1. Planteamiento del problema y pensamiento estadístico

Antes de aplicar cualquier modelo estadístico, es fundamental entender **qué tipo de problema se quiere resolver** y **qué se espera obtener como resultado**. La regresión logística no es solo una técnica matemática, sino una forma concreta de abordar problemas donde existe incertidumbre y toma de decisiones.

¿Qué es un problema de clasificación?

Un problema de clasificación es aquel en el que el objetivo es **asignar una observación a una categoría concreta**, a partir de cierta información disponible.

Algunos ejemplos habituales son:

- Determinar si una persona tiene o no una enfermedad.
- Predecir si un cliente va a comprar un producto.
- Decidir si un correo electrónico es spam.
- Evaluar si un estudiante abandonará sus estudios.

En todos estos casos, el resultado no es un número continuo, sino una **etiqueta o clase**. Por eso se habla de clasificación y no de predicción numérica.

Predicción de valores vs predicción de probabilidades

No todos los problemas predictivos son iguales. Es importante distinguir entre dos enfoques distintos.

Predicción de valores
Consiste en estimar un valor concreto, como un precio, una temperatura o una nota. El resultado es un número único y exacto. Este tipo de problemas suele resolverse con modelos de regresión lineal.

Predicción de probabilidades
En muchos contextos reales no es posible afirmar algo con total seguridad. En lugar de decir "sí" o "no", lo más razonable es estimar **la probabilidad de que ocurra un evento**. Por ejemplo, decir que un cliente tiene un 80 % de probabilidad de comprar.

Este enfoque reconoce que existe incertidumbre y permite tomar decisiones más informadas en función del riesgo.

La variable objetivo (target)

La variable objetivo es aquello que se quiere predecir.

- Una variable **continua** puede tomar infinitos valores dentro de un rango (por ejemplo, el precio de una casa o la altura de una persona).
- Una variable **categórica** solo puede tomar un número limitado de valores (por ejemplo, sí/no, aprobado/suspenso).

En los problemas de clasificación, la variable objetivo siempre es **categórica**.

Además, puede ser:

- **Binaria**, si solo hay dos categorías.
- **Multiclase**, si existen más de dos.

Esta distinción es importante porque determina qué tipo de modelo será adecuado más adelante.

La incertidumbre en la predicción

En la mayoría de problemas reales no se dispone de toda la información necesaria y los datos suelen contener ruido o variabilidad. Dos personas con características muy similares pueden comportarse de forma distinta. Por eso, cualquier predicción tiene un grado de incertidumbre.

En este contexto, resulta más realista hablar de probabilidades que de certezas. La probabilidad no afirma que algo ocurrirá, sino **qué tan probable es que ocurra**, lo cual es mucho más útil para tomar decisiones responsables.

¿Qué significa modelar una probabilidad?

Modelar una probabilidad implica construir un modelo que, a partir de ciertas variables explicativas, estime la probabilidad de que se produzca un determinado evento.

Por ejemplo:

Probabilidad de que una persona compre un producto dadas su edad, ingresos y hábitos de consumo.

Este tipo de modelado no pretende eliminar la incertidumbre, sino **cuantificarla**. El resultado es una estimación basada en datos, no una verdad absoluta.

Importancia de este enfoque en el trabajo

Comprender estos conceptos es esencial para:

- Justificar correctamente el uso de regresión logística frente a otros modelos.
- Explicar por qué se trabaja con probabilidades y no con predicciones deterministas.
- Evaluar críticamente los resultados y reconocer las limitaciones del modelo.

Si se entiende bien este planteamiento inicial, el resto del trabajo se construye de forma coherente y razonada.

2. Regresión lineal vs regresión logística

Este tema responde a una pregunta fundamental:

¿Por qué no podemos usar regresión lineal cuando el objetivo es clasificar?
La regresión logística surge precisamente para resolver los problemas que aparecen cuando se intenta aplicar regresión lineal a variables categóricas.

