Interpretability of classification models

Trabajo fin de grado en Matemáticas

Gregorio Blázquez Martínez
Tutor: Patrizio Guagliardo
Co-Tutor: Damián Álvarez Piqueras
Tutora Académica: Ana María Vargas Rey

Universidad Autónoma de Madrid

Junio 2024

Índice

Introducción

Contexto

- Los modelos de clasificación son esenciales en diversos campos como la economía y la salud.
- La precisión de los modelos ha mejorado gracias a los avances en aprendizaje automático.
- La interpretabilidad es crucial para la confianza y la adopción de estos modelos.

Contexto y Objetivos

Contexto

- Los modelos de clasificación son esenciales en diversos campos como la economía y la salud.
- La precisión de los modelos ha mejorado gracias a los avances en aprendizaje automático.
- La interpretabilidad es crucial para la confianza y la adopción de estos modelos.

Objetivos

- Proveer una base formal para la interpretabilidad de modelos de clasificación.
- Enfocarse en el método SHAP para explicar decisiones del modelo.
- Demostrar la aplicabilidad de SHAP en un caso real.

Modelo e Interpretabilidad

Modelo:

Sea $X = \{X_1, X_2, ..., X_n\}$ un conjunto de variables independientes (predictoras) y Y una variable dependiente (predicción) que se pretende explicar. Denominamos modelo a la función $f : \mathbb{R}^n \rightarrow \mathbb{R}$ que mapea un vector de características $\mathbf{x} = (x_1, x_2, ..., x_n)$ a un valor y que representa la variable dependiente. Lo formalizamos como:

$$Y = f(X_1, X_2, ..., X_n) + \epsilon$$

donde ϵ es un término de error que captura la variabilidad no explicada por el modelo.

Interpretabilidad: Aceptamos el uso extendido de este concepto como la capacidad de entender la predicción de un modelo en función de las variables independientes.

Modelos Interpretables

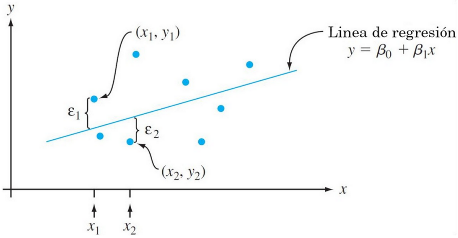
Regresión Lineal

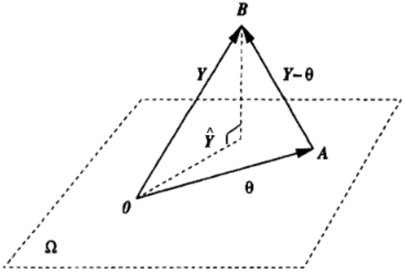
- Modelo.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Ajuste por mínimos cuadrados.

$$\hat{\boldsymbol{\beta}} = \arg \min_{\beta_0,\ldots,\beta_n} \sum_{i=1}^{k} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{n} \beta_j x_j^{(i)} \right) \right)^2$$





Interpretabilidad

La regresión lineal es altamente interpretable:

- Los coeficientes β_j representan el cambio en la variable dependiente al cambiar en una unidad la variable independiente x_j.
- Un β_j positivo indica una relación directa, mientras que un β_j negativo indica una relación inversa.
- La magnitud de β_j indica la fuerza de la relación.

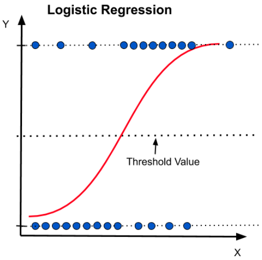
UAM Universidad Autónoma de Madrid

- Modelo:

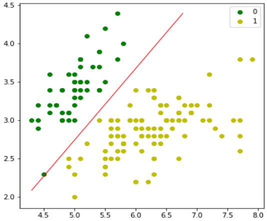
$$\mathbb{P}(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)\right)}$$

- Notación vectorial:

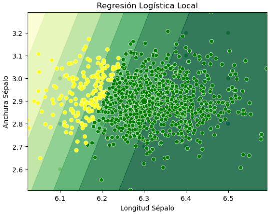
$$\mathbb{P}(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}\right)}$$



Función logística



Ejemplo de regresión logística



Ejemplo probabilidades

Interpretabilidad

La regresión logística es más compleja de interpretar que la regresión lineal:

- Los coeficientes β_j afectan las probabilidades de manera exponencial.

