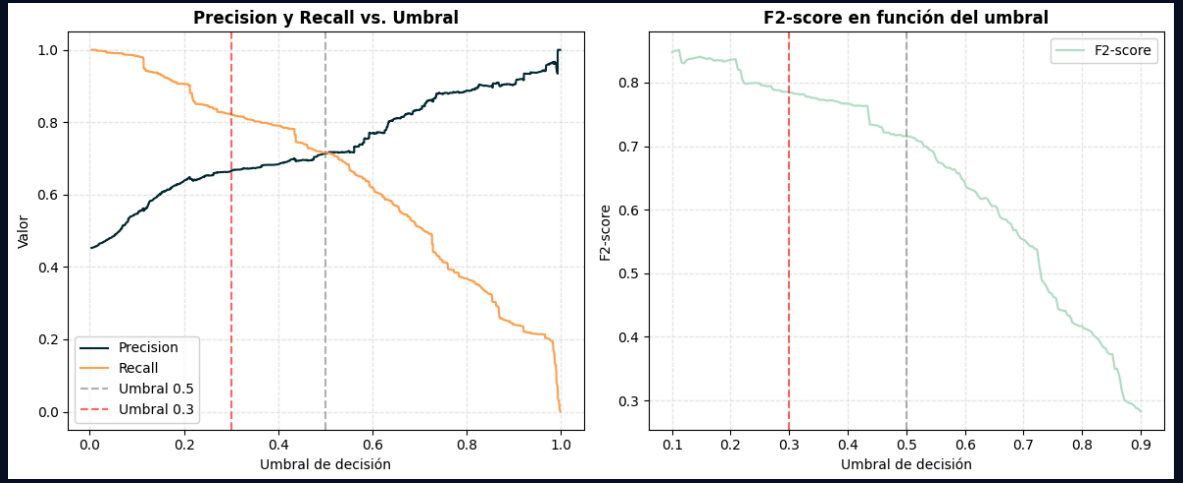
Regresión Logística

C=0.1, penalty='l2', solver = 'liblinear'



F1 : 71.5%

▲ Recall



Con umbral 0.3

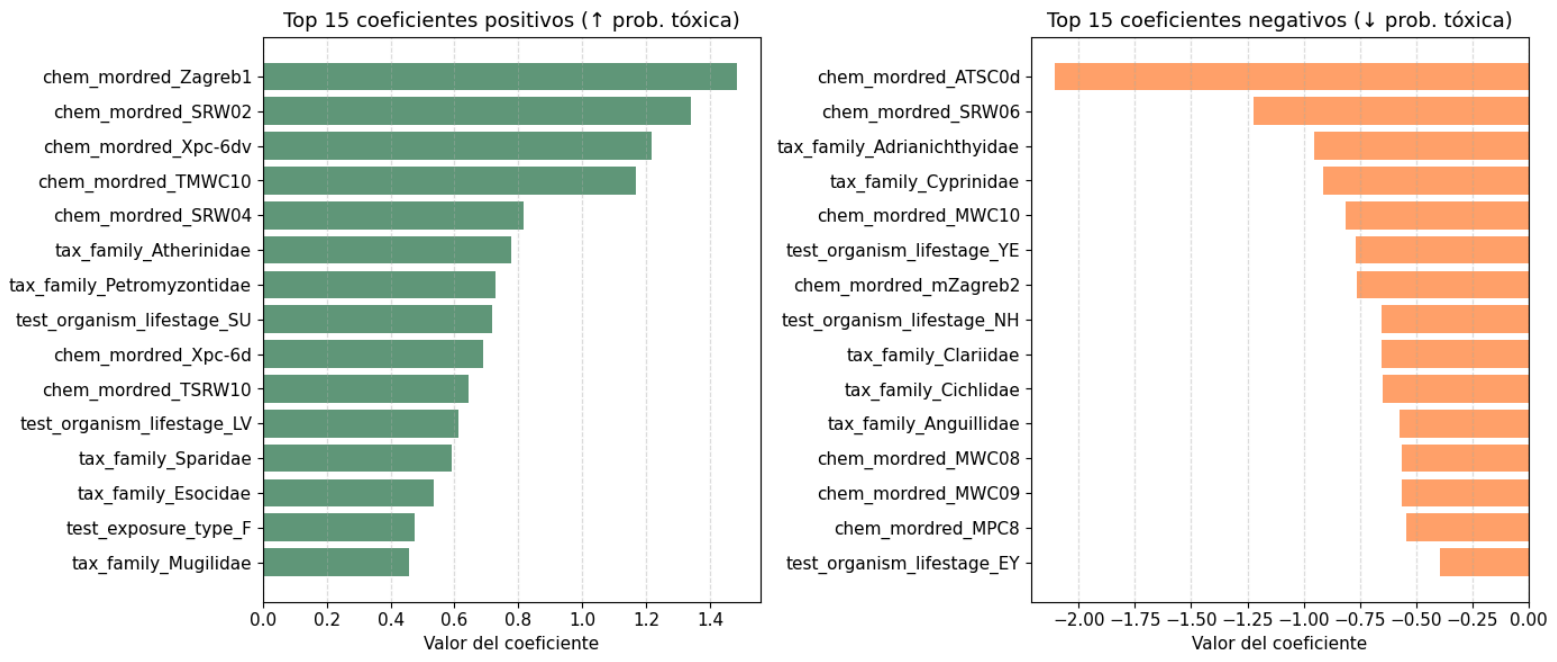
- Recall sube a 0.82
- Falsos negativos bajan un 37%, sin pérdida grave de precisión

Modelo flexible y adaptable

- Más estricto en entornos regulatorios
- Equilibrado en investigación.