2.1 Qué es la regresión lineal (recordatorio)

La regresión lineal es un modelo que busca describir la relación entre una variable dependiente continua y una o varias variables independientes mediante una ecuación del tipo:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \varepsilon$$

Características principales:

- La variable objetivo es **continua**
- El modelo predice **valores numéricos**
- La relación entre variables se asume aproximadamente lineal
- El error se modela como ruido alrededor del valor real

Ejemplo típico:

- Predecir el precio de una vivienda a partir de su tamaño y localización

2.2 Qué ocurre si intentamos usar regresión lineal para clasificar

Supongamos que queremos predecir una variable binaria:

- 1 = compra
- 0 = no compra

Si usamos regresión lineal, el modelo puede producir salidas como:

- −0.3
- 1.4
- 2.1

Esto genera varios problemas graves:

1. **Las predicciones no están acotadas**
   Una probabilidad solo tiene sentido entre 0 y 1. La regresión lineal no respeta este límite.
2. **No hay interpretación probabilística**
   Un valor como 1.3 no puede interpretarse como probabilidad.
3. **Supuestos incorrectos**
   La regresión lineal asume:
   a. Errores normalmente distribuidos
   b. Varianza constante
   Estos supuestos no se cumplen cuando la variable objetivo es binaria.
4. **Mala frontera de decisión**
   La separación entre clases no se ajusta bien a la realidad del problema.

Por estas razones, **la regresión lineal no es adecuada para problemas de clasificación**.

2.3 La idea clave detrás de la regresión logística

La regresión logística mantiene una parte de la regresión lineal, pero introduce un cambio crucial:

- En lugar de predecir directamente Y
- Predice una **probabilidad**

El modelo calcula primero una combinación lineal:

$z=\beta 0 +\beta 1\ X1 +\beta 2\ X2 +\ldots$

Y luego aplica una función especial que transforma cualquier número real en un valor entre 0 y 1.

Esa función es la **función logística** (o sigmoide), que se estudiará con más detalle en el tema siguiente.

2.4 Por qué la regresión logística sí funciona

La regresión logística soluciona los problemas de la regresión lineal porque:

- Garantiza que la salida esté entre 0 y 1
- Permite interpretar el resultado como una probabilidad
- Modela directamente la probabilidad de pertenecer a una clase
- Se adapta bien a variables objetivo categóricas

En lugar de decir:

“Este cliente comprará”

Dice:

“La probabilidad de que este cliente compre es del 72 %”

Esto encaja mucho mejor con la realidad y con la toma de decisiones.

2.5 Regresión logística como modelo de clasificación

Aunque su nombre incluye la palabra “regresión”, la regresión logística es un **modelo de clasificación**.

El término “regresión” se debe a que:

- Se estima una función matemática continua
- Se calculan coeficientes como en la regresión lineal

Pero el objetivo final no es un valor continuo, sino **clasificar observaciones a partir de probabilidades**.

2.6 Consecuencias prácticas para el trabajo

Este tema te permite justificar con claridad:

- Por qué no se puede usar regresión lineal en tu problema

- Por qué la variable objetivo obliga a usar un modelo de clasificación
- Por qué estimar probabilidades es más útil que predecir una clase directamente

Esto es especialmente importante en:

- Ejercicio 1.2 (justificación del modelo)
- Ejercicio 10 (reflexión crítica)

Resumen conceptual

- La regresión lineal predice valores continuos
- La clasificación requiere trabajar con probabilidades
- La regresión logística transforma una combinación lineal en una probabilidad
- El modelo respeta los límites y la interpretación probabilística

Si puedes explicar con tus palabras por qué la regresión lineal falla en clasificación y cómo la logística soluciona ese problema, este tema está correctamente asimilado.

3. Tipos de regresión logística

En este tema aprenderás a **elegir correctamente el tipo de regresión logística** según la naturaleza de la variable objetivo. Elegir mal el tipo de modelo no es un detalle técnico menor: implica formular mal el problema.

3.1 Análisis de la variable objetivo

Antes de pensar en el modelo, hay que mirar **cómo es la variable que queremos predecir**. Las preguntas clave son:

1. ¿Cuántas clases tiene?
2. ¿Existe algún tipo de orden natural entre esas clases?

De estas dos preguntas salen los distintos tipos de regresión logística.

3.2 Regresión logística binaria

Es la forma más sencilla y más utilizada.

Características:

- La variable objetivo tiene **dos clases**
- Normalmente se codifican como 0 y 1
- El modelo estima la probabilidad de pertenecer a una de las dos clases

Ejemplos:

- Compra / No compra
- Enfermo / Sano
- Abandona / No abandona

El modelo calcula:

$$P(Y = 1 \mid X)$$

y, a partir de esa probabilidad, se decide la clase final usando un umbral (por ejemplo, 0.5).