Interpretabilidad

La regresión logística es más compleja de interpretar que la regresión lineal:

- Los coeficientes β_j afectan las probabilidades de manera exponencial.
- Utilizando el término *odds*:

$$\text{odds} = \frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \exp\left(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n\right)$$

Interpretabilidad

La regresión logística es más compleja de interpretar que la regresión lineal:

- Los coeficientes β_j afectan las probabilidades de manera exponencial.
- Utilizando el término *odds*:

$$\text{odds} = \frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \exp\left(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n\right)$$

- Logaritmo de *odds*:

$$\log(\text{odds}) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n$$

Interpretabilidad

La regresión logística es más compleja de interpretar que la regresión lineal:

- Los coeficientes β_j afectan las probabilidades de manera exponencial.
- Utilizando el término *odds*:

$$\text{odds} = \frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \exp\left(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n\right)$$

- Logaritmo de *odds*:

$$\log(\text{odds}) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n$$

- Incremento en una unidad de x_j:

$$\text{odds}_{x_j+1} = \exp \beta_j \text{odds}_{x_j}$$

Métodos Interpretabilidad

LIME (Local Interpretable Model-Agnostic Explanations)

- **Why Should I Trust You?: Explaining the Predictions of Any Classifier**.

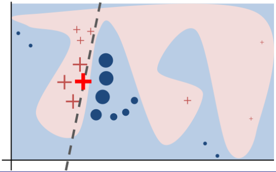
Marco Ribeiro

Sameer Singh

Carlos Guestrin

$$\min_{g} L(f, g, \pi_x) + \Omega(g) \tag{3.1}$$

- Ejemplo de LIME en Jupyter-Notebook



Lloyd Shapley

Juego coalicional:

Se tiene un conjunto N de n jugadores y una función superaditiva v que asigna subconjuntos de jugadores a números reales: $v : U = 2^N \to \mathbb{R}$, donde $v(\emptyset) = 0$. La función v se llama función característica.

Introducción a los Valores de Shapley

Lloyd Shapley

Juego coalicional:

Se tiene un conjunto N de n jugadores y una función superaditiva v que asigna subconjuntos de jugadores a números reales: $v : U = 2^N \to \mathbb{R}$, donde $v(\emptyset) = 0$. La función v se llama función característica.

Función valor:

Por la función valor $\phi[v]$ del juego v entendemos una función que asocia a cada jugador i un número real $\phi_i[v]$.

Lloyd Shapley

Axiomas de reparto justo:

- **Simetría:** $\phi_{\pi i}[\pi v] = \phi_i[v]$
- **Eficiencia:** $\sum_{i \in N} \phi_i[v] = v(N)$
- **Aditividad:** $\phi[v + w] = \phi[v] + \phi[w]$

Juego coalicional:

Se tiene un conjunto N de n jugadores y una función superaditiva v que asigna subconjuntos de jugadores a números reales: $v : U = 2^N \to \mathbb{R}$, donde $v(\emptyset) = 0$. La función v se llama función característica.

Función valor:

Por la función valor $\phi[v]$ del juego v entendemos una función que asocia a cada jugador i un número real $\phi_i[v]$.

Teorema de Shapley

Teorema de Shapley:

Para un juego coalicional dado v con un carrier finito N existe una única función valor que satisface los tres axiomas anteriores y viene dada por:

$$\phi_i[v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!}[v(S \cup \{i\}) - v(S)]$$

Los ϕ_i se denominan valores de Shapley o Shapley Values.

