Explainable AI – Unidad 2: Evaluación de Modelos Glass-Box en la Predicción de Reincidencia Penal

Josep Gabriel Fornes Reynes
Jordi Florit Ensenyat
Juan Esteban Rincón Marín

Septiembre 2025

1 Introducción

1.1 Resumen del problema y enfoque

El objetivo de esta entrega es entrenar, comparar y explicar distintos modelos de *Machine Learning* para predecir la variable `recid`, que indica la reincidencia de un individuo tras una condena. El flujo de trabajo combina un análisis exploratorio inicial, el diseño de un `Pipeline` reproducible y la evaluación de tres enfoques **glass-box**: un modelo **lógico** (árbol de decisión), un modelo **logístico** (regresión logística) y un modelo **aditivo** (*Explainable Boosting Machine*, EBM). Cada uno representa una forma diferente de equilibrio entre interpretabilidad y capacidad predictiva.

El proyecto busca no solo optimizar el rendimiento predictivo, sino también estudiar la **explicabilidad global y local** de los modelos, analizando cómo cada uno construye sus decisiones y qué variables tienen mayor influencia en el resultado final.

1.2 Datos (características y procesamiento)

Se trabaja con el conjunto `recidivism.csv`, compuesto por observaciones individuales donde la variable objetivo `recid` es binaria (reincide / no reincide). El conjunto presenta un equilibrio de clases razonable (relación ≈ 1.14 entre clases), y combina variables **numéricas** (edad, número de antecedentes, delitos juveniles) y **categóricas** (raza, sexo, tipo de cargo, empleo).

El análisis exploratorio mostró correlaciones moderadas entre las variables (entre –0.2 y +0.3), sin indicios de multicolinealidad. Esto permitió utilizar tanto modelos logísticos como aditivos. Además, el estudio de linealidad evidenció que la edad y el número de antecedentes influyen en la reincidencia de forma no lineal, lo que justificó la incorporación del modelo EBM.

El preprocesamiento se encapsuló dentro de un `Pipeline` que separa variables numéricas (escaladas mediante `StandardScaler`) y categóricas (codificadas con `OneHotEncoder`). Se añadió una función personalizada para unificar variables redundantes derivadas de la codificación, generando comparaciones binarias más estables como `is_AfricanAmerican_vs_Caucasian`, también `is_Felony_vs_Misdemeanor` y `is_Male_vs_Female`. De este modo, el pipeline garantiza reproducibilidad, evita *data leakage* y aplica las mismas transformaciones durante entrenamiento y validación.

1.3 Resultados (visión general)

Los tres modelos ofrecen perspectivas complementarias sobre el problema. El **árbol de decisión** baseline sirvió como referencia y, tras el ajuste mediante `GridSearchCV`, alcanzó una **accuracy media de 0.719** en validación cruzada, manteniendo buena interpretabilidad y reglas compactas. La **regresión logística** mostró un rendimiento estable (**Train = 0.732, 5-fold = 0.702**) y coeficientes consistentes con la interpretación esperada: la edad y el empleo estable reducen la probabilidad de reincidencia, mientras que el número de antecedentes la incrementa. Finalmente, el **modelo aditivo (EBM)** se consolidó como el más equilibrado del estudio, con un **Train accuracy de 0.748** y una **5-fold CV accuracy de 0.723**, logrando la mejor generalización y capturando relaciones no lineales suaves sin perder interpretabilidad.

En conjunto, el EBM mostró la brecha más reducida entre entrenamiento y validación, indicando una excelente capacidad de generalización y manteniendo una explicabilidad global clara (curvas de efecto por variable) y local coherente con las tendencias generales. Esto lo posiciona como el modelo más completo para la predicción y análisis explicable de la reincidencia penal.

2 Metodología

2.0.1 Árbol baseline (referencia)

Como primer paso se entrena un `DecisionTreeClassifier` con parámetros por defecto dentro del `Pipeline`. Este modelo sirve como punto de partida para evaluar dos aspectos fundamentales: (i) si la representación y el preprocesado aplicados permiten al modelo aprender patrones relevantes sin una ingeniería de características adicional, y (ii) cuánto aporta realmente el ajuste de hiperparámetros al rendimiento y la interpretabilidad.

La elección de un árbol como baseline se justifica por su **transparencia inmediata** (estructura legible y reglas explícitas) y su bajo coste computacional, lo que facilita validar rápidamente el flujo completo de entrenamiento y evaluación.

2.0.2 Árbol ajustado (GridSearchCV)

En un segundo paso se ajusta el mismo clasificador mediante `GridSearchCV` con validación cruzada estratificada de 5 particiones. La búsqueda explora hiperparámetros que equilibran la complejidad y la legibilidad del modelo, concretamente:

- **`max_depth`**: limita la profundidad del árbol para evitar sobreajuste.
- **`min_samples_leaf`** y **`min_samples_split`**: controlan el tamaño mínimo de las hojas, estabilizando las reglas y mejorando la generalización.
- **`criterion` (gini, entropy, log_loss)**: distintos criterios de impureza; `log_loss` produce probabilidades más calibradas, mientras que `gini`/`entropy` son más eficientes.
- **`max_features` (None, sqrt, log2)**: actúa como regularizador al limitar el número de atributos por división.
- **`splitter` (best, random)**: `best` busca divisiones deterministas óptimas; `random` introduce aleatoriedad útil en algunos contextos exploratorios.

El ajuste se integra dentro del `Pipeline`, garantizando que los mismos pasos de preprocesamiento se apliquen en cada iteración de la validación cruzada, evitando cualquier *data leakage*. Este proceso permite comparar directamente el modelo base y el ajustado en términos de rendimiento y legibilidad, manteniendo la reproducibilidad completa.