Cuándo usarla:

- Siempre que el problema tenga solo dos posibles resultados
- Es el caso más común en aplicaciones reales

3.3 Regresión logística multinomial

Se utiliza cuando la variable objetivo tiene **más de dos clases** y **no existe un orden natural entre ellas**.

Ejemplos:

- Tipo de producto comprado: A, B o C
- Categoría de noticia: política, deportes, economía
- Tipo de cliente: básico, premium, VIP (si no hay jerarquía clara)

En este caso:

- El modelo estima una probabilidad para cada clase
- La suma de todas las probabilidades es 1
- Se asigna la clase con mayor probabilidad

Conceptualmente, es una extensión de la regresión logística binaria a múltiples clases.

3.4 Regresión logística ordinal

Este tipo se usa cuando:

- Hay **más de dos clases**
- Existe un **orden natural** entre ellas

Ejemplos:

- Bajo, medio, alto
- Muy insatisfecho, insatisfecho, neutro, satisfecho, muy satisfecho
- Nivel de riesgo: bajo, medio, alto

Aquí las clases no son intercambiables, ya que el orden contiene información importante.

La regresión logística ordinal:

- Tiene en cuenta ese orden
- Modela probabilidades acumuladas
- Penaliza errores grandes más que errores pequeños

No es tan común como la binaria o la multinomial, pero es más adecuada cuando el orden importa.

3.5 Enfoque One-vs-Rest (OvR)

El enfoque One-vs-Rest no es un tipo de regresión logística distinto, sino una **estrategia de entrenamiento**.

Funciona así:

- Se construye un modelo binario por cada clase
- Cada modelo aprende a distinguir “esta clase” vs “todas las demás”
- La clase final es la que obtiene la mayor probabilidad

Ejemplo con 3 clases:

- Modelo 1: clase A vs no A
- Modelo 2: clase B vs no B
- Modelo 3: clase C vs no C

Este enfoque es:

- Fácil de implementar

- Muy usado en la práctica
- Una alternativa a la logística multinomial

3.6 Comparación conceptual

Tipo	Nº clases	Orden	Cuándo usar
Logística binaria	2	No	Problemas sí/no
Logística multinomial	>2	No	Clases sin orden
Logística ordinal	>2	Sí	Clases ordenadas
One-vs-Rest	>2	No	Alternativa práctica

3.7 Importancia para el trabajo

Este tema es clave para:

- Ejercicio 2.1: describir correctamente la variable objetivo
- Ejercicio 2.2: justificar la elección del modelo
- Evitar errores conceptuales como tratar un problema ordinal como multinomial

En un trabajo bien hecho, la elección del tipo de regresión logística **no se da por supuesta**, se explica.

Cierre conceptual

La elección del modelo no depende del algoritmo, sino de **la naturaleza del problema**.

Analizar bien la variable objetivo es el primer paso para construir un modelo coherente.

4. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos es la fase en la que se estudian los datos antes de construir cualquier modelo. Su objetivo principal es entender **qué información contienen los datos**, cómo se distribuyen y qué problemas pueden afectar al modelo.

4.1 Análisis univariante

El análisis univariante consiste en estudiar cada variable de forma individual.

4.1.1 Tipo de variables

Cada variable puede ser:

- Continua (por ejemplo, edad, ingresos)
- Discreta (por ejemplo, número de compras)
- Categórica (por ejemplo, tipo de cliente)

Identificar el tipo de variable es importante porque determina cómo se analiza y cómo se representará gráficamente.

4.1.2 Distribución de una variable

La **distribución** describe cómo se reparten los valores de una variable.

Distribución simétrica

Una distribución es simétrica cuando:

- Los valores se reparten de forma equilibrada alrededor de un punto central
- La media y la mediana suelen ser similares
- La parte izquierda y derecha del gráfico tienen una forma parecida

Ejemplo típico:

- Alturas de una población
- Errores aleatorios

En una distribución simétrica, los valores extremos aparecen de forma equilibrada a ambos lados.