Interpretación

Análisis de **coeficientes**:cada coeficiente representa el efecto de una variable sobre la probabilidad de toxicidad.



DESCRIPTORES QUÍMICOS

Zagreb1
Xpc-6dv
SRW02

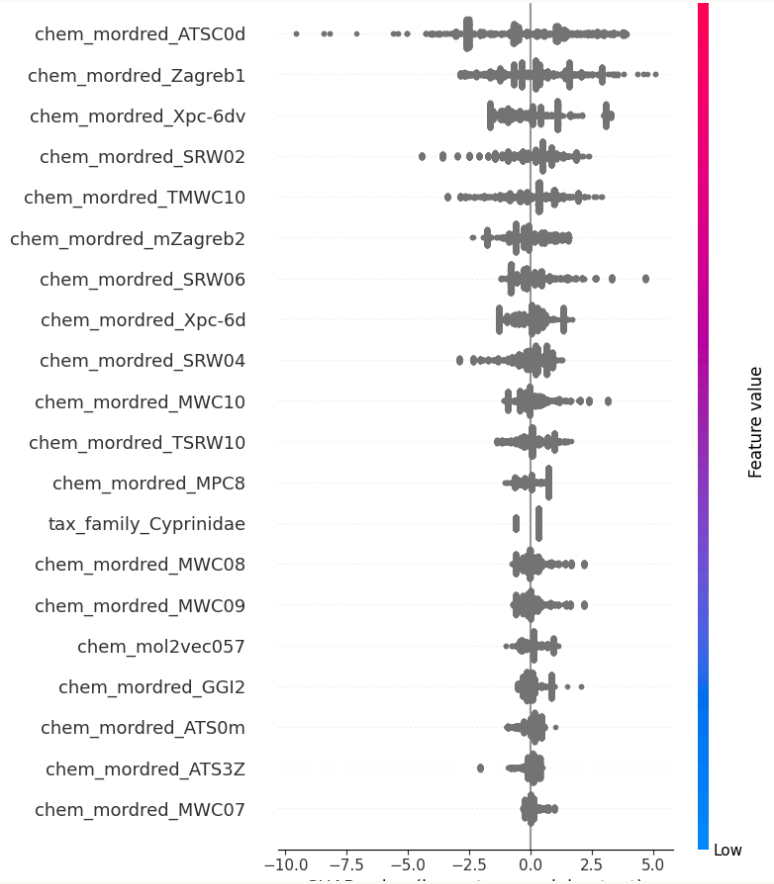
FACTORES BIOLÓGICOS

Etapa de vida del organismo

Familia taxonómica.

Interpretación: SHAP

Permite analizar el impacto individual de cada variable sobre cada predicción.



Summary plot: cada punto representa una muestra individual

Zagreb1
Xpc-6dv
SRW02

MISMAS VARIABLES QUÍMICAS

ATSC0d

VALORES BAJOS, MAYOR PROBABILIDAD DE TOXICIDAD

Mayor interpretabilidad a nivel individual

Interpretación: SHAP

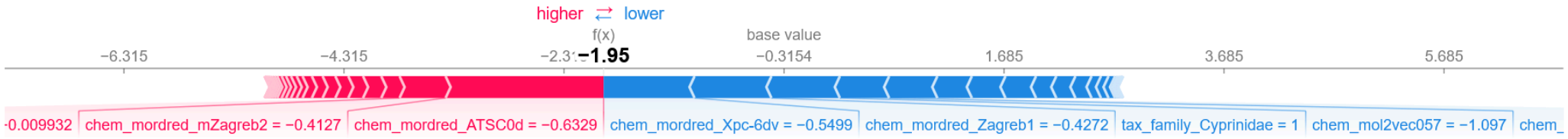
Permite analizar el impacto individual de cada variable sobre cada predicción.

¿Qué empujó al modelo a clasificar una sustancia como **tóxica**?

- Justificación de decisiones regulatorias.
- Priorización de sustancias para evaluación experimental.

Confianza

Sistema de apoyo de decisiones



Force plot

Conclusión

Construcción de un modelo **preciso, reproducible** e **interpretable** para **predecir la toxicidad aguda en peces**, reduciendo la necesidad de ensayos in vivo.

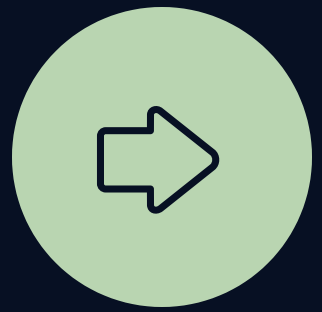
Regresión Logística

- F1-score de **71.5% en test**
- **Estabilidad entre folds**
- Comportamiento robusto ante datos nuevos

SHAP

Ajuste de umbral de decisión

- Adaptación del modelo según el riesgo aceptable
- Mayor utilidad en contextos regulatorios



API

- Recepción de estructuras químicas
- Devolución de predicciones
- Cribado temprano de sustancias
- Normativas REACH

Gracias

Hacia una ecotoxicología más ética, rápida y accesible