2.0.3 Estudio de linealidad

Antes de entrenar los modelos logístico y aditivo, se realiza un estudio de la **linealidad de las variables** y sus relaciones con la variable objetivo. Para ello se calculan matrices de correlación entre variables numéricas y se representan las relaciones individuales con respecto a `recid`.

Este análisis tiene un doble propósito: por un lado, identificar variables cuya relación con el objetivo puede modelarse adecuadamente con un modelo logístico (como la regresión logística), y por otro, detectar aquellas que muestran **comportamientos no lineales**, donde modelos aditivos como el *GAM/EBM* podrían capturar mejor las tendencias.

2.0.4 Creación del dataset unificado

A partir de las conclusiones del estudio de linealidad, se observó que algunas variables presentaban **alta correlación o redundancia**. En particular, se identificaron pares de variables mutuamente excluyentes, generadas por la codificación `One-HotEncoder`, que transmitían esencialmente la misma información:

- `race_Caucasian` y `race_African-American`: una persona afroamericana no es caucásica y viceversa.
- `c_charge_degree_F` y `c_charge_degree_M`: el tipo de cargo criminal (felony/misdemeanor) solo puede pertenecer a una de las dos categorías.
- `sex_Male` y `sex_Female`: igualmente excluyentes entre sí.

Para reducir la redundancia y mejorar la estabilidad de los modelos logísticos y aditivos, se decidió crear un nuevo conjunto de datos aplicando una función de preprocesado personalizada. Esta función elimina una de las columnas redundantes y renombra la restante con una etiqueta más descriptiva, tal como se muestra a continuación:

- `race_African-American` $\rightarrow$ `is_AfricanAmerican_vs_Caucasian`
- `c_charge_degree_F` $\rightarrow$ `is_Felony_vs_Misdemeanor`
- `sex_Male` $\rightarrow$ `is_Male_vs_Female`

De esta forma, cada nueva variable representa explícitamente una comparación binaria, simplificando la interpretación de los coeficientes en los modelos logísticos y evitando problemas de multicolinealidad. Este preprocesamiento se integró dentro de un `Pipeline` mediante un `FunctionTransformer`, asegurando que las transformaciones se apliquen de manera coherente durante todas las fases de entrenamiento, validación y predicción.

2.0.5 Modelos logístico y aditivo

Tras la creación del dataset unificado, se entrenaron los modelos correspondientes a las otras dos familias explicables: el modelo **logístico** (Regresión Logística) y el modelo **aditivo** (*Explainable Boosting Machine*, EBM). Ambos se implementaron dentro de `Pipelines` equivalentes al del árbol, garantizando un flujo de preprocesado homogéneo y validación cruzada estratificada en 5 pliegues.

- **Modelo logístico (Regresión Logística)**: este modelo constituye una aproximación sencilla y robusta para problemas de clasificación binaria, en la que la predicción se basa en una combinación lineal de las variables explicativas. Cada coeficiente representa la contribución (positiva o negativa) de una variable a la probabilidad de reincidencia, lo que permite una interpretación directa y cuantitativa de su influencia.

  En este contexto, la regresión logística sirve como punto de referencia para evaluar hasta qué punto la relación entre las variables del conjunto `recidivism.csv` puede aproximarse mediante dependencias lineales. Su principal ventaja es la **transparencia global**: los coeficientes del modelo se pueden ordenar y comparar fácilmente, identificando qué factores aumentan o reducen la probabilidad de reincidencia. Sin embargo, su capacidad para capturar relaciones no lineales o interacciones entre variables es limitada, lo que puede reducir su rendimiento frente a modelos más flexibles.

- **Modelo aditivo (GAM / Explainable Boosting Machine)**: el modelo aditivo generalizado (*Generalized Additive Model*, GAM), que ha sido implementado con la clase `ExplainableBoostingClassifier` de la librería `interpret.glassbox`, combina interpretabilidad y capacidad predictiva en una única estructura. A diferencia del modelo logístico, el GAM no asume que las relaciones entre variables y el objetivo sean lineales: cada variable se modela de manera independiente mediante una función de forma libre (no paramétrica), y el resultado final se obtiene sumando los efectos de todas las variables.

  Esta arquitectura permite capturar **relaciones no lineales suaves** —por ejemplo, umbrales de edad o saturación de efectos por número de antecedentes— sin perder la capacidad de explicar los resultados. Además, los efectos individuales de cada variable pueden visualizarse fácilmente en forma de curvas, lo que aporta una interpretación global clara y comparable con los coeficientes de un modelo logístico.

  En el proyecto se probaron dos variantes: una sin interacciones (GAM puro) y otra con un número limitado de interacciones entre pares de variables (`interactions`=5). Esta comparación permite analizar el **compromiso entre simplicidad e incremento de rendimiento**: el modelo con interacciones tiende a captar dependencias más complejas, pero puede volverse menos legible.

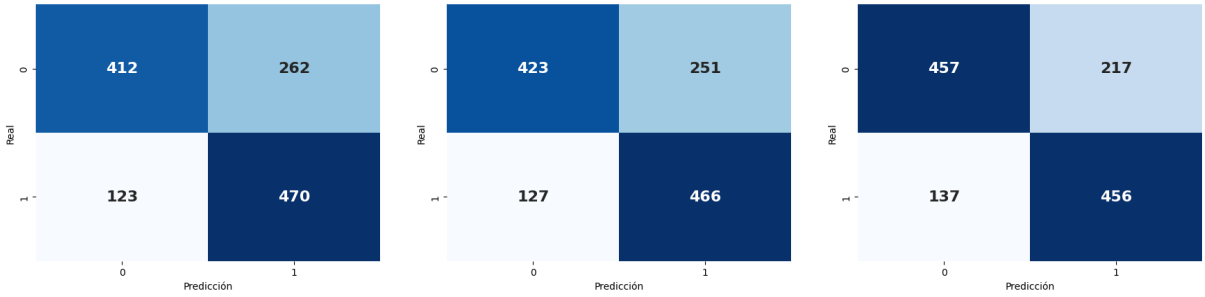
Para ambos modelos, el proceso de búsqueda de hiperparámetros incluyó rejillas con variaciones de `C` y `penalty` en la regresión logística, y de `learning_rate`, `max_bins` e `interactions` en el EBM. En todos los casos se empleó una validación cruzada estratificada de cinco pliegues (`StratifiedKFold`) para garantizar comparaciones justas entre modelos y evitar sesgos por desbalanceo de clases.