Distribución asimétrica

Una distribución es asimétrica cuando:

- Los valores se concentran más en un lado
- Existe una "cola" más larga hacia la izquierda o hacia la derecha

Tipos comunes:

- Asimetría positiva (cola hacia la derecha): muchos valores pequeños y pocos muy grandes

Ejemplo: ingresos

- Asimetría negativa (cola hacia la izquierda): muchos valores grandes y pocos muy pequeños

La asimetría indica que:

- La media y la mediana no coinciden
- Puede haber valores extremos que influyan mucho en el modelo

4.1.3 Valores extremos (outliers)

Los outliers son observaciones muy alejadas del resto. No siempre son errores, pero pueden:

- Distorsionar la escala de la variable
- Influir de forma excesiva en el ajuste del modelo

Por eso, en el EDA se identifican y se analizan, sin decidir todavía qué hacer con ellos.

4.1.4 Balanceo de la variable objetivo

En clasificación, es fundamental estudiar la distribución de la variable objetivo.

- **Clases balanceadas**:

Las clases tienen un número similar de observaciones.

- **Clases desbalanceadas**:

Una clase es mucho más frecuente que la otra.

Ejemplo:

- 95 % no compra, 5 % compra → desbalance fuerte

Consecuencias del desbalance

Cuando las clases están desbalanceadas:

- El modelo puede aprender a predecir siempre la clase mayoritaria
- Métricas como el accuracy pueden ser engañosas
- Los errores en la clase minoritaria pueden pasar desapercibidos

Por eso, identificar el desbalance es clave para:

- Elegir métricas adecuadas
- Ajustar el umbral de decisión
- Interpretar correctamente los resultados

4.2 Análisis bivariante

El análisis bivariante estudia la relación entre dos variables.

4.2.1 Relación entre variables explicativas y la variable objetivo

Aquí se observa cómo cambia la distribución de una variable explicativa según la clase.

Ejemplo:

- Comparar edades entre quienes compran y quienes no
- Ver si ciertos rangos se asocian más a una clase que a otra

Este análisis ayuda a detectar:

- Variables potencialmente predictoras
- Variables con poca relación con el objetivo

4.2.2 Relación entre variables explicativas (multicolinealidad)

La multicolinealidad ocurre cuando:

- Dos o más variables explicativas están fuertemente relacionadas entre sí

Esto no suele afectar mucho a la capacidad predictiva, pero sí:

- A la estabilidad de los coeficientes
- A la interpretación del modelo

Por eso es importante detectarla y tenerla en cuenta.

Importancia del EDA en regresión logística

El análisis exploratorio permite anticipar problemas como:

- Escalas muy distintas entre variables
- Influencia de valores extremos
- Desbalance de clases
- Relaciones redundantes entre variables

Todo esto influye directamente en el comportamiento del modelo.

Cierre conceptual

El EDA no consiste solo en hacer gráficos, sino en **interpretar lo que los datos cuentan**. Entender cómo se distribuyen las variables y cómo se reparten las clases es esencial para construir un modelo coherente y para justificar todas las decisiones posteriores.

5. La regresión logística desde la teoría

En este punto dejamos de lado la exploración de los datos y comenzamos a entender **cómo funciona realmente la regresión logística** y cómo se interpreta. Es la base conceptual que te permitirá conectar los datos con el modelo y sus resultados.

5.1 La función logística

La regresión logística se basa en la **función logística** (también llamada sigmoide):

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Donde (z) es una combinación lineal de las variables explicativas:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Qué hace la función logística

- Convierte cualquier número real (z) en un valor entre 0 y 1.
- Este valor se interpreta como **la probabilidad de que ocurra un evento** (por ejemplo, que un cliente compre o que un paciente tenga una enfermedad).
- A diferencia de la regresión lineal, no produce valores fuera del rango [0,1].

Concepto de modelar una probabilidad

Cuando decimos que la regresión logística **modela una probabilidad**:

- El objetivo no es predecir directamente la clase (0 o 1), sino la **probabilidad de pertenecer a la clase 1**.
- Una vez estimada la probabilidad, se puede aplicar un **umbral de decisión** (por defecto 0.5) para clasificar la observación.

5.2 Interpretación de los coeficientes

Los coeficientes $\beta_i$en la regresión logística no se interpretan igual que en regresión lineal.