Interpretación

La contribución marginal de i a cada coalición, promediada sobre todas las permutaciones posibles en las que se puede formar la coalición.

Ejemplo:

- Dos concursantes (A y B) con premios 10000€, 7500€, 5000€y 0€
- A y B juntos: 10000€
- A solo: 7500€, B solo: 5000€
- Si ninguno se presenta: 0€
- Contribuciones marginales (aportaciones) de A: 7500€y 5000€
- Promedio: $\frac{7500+5000}{2} = 6250$€

SHAP (SHapley Additive exPlanations)

- **A Unified Approach to Interpreting Model Predictions**

Scott Lundberg

Su-In Lee

SHAP (SHapley Additive exPlanations)

- **A Unified Approach to Interpreting Model Predictions**

Scott Lundberg

Su-In Lee

Definición 3.12:

Denominamos *método de atribución de características aditivas* a aquellos métodos con un modelo de explicación que es una función lineal de variables binarias, definido como:

$$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

donde $\mathbf{z}'$ es un vector binario de tamaño M y ϕ_i es un valor en $\mathbb{R}$.

Existencia única bajo propiedades

1. **Precisión Local**: $\hat{f}(x) = g(x') = \phi_0 + \sum_{j=1}^{n} \phi_j x'_j$
2. **Ausencia**: $x'_j = 0 \Rightarrow \phi_j = 0$
3. **Consistencia**: $\phi_j(\hat{f}', x) \geq \phi_j(\hat{f}, x)$

Teorema 3.13:

Solo hay un modelo de explicación posible g que sigue la Definición 3.12 y satisface las Propiedades 1, 2 y 3:

$$\phi_i(f; x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f(h_x(z')) - f(h_x(z' \setminus i))] \tag{2}$$

donde $|z'|$ es el número de entradas no nulas en z', y $z' \subseteq x'$ representa todos los vectores z' donde las entradas no nulas son un subconjunto de las entradas no nulas en x'.

KernelSHAP y LIME

Teorema 3.14:

Bajo la definición de LIME como en la definición 3.12, las formas específicas de π_x, L y Ω que hacen que las soluciones de la Ecuación 3.1 sean consistentes con las Propiedades 1 a 3 son:

$$\Omega(g) = 0; \quad \pi_{x'}(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

$$L(f;g;\phi_{x_0}) = \sum_{z' \in Z} \left(f(h_{x'}^{-1}(z')) - g(z')\right)^2 \pi_{x'}(z')$$

donde $|z'|$ es el número de elementos no nulos en z'.

KernelSHAP y LIME

Teorema 3.14:

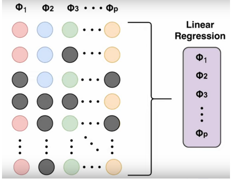
Bajo la definición de LIME como en la definición 3.12, las formas específicas de π_x, L y Ω que hacen que las soluciones de la Ecuación 3.1 sean consistentes con las Propiedades 1 a 3 son:

$$\Omega(g) = 0; \quad \pi_{x'}(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

$$L(f;g;\phi_{x_0}) = \sum_{z' \in Z} \left(f(h_{x'}^{-1}(z')) - g(z')\right)^2 \pi_{x'}(z')$$

donde $|z'|$ es el número de elementos no nulos en z'.

Estimación con KernelSHAP



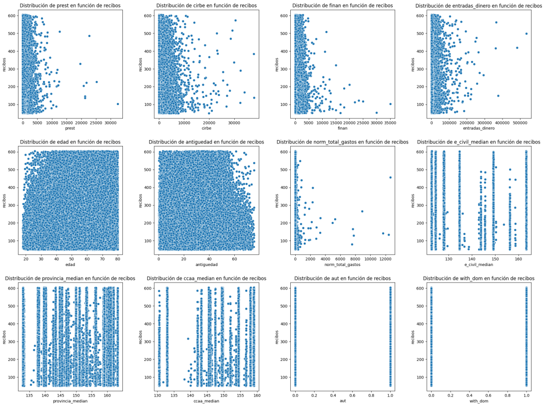
Implementación Práctica

Implementación

- Ejemplo de aplicabilidad de los conceptos estudiados.
- Modelo de clasificación y técnicas de interpretabilidad aplicadas a un conjunto de datos real.
- Implementación detallada en un Jupyter Notebook.