3 Resultados

Lectura breve. El modelo **DecisionTree (baseline)** mostró un rendimiento inicial modesto, sirviendo como referencia para validar el flujo de trabajo. El **DecisionTree (tuned)** logró una

Classifier	Best parameters	Train acc	5-fold CV accuracy
DecisionTree (baseline)	*defaults*	0.670	0.659
DecisionTree (tuned)	`criterion=gini, max_depth=3, min_samples_leaf=1, min_samples_split=2, max_features=None, class_weight=None, splitter=best`	0.700	0.719
LogisticRegression	`C=1.0, penalty='l2', solver='lbfgs'`	0.732	0.702
ExplainableBoostingMachine (GAM)	`learning_rate=0.05, max_bins=256, interactions=5`	0.748	0.723

Table 1: Resumen de resultados por modelo. Los valores de *Train acc* y *5-fold CV accuracy* reflejan la media sobre el conjunto de entrenamiento y la validación cruzada, respectivamente.



(a) Árbol de decisión (ajustado) (b) Regresión logística (c) GAM / EBM

Figure 1: Comparativa de matrices de confusión normalizadas (0–1) para los tres modelos. El EBM presenta la mayor proporción de verdaderos positivos y falsos negativos reducidos.

mejora significativa, con una **5-fold CV accuracy de 0.719**, demostrando que un árbol poco profundo y bien regularizado puede alcanzar una generalización estable.

La **Regresión Logística** ofreció un rendimiento muy equilibrado (**Train = 0.732, 5-fold = 0.702**), mostrando coeficientes coherentes con la interpretación esperada: la edad y el empleo estable reducen la probabilidad de reincidencia, mientras que el número de antecedentes la incrementa.

Por último, el modelo **aditivo (GAM / EBM)** se consolidó como el más robusto del estudio (**Train = 0.748, 5-fold = 0.723**), manteniendo una diferencia mínima entre entrenamiento y validación. Este comportamiento indica una alta capacidad de generalización, combinando explicabilidad global (curvas de efecto) con consistencia local en las predicciones. A nivel de error, el EBM reduce falsos negativos sin aumentar excesivamente los falsos positivos, lo que lo posiciona como el modelo más equilibrado en rendimiento e interpretabilidad.

4 Modelos

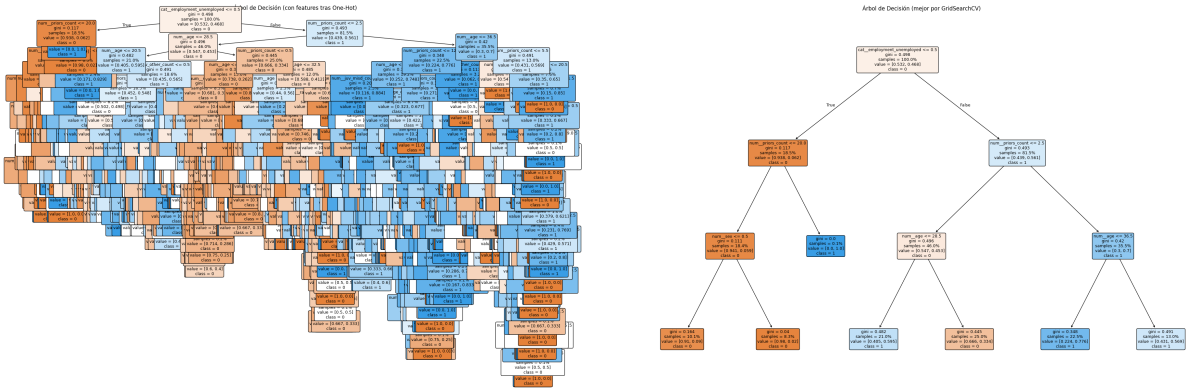
4.1 Comparación del árbol baseline y ajustado

El modelo **baseline**, entrenado con los parámetros por defecto, sirvió como referencia para validar el flujo de datos, el preprocesamiento y la capacidad del árbol para capturar patrones básicos en el conjunto `recidivism.csv`. En validación, el baseline alcanzó una **Accuracy (valid) = 0.659**, mostrando una estructura más profunda y fragmentada, con ramas menos coherentes entre sí y una ligera tendencia al sobreajuste. Aun así, permitió comprobar que las variables más relevantes —principalmente el número de antecedentes y la situación laboral— aparecían

de forma recurrente en las divisiones iniciales, lo que validó el correcto preprocesamiento de los datos.

Tras aplicar `GridSearchCV` con validación estratificada de 5 pliegues, el árbol **ajustado** (*tuned*) logró un mejor compromiso entre profundidad, precisión y legibilidad. El modelo redujo su complejidad a aproximadamente tres niveles y mejoró la consistencia de las reglas, alcanzando una **Accuracy (test) ≈ 0.70** y una **5-fold CV accuracy ≈ 0.719**. Las divisiones principales se concentraron en las variables `priors_count`, `employment_unemployed` y `age`, reflejando una jerarquía de decisión más lógica y estable.