- **Signo**:
  - Positivo → aumenta la probabilidad de la clase 1 al aumentar la variable
  - Negativo → disminuye la probabilidad
- **Magnitud**:
  - Indica cuánto cambia la **log-odds** por unidad de cambio en la variable
- **Log-odds**:
  - El modelo en realidad predice el **logaritmo de las odds**, no directamente la probabilidad.

5.3 Odds y Odds Ratio

Qué son las odds

$$\text{odds} = \frac{p}{1-p}$$

- Representa la razón entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra.
- Ejemplo: si (p=0.8), las odds = 0.8 / 0.2 = 4 → "4 veces más probable que ocurra a que no ocurra".

Qué es un odds ratio

- Es la razón entre las odds para diferentes niveles de una variable.
- Permite interpretar la **influencia relativa de cada variable** en la probabilidad del evento.

Ejemplo práctico:

- Un odds ratio de 2 para la variable "fuma" significa que fumar **duplica las probabilidades** de que ocurra el evento (por ejemplo, enfermedad).

5.4 Diferencia con la regresión lineal

- La regresión lineal predice un **valor numérico continuo**; la logística predice **una probabilidad**.
- La regresión logística transforma la combinación lineal de variables mediante la función sigmoide para mantener los valores entre 0 y 1.
- Los coeficientes en regresión lineal se interpretan directamente en unidades de la variable dependiente; en logística se interpretan en **odds/log-odds**, lo que requiere transformación para entender el efecto sobre la probabilidad

6. Preparación de los datos para regresión logística

Antes de entrenar un modelo, es fundamental **asegurarse de que los datos estén limpios y bien estructurados**. La regresión logística es sensible a ciertos problemas, así que esta preparación influye directamente en la calidad del modelo.

6.1 Tratamiento de valores faltantes

Qué son:

- Datos que están ausentes en alguna variable de alguna observación.

Por qué importan:

- Valores faltantes pueden sesgar los resultados o impedir que el modelo se entrene correctamente.

Opciones para tratarlos:

1. **Eliminación de filas o columnas**
    a. Si hay pocas observaciones con valores faltantes, se pueden eliminar sin afectar el análisis.
2. **Imputación**
    a. Numéricas: media, mediana, o técnicas más sofisticadas (KNN, regresión).
    b. Categóricas: moda o categoría especial "desconocido".

Decisión:

- Depende de la cantidad de datos faltantes y de la importancia de la variable.

6.2 Outliers

Qué son:

- Observaciones con valores **extremos o inusuales** comparados con la mayoría de los datos.

Por qué importan:

- Pueden **distorsionar el modelo**: las estimaciones de los coeficientes pueden ser influenciadas fuertemente por estos valores.

Cómo tratarlos:

1. **Eliminarlos** si claramente son errores o no representan la población.
2. **Transformarlos** usando escalas logarítmicas u otras.
3. **Conservarlos** si son datos reales y significativos; la regresión logística puede tolerar algunos outliers mejor que la lineal.

6.3 Escalado y codificación de variables

Escalado de variables:

- Variables con rangos muy distintos pueden afectar la convergencia del modelo.
- Escalado estándar (media=0, desviación=1) o Min-Max (0-1) puede ser útil, sobre todo si se usan métodos de regularización.

Codificación de variables categóricas:

- La regresión logística requiere **variables numéricas**.
- Técnicas comunes:
    - **One-hot encoding** → crea columnas binarias para cada categoría.
    - **Label encoding** → asigna números a las categorías (solo si hay orden).

Decisión:

- Para variables nominales (sin orden), usar one-hot.
- Para variables ordinales, usar label encoding o asignar números respetando el orden.

6.4 Por qué es importante

- Preparar los datos correctamente garantiza que el modelo **aprenda patrones reales y no artefactos de los datos**.
- Facilita la **interpretación de los coeficientes** y la confiabilidad de las predicciones.

7. Entrenamiento del modelo de regresión logística

Entrenar un modelo de regresión logística significa **determinar los coeficientes de las variables explicativas** para que la función logística prediga correctamente la probabilidad de que ocurra la variable objetivo. Es decir, buscamos los valores de los parámetros que hacen que las predicciones del modelo se acerquen lo más posible a la realidad observada.