Previo a SHAP

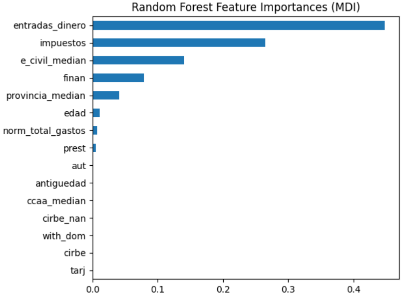
- Problema de regresión complejo.
- Dificultad para interpretar relaciones entre variables y target.
- Visualización inicial de los datos:



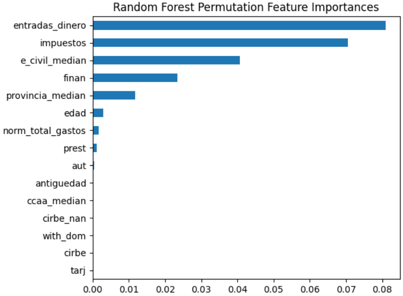
Previsualización de los datos.

Importancia de Características

- Modelos entrenados con *Feature Importance (FI)* y *Permutation Feature Importance (PFI)*.
- Identificación de variables más importantes.



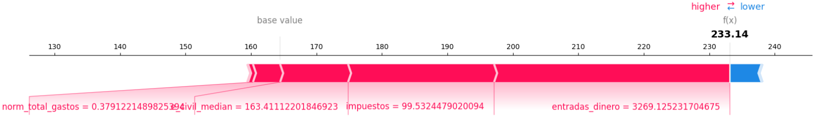
(a) Feature Importance.



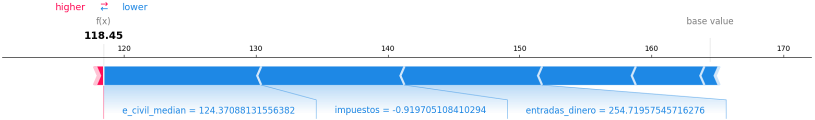
(b) Permutation Feature Importance.

Resultados importancia de características previo a SHAP.

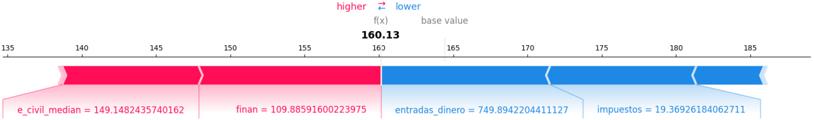
SHAP: Force Plots Locales



Force Plot para una predicción alta.

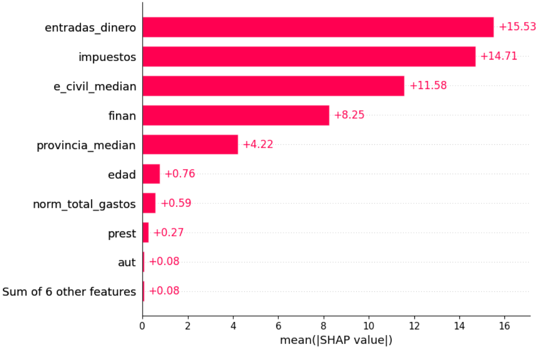


Force Plot para una predicción baja.



Force Plot para una predicción cercana a la predicción base.