En conjunto, el modelo ajustado mantiene la claridad visual del baseline, pero con reglas más compactas, generalización superior y un comportamiento más equilibrado entre clases, lo que confirma que un control adecuado de hiperparámetros puede mejorar simultáneamente la interpretabilidad y el rendimiento.



(a) Árbol baseline.

(b) Árbol ajustado (GridSearch).

Figure 2: Comparativa visual de los árboles baseline y ajustado. El modelo ajustado presenta menor profundidad y una jerarquía de decisiones más estable.

4.2 Estudio de linealidad y unificación de variables

El análisis de correlaciones mostró valores moderados (entre -0.2 y $+0.3$), lo que indica la **ausencia de multicolinealidad fuerte** y sugiere que las variables son en general independientes. Esto permite aplicar tanto modelos logísticos como aditivos sin comprometer la estabilidad de los coeficientes.

Con el objetivo de profundizar en las relaciones entre las variables numéricas y la probabilidad de reincidencia, se realizó un **estudio de linealidad** para observar el comportamiento individual de cada predictor frente a la variable objetivo. La Figure 3 muestra cómo, aunque algunas relaciones son aproximadamente proporcionales, otras presentan patrones no lineales que justifican el uso de un modelo aditivo (GAM).

- **age**: relación inversa clara con la reincidencia; a menor edad, mayor probabilidad de reincidencia.
- **priors_count**: efecto creciente con tendencia a saturarse a partir de ∼20 antecedentes.
- **juv_fel_count**: leve relación positiva, aunque con dispersión limitada.

- **juv_misd_count**: comportamiento no lineal e irregular, reflejando un posible efecto umbral.
- **juv_other_count**: correlación débil y poco estructurada con el objetivo.

Estos resultados confirman que **la edad y los antecedentes previos** son los principales determinantes de la reincidencia, pero su relación no es estrictamente lineal. Por ello, además del modelo logístico, se incorporó un modelo aditivo capaz de capturar estas transiciones suaves de riesgo.

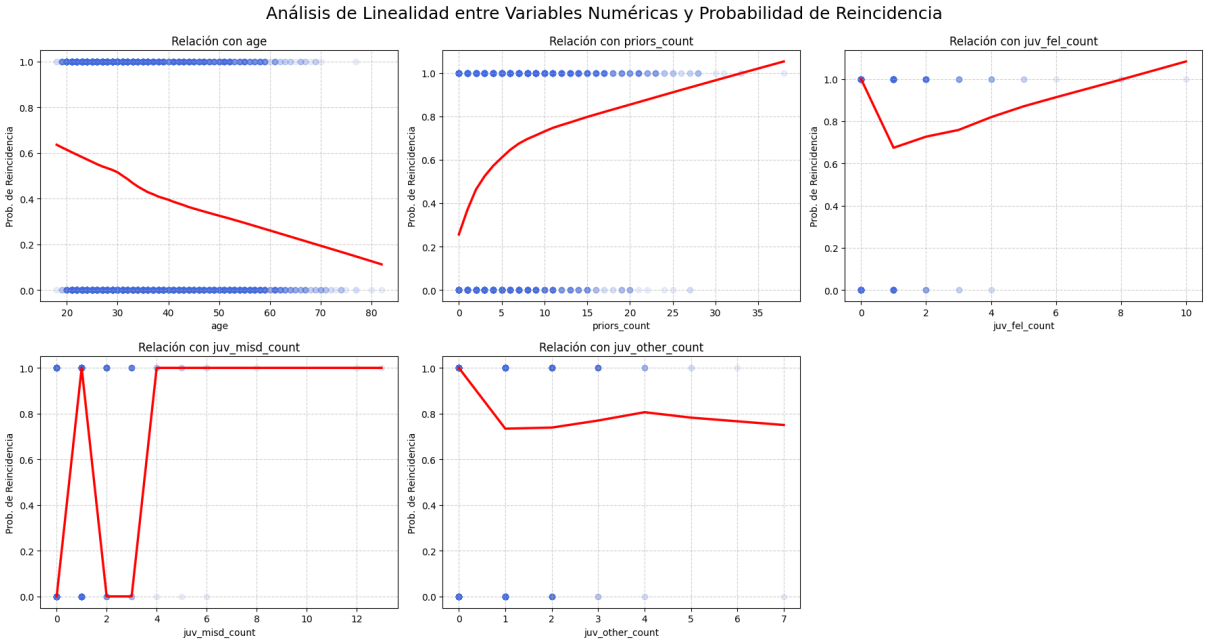


Figure 3: Análisis de linealidad entre variables numéricas y probabilidad de reincidencia. Las líneas rojas representan la tendencia media (efecto parcial) y los puntos azules los valores observados. Se observa un patrón no lineal claro en `priors_count` y `age`, lo que motiva el uso del modelo aditivo (GAM).

Complementariamente, se analizó la **tasa media de reincidencia para variables categóricas y binarias**, con el fin de comprender mejor el comportamiento de los grupos sociales y judiciales presentes en el conjunto de datos. Como muestra la Figure 4, los individuos desempleados presentan una tasa de reincidencia notablemente superior (0.56) frente a los empleados (0.06), mientras que las diferencias por sexo y tipo de cargo son más moderadas pero consistentes. Asimismo, el grupo `is_AfricanAmerican_vs_Caucasian` muestra una ligera mayor tasa de reincidencia en la población afroamericana (0.52 frente a 0.40), coherente con los patrones descritos en la literatura previa sobre sesgos y factores socioeconómicos.

Estos patrones refuerzan la idea de que los factores socioeconómicos y demográficos condicionan la probabilidad de reincidencia, y justifican la unificación de variables redundantes para facilitar la interpretación de los modelos explicativos.