7.1 Ajuste del modelo

Qué implica entrenar el modelo:

1. **Seleccionar variables explicativas**
    a. Podemos usar todas las variables disponibles o solo un subconjunto relevante.
    b. Las variables irrelevantes o muy correlacionadas entre sí (**multicolinealidad**) pueden afectar la interpretación de los coeficientes y el rendimiento del modelo.
2. **Elegir el método de entrenamiento**
   La regresión logística no se ajusta usando mínimos cuadrados ordinarios como en la regresión lineal, porque la variable objetivo es categórica. Los métodos más usados son:
    a. **Máxima Verosimilitud (Maximum Likelihood Estimation)**
       Busca los coeficientes que **maximizan la probabilidad de observar los datos reales** dado el modelo. Es el método estadístico estándar en regresión logística. Siempre se usa de forma implícita y se combina con un algoritmo (puntos b y c). Aquí es donde se unen el modelo estadístico (regresión logística), el criterio estadístico que define qué coeficientes son mejores (Máxima Verosimilitud) y el algoritmo de entrenamiento que permite encontrarlos (Descenso de Gradiente, Newton-Raphson, IRLS, etc)
    b. **Descenso de Gradiente (Gradient Descent)**
       Algoritmo iterativo que ajusta los coeficientes paso a paso. Calcula la pendiente de la función de verosimilitud y actualiza los coeficientes en la dirección que reduce el error.
        i. Requiere definir una **tasa de aprendizaje**.
        ii. Útil para datasets grandes.
    c. **Métodos Iterativos como Newton-Raphson o IRLS (Iteratively Reweighted Least Squares)**
       Métodos más sofisticados que usan la curvatura de la función de verosimilitud para **converger más rápido** a los coeficientes óptimos.
        i. Eficientes para datasets pequeños o medianos.
        ii. Se basan en derivadas de segundo orden.
3. **Entrenamiento paso a paso**
   El algoritmo funciona así:
    a. Se eligen valores iniciales para los coeficientes.

b. Se calcula la probabilidad predicha de cada observación usando la función logística.
c. Se mide qué tan lejos están las predicciones de los datos reales.
d. Se ajustan los coeficientes según el método elegido.
e. Se repite hasta que los coeficientes **convergen** a los valores óptimos.

7.2 Interpretación inicial de los coeficientes

Una vez entrenado el modelo, cada variable explicativa tiene un **coeficiente**. Su interpretación es diferente a la regresión lineal:

1. **Signo del coeficiente**
    a. Positivo: a medida que la variable aumenta, aumenta la probabilidad de la clase de interés (por ejemplo, "1").
    b. Negativo: a medida que la variable aumenta, disminuye la probabilidad de la clase de interés.
2. **Magnitud del coeficiente**
    a. Indica la fuerza de la relación entre la variable y la probabilidad de la clase.
    b. No se interpreta directamente como un cambio en la variable objetivo.
    c. Para interpretaciones más prácticas, se calcula el **odds ratio**:

$$odds\ ratio = e\beta$$

Este valor indica cómo se multiplican las probabilidades por cada unidad que aumenta la variable.

3. **Variables más influyentes**
    a. Aquellas con coeficientes grandes en magnitud tienen mayor impacto en la predicción.
    b. Es importante considerar la escala de la variable para evaluar correctamente su efecto.

7.3 Notas prácticas

- Revisar la **multicolinealidad**, ya que puede inflar coeficientes y confundir la interpretación.
- Si el modelo no converge:
    - Escalar las variables numéricas.
    - Revisar outliers y valores faltantes.
- El entrenamiento es solo el primer paso; después hay que **evaluar el modelo** y ajustar el **umbral de decisión** para interpretar correctamente las probabilidades.

8. Probabilidades y umbral de decisión

En regresión logística, el resultado principal del modelo **no es una clase**, sino una **probabilidad**. El modelo responde a la pregunta:
"Dadas estas características, ¿qué probabilidad hay de que el caso pertenezca a la clase de interés?"

Por ejemplo:
"La probabilidad de que este cliente haga default es 0.72"
"La probabilidad de que este paciente tenga la enfermedad es 0.18"

Esto es una ventaja importante frente a otros modelos de clasificación, porque permite **razonar con incertidumbre**. Sin embargo, en la práctica casi siempre necesitamos tomar una decisión concreta. Para eso se introduce el concepto de **umbral de decisión**.