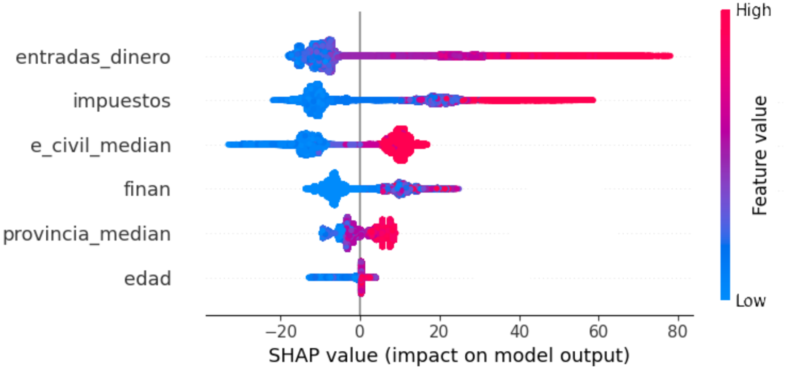
- **Bar Plot**: Importancia de las características con SHAP.



Importancia de las características con SHAP.

SAP: Explicaciones Globales

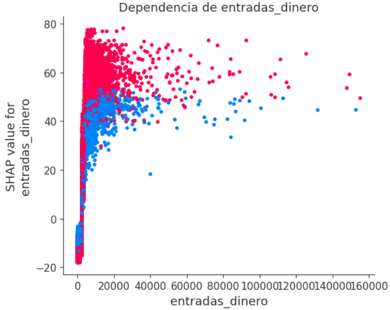
- **Summary Plot**: Distribución de los SHAP values.



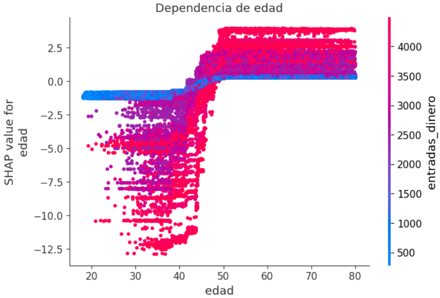
Summary Plot SHAP.

SHAP: Explicaciones Globales

- **Dependence Plots**: Distribución y relaciones entre características.



(a) Dependencia de entradas dinero.



(b) Dependencia de edad.

Dependence Plots específicos.

Conclusiones

- **Importancia de la Interpretabilidad:**
  - Precisión en modelos de clasificación.
  - Necesidad de comprender y confiar en las decisiones del modelo.
  - Contribución a una toma de decisiones más informada.

Conclusiones

- **Importancia de la Interpretabilidad:**
  - Precisión en modelos de clasificación.
  - Necesidad de comprender y confiar en las decisiones del modelo.
  - Contribución a una toma de decisiones más informada.
- **SHAP y Explicaciones de Predicciones:**
  - Locales y globales.
  - Agnósticas al modelo.

Conclusiones

- **Importancia de la Interpretabilidad:**
  - Precisión en modelos de clasificación.
  - Necesidad de comprender y confiar en las decisiones del modelo.
  - Contribución a una toma de decisiones más informada.
- **SHAP y Explicaciones de Predicciones:**
  - Locales y globales.
  - Agnósticas al modelo.
- **Resumen:**
  - SHAP como herramienta poderosa y versátil.
  - Mejora en la interpretabilidad y confianza en predicciones.
  - Crucial en campos como la economía y la salud.
  - Base matemática sólida y accesible para profesionales.

Universidad Autónoma de Madrid

Anexos

Enlaces Complementarios

- **Ejemplo de LIME en Jupyter-Notebook**:
  Notebook que implementa de mamnera muy simplificada y breve el funcionamiento de LIME.
- **Desarrollo en Jupyter-Notebook**:
  Notebook con la implementación práctica y referenciada en este TFG. Permite ejemplificar la aplicabilidad de SHAP en un caso real.