Asimismo, se detectaron **variables redundantes** derivadas de la codificación One-HotEncoder, como race_Caucasian / race_African-American, c_charge_degree_F / c_charge_degree_M, y sex_Male / sex_Female. Estas variables fueron unificadas mediante una función de preprocesado que las combinó en comparaciones binarias más interpretables: is_AfricanAmerican_vs_Caucasian,

is_Felony_vs_Misdemeanor y is_Male_vs_Female, reduciendo redundancias y mejorando la estabilidad del modelo logístico.

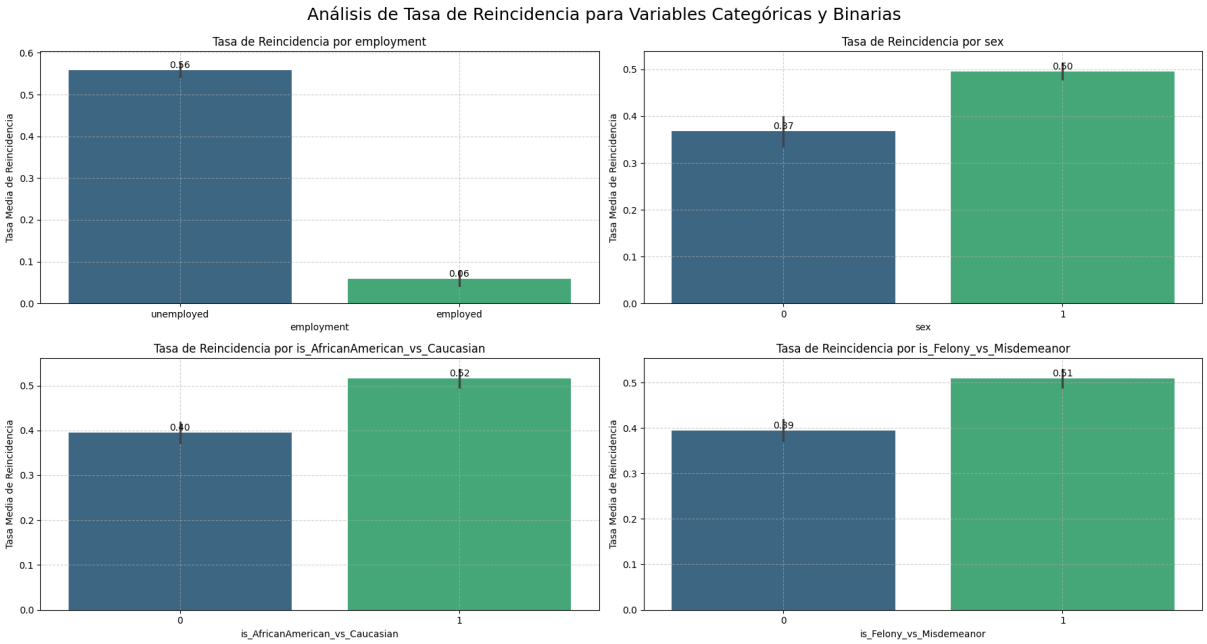


Figure 4: Tasa media de reincidencia por variables categóricas y binarias. Se observa una diferencia marcada por empleo, con tasas de reincidencia más altas entre personas desempleadas. Las diferencias por sexo, raza y tipo de cargo son más suaves pero mantienen coherencia en la dirección del efecto.

4.3 Modelo logístico

El modelo **logístico** actúa como una referencia intermedia entre el árbol y el GAM. Ofrece una lectura directa de los coeficientes (signo y magnitud) y permite cuantificar el impacto de cada variable sobre la probabilidad de reincidencia. En validación se observó una **Accuracy (valid)** $\approx$ **0.70** y un comportamiento equilibrado entre clases. Los coeficientes positivos se asocian a `priors_count` y a la condición de desempleo, mientras que los negativos a `age` y a la categoría de empleo estable, que funcionan como factores protectores.

El ajuste mediante `GridSearchCV` exploró distintos valores de regularización `C`, penalización `L2` y solvers, identificando un modelo estable y bien calibrado. Aunque su estructura lineal limita la capacidad de capturar interacciones o efectos no lineales, el modelo destaca por su **claridad interpretativa y consistencia estadística**, manteniendo una brecha reducida entre entrenamiento y validación.

4.4 Modelo aditivo (GAM / EBM)

El **GAM/EBM**, implementado mediante `ExplainableBoostingClassifier`, combina la flexibilidad de los modelos no lineales con la interpretabilidad de los modelos aditivos. Cada predictor se modela mediante una función de forma libre y el resultado se obtiene como la suma de los efectos individuales. En validación, el EBM alcanzó una **Accuracy (valid)** $\approx$ **0.72**, mejorando tanto al árbol ajustado como al modelo logístico.

Las curvas globales del EBM muestran un patrón coherente con el análisis exploratorio: el riesgo de reincidencia aumenta con el número de antecedentes y disminuye con la edad, de forma no lineal. La versión con un número limitado de interacciones (`interactions = 5`) aportó una ligera mejora en precisión sin comprometer la explicabilidad, destacando combinaciones intuitivas como edad y antecedentes.

En conjunto, el modelo aditivo se consolidó como el **más equilibrado** del estudio: logró la mejor generalización, explicó correctamente las tendencias observadas y mantuvo una estructura fácilmente interpretable tanto a nivel global (efectos parciales) como local (contribuciones por individuo).

Lectura conjunta. El **árbol ajustado** supera al baseline y estabiliza las reglas de decisión; el **modelo logístico** aporta claridad interpretativa y resultados consistentes; y el **GAM/EBM** logra el mejor equilibrio entre rendimiento y explicabilidad. La progresión entre los tres modelos refleja cómo un incremento gradual en complejidad estructural permite capturar mejor los patrones del fenómeno sin perder transparencia ni robustez.