Qué es un umbral de decisión

El umbral de decisión es un valor entre 0 y 1 que se utiliza para convertir probabilidades en clases.

La regla general es:

- Si la probabilidad predicha ≥ umbral → se asigna la clase positiva
- Si la probabilidad predicha < umbral → se asigna la clase negativa

Formalmente:

$y = 1 \text{ si } P(Y = 1 \mid X) \geq \tau, \quad y = 0 \text{ si } P(Y = 1 \mid X) < \tau$

donde τ es el umbral.

El umbral 0.5: por qué es el valor por defecto

El umbral más usado es 0.5 porque:

- Divide el espacio de probabilidad en dos partes iguales

- Funciona bien cuando:
    - Las clases están razonablemente balanceadas
    - El coste de equivocarse en una clase u otra es similar
    - El objetivo es maximizar la exactitud global

Conceptualmente, usar 0.5 significa:
"Clasifico como positiva solo cuando el modelo está más seguro de que sí que de que no".

Cuándo el umbral 0.5 no es adecuado

El umbral 0.5 **no es una ley**, es solo una convención. Hay muchos casos en los que no es la mejor opción.

Algunos ejemplos:

- Clases muy desbalanceadas (por ejemplo, 95 % negativos, 5 % positivos)
- Problemas donde un tipo de error es mucho más grave que el otro
- Contextos de riesgo, salud, fraude o seguridad

En estos casos, usar 0.5 puede llevar a decisiones claramente malas, aunque el modelo esté bien entrenado.

Falsos positivos y falsos negativos

Para entender cómo elegir un umbral, hay que comprender los dos tipos de error principales:

- **Falso positivo (FP)**
  El modelo predice clase positiva, pero la realidad es negativa.
  Ejemplo: marcar como fraude una transacción legítima.
- **Falso negativo (FN)**
  El modelo predice clase negativa, pero la realidad es positiva.
  Ejemplo: no detectar una enfermedad que sí está presente.

Estos errores no tienen el mismo coste en todos los problemas.

Coste del error y elección del umbral

La elección del umbral debe basarse en **qué error es más costoso** en el contexto real del problema.

Ejemplos:

- En medicina:
  - Un falso negativo puede ser muy grave
  - Se suele bajar el umbral para detectar más casos, aunque aumenten los falsos positivos
- En spam:
  - Un falso positivo puede ser molesto
  - Se suele subir el umbral para evitar bloquear correos legítimos
- En concesión de créditos:
  - Un falso negativo implica pérdidas económicas
  - El umbral se ajusta según la tolerancia al riesgo de la empresa

Bajar el umbral:

- Aumenta los verdaderos positivos
- Aumenta también los falsos positivos

Subir el umbral:

- Reduce los falsos positivos
- Aumenta los falsos negativos

Esto se conoce como un **trade-off**, no existe una solución perfecta.

Sensibilidad y especificidad (idea conceptual)

Cambiar el umbral modifica dos propiedades importantes del modelo:

- **Sensibilidad** **(recall)**
  Capacidad de detectar correctamente los positivos
  Aumenta cuando el umbral baja
- **Especificidad**
  Capacidad de identificar correctamente los negativos
  Aumenta cuando el umbral sube

Elegir un umbral es, en el fondo, decidir qué priorizar.

Umbral como decisión, no como propiedad del modelo

Un punto clave:
El umbral **no forma parte del modelo**, sino de la decisión que tomamos con él.

El modelo aprende probabilidades a partir de los datos.
El umbral refleja criterios externos:

- Costes
- Riesgo
- Contexto
- Objetivos del negocio o del estudio

Por eso, dos personas pueden usar el mismo modelo con **umbrales distintos** y ambas estar actuando correctamente según su contexto.

Idea final para pensamiento crítico

Un buen modelo de regresión logística no es el que "acierta más con umbral 0.5", sino el que:

- Produce probabilidades coherentes
- Permite ajustar decisiones según el contexto
- Hace explícita la incertidumbre

Entender el umbral de decisión es entender que **modelar no es decidir**, sino ayudar a decidir mejor.

9. Evaluación de modelos de clasificación

Una vez entrenado el modelo y elegido un umbral de decisión, el siguiente paso es evaluar si el modelo **realmente cumple su objetivo**. Evaluar no es solo "ver cuántos aciertos tiene", sino entender **cómo se equivoca**, **en qué casos** y **si esos errores son aceptables** para el problema concreto.