Teorema de Gauss-Markov Regresión Lineal

Estimador obtenido:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y}$$

Teorema de Gauss-Markov

Bajo las condiciones de $E[\epsilon] = \mathbf{0}$ y $\mathrm{Var}[\epsilon] = \sigma^2 \mathbf{I}$, el estimador de mínimos cuadrados es el mejor estimador lineal insesgado.
Es decir, tiene la menor varianza entre todos los estimadores lineales insesgados.

Obtención de los Pesos Regresión Logística

- Queremos maximizar la función de verosimilitud:

$$L(\beta) = \prod_{i=1}^{k} (\beta^{\intercal}\mathbf{x^{(i)}})^{y^{(i)}} * (1 - \beta^{\intercal}\mathbf{x^{(i)}})^{1-y^{(i)}}$$

- Maximizando la log-verosimilitud:

$$l(\beta) = \sum_{i=1}^{k} y^{(i)}\beta^{\intercal}\mathbf{x^{(i)}} - \log\left(1 + \exp\left(\beta^{\intercal}\mathbf{x^{(i)}}\right)\right)$$

- Método Newton-Rapshon:

$$\beta^{\mathbf{(s+1)}} = \beta^{\mathbf{(s)}} - [H_l(\beta^{(s)})]^{-1}\nabla l(\beta^{\mathbf{(s)}})$$

- Tras los cálculos:

$$\beta^{\mathbf{(s+1)}} = \beta^{\mathbf{(s)}} - [\mathbf{X}\mathbf{W}^{(s)}\mathbf{X}^{\intercal}]^{-1}\mathbf{X}(\mathbf{Y} - \hat{\mathbf{Y}}^{\mathbf{(s)}})$$

Función Característica:

$$v_x(S) = \int \hat{f}(x_1, \ldots, x_n)\, d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Shapley Values en Modelos de Clasificación

Función Característica:

$$v_x(S) = \int \hat{f}(x_1, \ldots, x_n)\, d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Valores de Shapley:

$$\phi_j(v) = \sum_{S \subseteq \{x_1, \ldots, x_n\} \setminus \{x_j\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(v(S \cup \{x_j\}) - v(S)\right)$$

Shapley Values en Modelos de Clasificación

Función Característica:

$$v_x(S) = \int \hat{f}(x_1, \dots, x_n)\, d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Valores de Shapley:

$$\phi_j(v) = \sum_{S \subseteq \{x_1, \dots, x_n\} \setminus \{x_j\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(v(S \cup \{x_j\}) - v(S)\right)$$

Propiedades:

- **Eficiencia:** $\sum_{j=1}^{n} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$
- **Simetría:** Si $v(S \cup \{x_j\}) = v(S \cup \{x_k\})$, entonces $\phi_j = \phi_k$
- **Jugador Nulo:** Si $v(S \cup \{x_j\}) = v(S)$, entonces $\phi_j = 0$
- **Aditividad:** $\phi[v + w] = \phi[v] + \phi[w]$

Futuros Trabajos

- **Ampliación de Técnicas de Interpretabilidad:**
  - Explorar otras técnicas emergentes complementarias a LIME y SHAP.
  - Integración y comparación de estas técnicas para un panorama más completo.
- **Aplicaciones en Diferentes Sectores:**
  - Investigar la aplicabilidad en marketing, seguridad informática, energía y ciencias ambientales.
- **Automatización y Herramientas de Interpretabilidad:**
  - Desarrollo de herramientas para automatizar el proceso de interpretabilidad.
  - Generación de reportes automatizados claros y comprensibles.
- **Validación y Robustez de Interpretaciones:**
  - Desarrollar métodos para validar y evaluar la calidad de las interpretaciones.
  - Crear métricas para medir fidelidad y coherencia de las explicaciones.
- **Consideraciones Éticas y de Sesgo:**
  - Identificar y mitigar sesgos en los modelos de clasificación.
  - Desarrollar metodologías para asegurar equidad y transparencia.

UAM
Universidad Autónoma de Madrid