5 Discusión: Explicabilidad global y local

5.1 Explicabilidad global

A nivel global, los tres modelos presentan distintos grados de transparencia y profundidad interpretativa, cada uno con fortalezas específicas según su estructura.

En el **árbol de decisión**, la explicabilidad global se manifiesta de forma visual y directa a través de las reglas. Su estructura jerárquica permite identificar con claridad las variables dominantes: la situación laboral, el número de antecedentes y la edad. El árbol ajustado consolida esta jerarquía en una forma fácilmente interpretable, situando la **condición de empleo** como nodo raíz y primera división crítica. Los individuos desempleados presentan una mayor probabilidad de reincidencia, mientras que una situación laboral estable se asocia con un menor riesgo, reflejando un patrón coherente con la literatura sobre reincidencia.

(Figure 5).

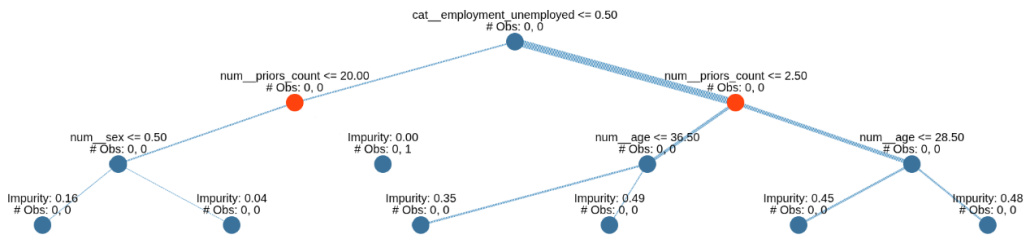


Figure 5: Importancia global de las variables en el árbol ajustado. Se observa la predominancia de las variables relacionadas con el empleo, los antecedentes y la edad.

El **modelo logístico** mantiene una interpretabilidad transparente, pero desde un enfoque más cuantitativo. La magnitud y el signo de los coeficientes permiten medir la contribución exacta de cada variable: el desempleo y los antecedentes incrementan la probabilidad de reincidencia, mientras que el empleo y la edad actúan como factores protectores. Esta relación se

aprecia en la Figure 6, donde se muestra el impacto global de cada característica.

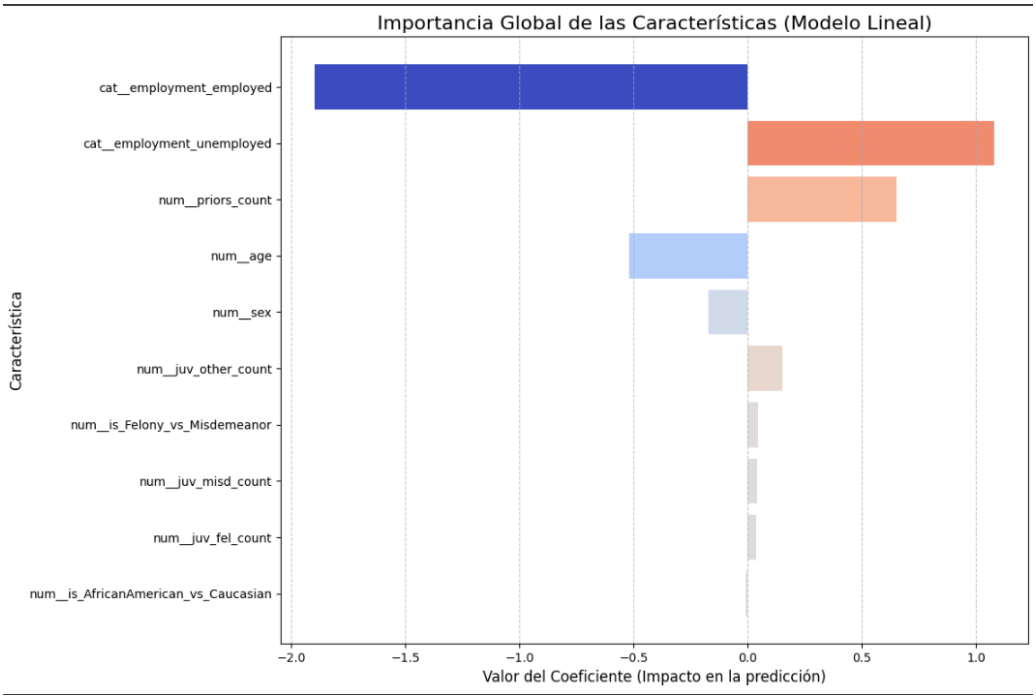


Figure 6: Importancia global de las características en el modelo logístico. Los coeficientes negativos (en azul) representan factores protectores frente a la reincidencia.

Por su parte, el **GAM/EBM** se posiciona como el modelo con la explicabilidad más rica. A diferencia del árbol o del modelo lineal, el EBM permite observar el efecto de cada variable en forma de curvas de impacto, que reflejan transiciones suaves y umbrales de cambio. El empleo vuelve a ser la variable dominante, seguida de los antecedentes y la edad, confirmando las tendencias observadas en el análisis exploratorio. La Figure 7 muestra la importancia global de cada término y las interacciones más relevantes detectadas.

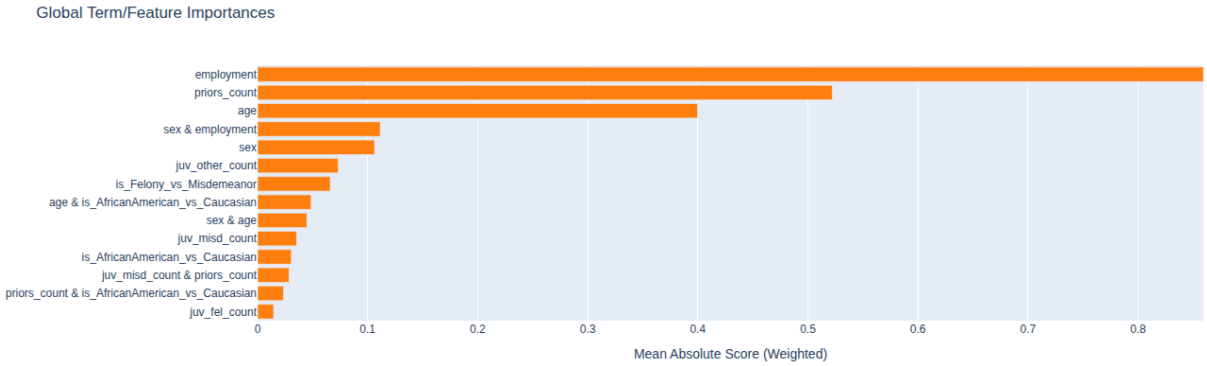


Figure 7: Importancia global de las características y términos en el modelo EBM. Se observan los mayores pesos asociados al empleo, los antecedentes y la edad, junto a interacciones significativas como sexo–empleo.

5.2 Explicabilidad local

A nivel local, los tres modelos permiten analizar cómo llegan a sus predicciones para casos individuales, pero lo hacen de formas muy distintas: el árbol mediante rutas de decisión explícitas, la regresión logística a través de pesos lineales y el EBM descomponiendo la predicción en contribuciones no lineales.

En el **árbol de decisión**, la explicación de una predicción individual se obtiene recorriendo los nodos hasta la hoja final. Cada condición define un punto de decisión que conduce a la clase final. La Figure 8 muestra un ejemplo real de esta lógica: el modelo predice reincidencia para un individuo desempleado, joven y con varios antecedentes. Las ramas coloreadas indican los nodos activos en la decisión y la intensidad del riesgo asociado a cada condición. Este tipo de representación facilita una lectura intuitiva y clara del razonamiento del modelo, aunque pierde precisión en escenarios con reglas más difusas o variables correlacionadas.

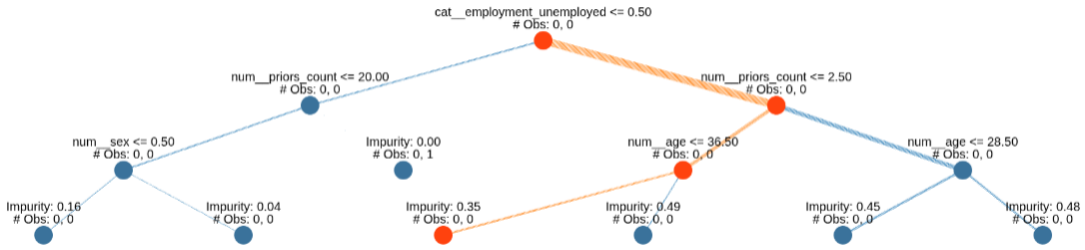


Figure 8: Explicabilidad local en el árbol de decisión. Se observa la ruta de nodos que lleva a la predicción final (reincidencia), con las divisiones activas resaltadas en color.

En la **regresión logística**, la explicabilidad local se basa en la suma ponderada de los coeficientes globales. Cada variable contribuye de manera positiva o negativa al logit de salida, lo que permite entender cómo se acumulan los efectos individuales. En la Figure 9, correspondiente a un individuo con tres antecedentes, edad media y situación de desempleo, las barras naranjas reflejan los factores que incrementan la probabilidad de reincidencia, mientras que las azules indican efectos protectores. El modelo muestra un comportamiento coherente con su lógica global: el desempleo y los antecedentes dominan la decisión, mientras que la edad ejerce un efecto amortiguador.

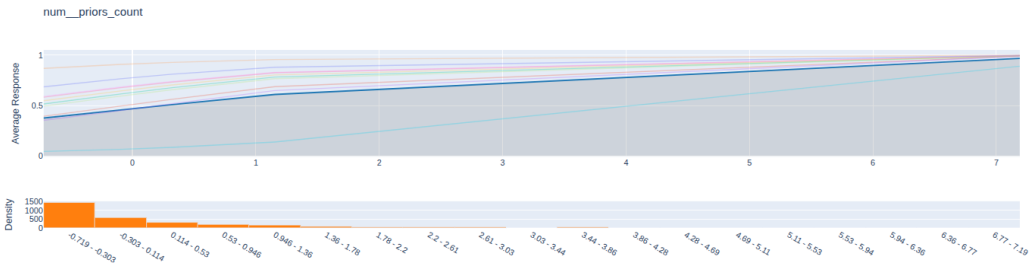


Figure 9: Explicabilidad local del modelo logístico para un caso individual. Las contribuciones positivas (naranja) aumentan la probabilidad de reincidencia, mientras que las negativas (azul) la reducen.

El **modelo aditivo (GAM / EBM)** ofrece la explicabilidad local más rica y matizada. A diferencia del modelo lineal, las contribuciones de cada variable no son constantes, sino depen-

dientes de su rango de valores. La Figure 10 muestra un ejemplo donde el modelo predice reincidencia con una probabilidad del 65.6%. El empleo (desempleado) y el número de antecedentes son los principales factores de riesgo, mientras que la edad actúa como un amortiguador parcial. Esta capacidad de modelar relaciones no lineales y explicarlas visualmente distingue al EBM como el modelo más transparente y completo a nivel local.

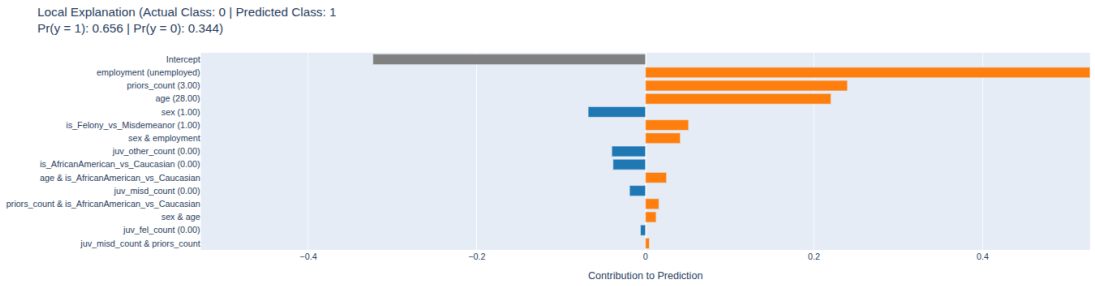


Figure 10: Explicabilidad local del modelo EBM. Cada barra representa el impacto de una variable en la predicción individual, con una descomposición no lineal que refleja tanto magnitud como dirección del efecto.

Síntesis. En conjunto, los tres modelos presentan distintos niveles de granularidad en su capacidad explicativa: el árbol destaca por su claridad lógica, la regresión logística por su coherencia estadística, y el EBM por su riqueza interpretativa y precisión local. Esta progresión evidencia cómo un incremento gradual en la complejidad del modelo puede mejorar la calidad de las explicaciones sin comprometer su transparencia.

5.3 Comparativa general de interpretabilidad

La comparación global y local entre los tres modelos muestra un gradiente claro de complejidad e interpretabilidad:

- El **árbol de decisión** destaca por su **transparencia estructural**: es el modelo más fácil de entender, aunque menos flexible. Es ideal para justificar decisiones concretas o comunicar reglas de forma visual, pero sufre cierta pérdida de estabilidad cuando las relaciones entre variables son más sutiles.
- La **regresión logística** proporciona una **interpretabilidad cuantitativa y estable**, permitiendo medir la influencia de cada variable de manera directa a través de sus coeficientes. Sin embargo, su estructura estrictamente lineal le impide capturar bien las no linealidades o interacciones presentes en los datos. Es, por tanto, el modelo más "matemáticamente explicable", aunque con menor capacidad de adaptación.
- El **GAM/EBM** combina ambas virtudes: mantiene una representación clara —basada en efectos univariantes y posibles interacciones— y logra la mejor precisión sin perder legibilidad. Su estructura modular permite analizar cada variable por separado, ofreciendo explicaciones globales mediante curvas de efecto y locales mediante contribuciones individuales. Esto lo convierte en el modelo más equilibrado entre precisión y transparencia.

En resumen, los tres enfoques demuestran que es posible alcanzar distintos equilibrios entre explicabilidad y rendimiento. El árbol ofrece comprensión inmediata, el modelo logístico aporta estabilidad y coherencia estadística, y el GAM/EBM integra flexibilidad y transparencia, consolidándose como la alternativa más completa del estudio en términos de interpretabilidad integral.

6 Conclusión

El estudio comparó tres modelos explicables para la predicción de reincidencia penal: un árbol de decisión, una regresión logística y un modelo aditivo (Explainable Boosting Machine, EBM). Los resultados muestran que todos alcanzan un rendimiento competitivo, con **accuracies medias de 0.719, 0.702 y 0.723** respectivamente, manteniendo una buena estabilidad entre entrenamiento y validación. El uso de un pipeline reproducible garantizó coherencia en el preprocesamiento y permitió comparar de forma justa la capacidad predictiva y explicativa de cada enfoque.

A nivel de **explicabilidad global**, los tres modelos coincidieron en resaltar las mismas variables clave: la situación laboral, el número de antecedentes y la edad. El árbol ofreció una lectura jerárquica y visual a través de reglas de decisión claras, la regresión logística tradujo las relaciones en coeficientes de interpretación directa y el EBM amplió esta visión mediante curvas de efecto que capturan relaciones no lineales con notable claridad. Esta convergencia entre modelos refuerza la fiabilidad de los patrones identificados y demuestra la coherencia del conjunto de datos.

En cuanto a la **explicabilidad local**, cada modelo aportó una perspectiva complementaria: el árbol permite seguir paso a paso la ruta de decisión de un individuo, la regresión logística descompone las predicciones en contribuciones lineales de cada variable, y el EBM combina ambas visiones, mostrando de forma visual cómo cada factor incrementa o reduce el riesgo individual de reincidencia. Esta capacidad de justificar cada decisión con detalle convierte al EBM en la opción más completa desde el punto de vista explicativo.

En conjunto, los resultados confirman que la transparencia y la precisión pueden coexistir. El árbol destaca por su claridad interpretativa, la regresión logística por su consistencia estadística y el EBM por integrar ambas virtudes, alcanzando el mejor equilibrio entre rendimiento, estabilidad y capacidad explicativa. Por todo ello, el EBM se posiciona como el modelo más sólido del estudio, al lograr unir de forma efectiva la **precisión predictiva y la explicabilidad**, demostrando que la inteligencia artificial puede ser al mismo tiempo rigurosa y comprensible.

Referencias

- Josep Gabriel Fornes Reynes; Jordi Florit Ensenyat; Juan Esteban Rincón Marín. *XAI – Proyecto Tema 2 (Repositorio)*. GitHub. Disponible en: https://github.com/PepBiel/IAExplicable_Proyectos/tree/main/Proyecto_Tema2 (acceso: Septiembre 2025).