En modelos de clasificación, la evaluación siempre debe interpretarse en relación con el **contexto del problema**, no como un número aislado.

La matriz de confusión: la base de todo

La herramienta fundamental para evaluar un modelo de clasificación es la **matriz de confusión**. Resume, en forma de tabla, cómo se comparan las predicciones del modelo con la realidad.

En un problema binario:

- **Verdaderos positivos (TP)**
  Casos positivos correctamente clasificados como positivos.
- **Verdaderos negativos (TN)**
  Casos negativos correctamente clasificados como negativos.
- **Falsos positivos (FP)**
  Casos negativos clasificados erróneamente como positivos.
- **Falsos negativos (FN)**
  Casos positivos clasificados erróneamente como negativos.

La matriz de confusión permite ver no solo cuántos errores hay, sino **qué tipo de errores** comete el modelo.

Accuracy (exactitud)

La accuracy mide la proporción total de predicciones correctas:

$$Accuaracy = \frac{(TP + TN)}{TP + TN + FP + FN}$$

Es una métrica intuitiva, pero puede ser engañosa.

Ejemplo clásico:

- Dataset con 95 % negativos
- Un modelo que siempre predice "negativo" tiene 95 % de accuracy
- Pero no sirve para detectar positivos

Por eso, la accuracy **no debe usarse sola**, especialmente con clases desbalanceadas.

Precision

La precision mide qué proporción de las predicciones positivas fueron realmente correctas:

$$\text{Pr}\, e\, cision = TP/(TP + FP)$$

Responde a la pregunta:
"Cuando el modelo dice que es positivo, ¿con qué frecuencia acierta?"

Es importante cuando:

- Los falsos positivos son costosos
- Queremos confianza en las predicciones positivas

Ejemplo: detección de fraude, spam, alertas automáticas.

Recall (sensibilidad)

El recall mide qué proporción de los positivos reales fueron detectados:

$$Recall = \frac{TP}{FN + TP}$$

Responde a la pregunta:
"De todos los positivos reales, ¿cuántos detecta el modelo?"

Es importante cuando:

- Los falsos negativos son muy costosos
- No queremos "perder" casos importantes

Ejemplo: diagnóstico médico, detección de fallos críticos.

Precision vs Recall: el compromiso

Precision y recall suelen estar en tensión:

- Subir el recall suele bajar la precision
- Subir la precision suele bajar el recall

Este equilibrio depende directamente del **umbral de decisión** y del contexto del problema. No existe un valor "correcto" universal.

F1-score

El F1-score combina precision y recall en una sola métrica:

$$F1 = 2 \cdot \left(\frac{(precision \cdot Recall)}{precision + Recall} \right)$$

Se usa cuando:

- Importan tanto los falsos positivos como los falsos negativos
- Se busca un equilibrio entre ambos

Es especialmente útil en datasets desbalanceados.

Curva ROC

La curva ROC analiza el comportamiento del modelo **para todos los umbrales posibles**, no solo uno fijo.

Representa:

- Eje X: tasa de falsos positivos (FPR)
- Eje Y: tasa de verdaderos positivos (TPR o recall)

Cada punto de la curva corresponde a un umbral distinto.

La ROC permite responder:
"¿Qué tan bien separa el modelo las clases, independientemente del umbral?"

AUC (Area Under the Curve)

El AUC es el área bajo la curva ROC.

- AUC = 0.5 → modelo sin capacidad predictiva (equivalente a azar)
- AUC = 1.0 → separación perfecta entre clases

Interpretación intuitiva:
"El AUC es la probabilidad de que el modelo asigne mayor probabilidad a un positivo real que a un negativo real."

Es una métrica robusta para comparar modelos.

Cuándo usar cada métrica

No existe una métrica universalmente mejor. La elección depende del problema:

- Accuracy: solo si las clases están balanceadas y los errores cuestan lo mismo
- Precision: cuando los falsos positivos son críticos
- Recall: cuando los falsos negativos son críticos
- F1-score: cuando se necesita equilibrio
- ROC/AUC: para comparar modelos y analizar capacidad discriminativa

Evaluar es decidir si el modelo es útil

Evaluar un modelo no significa demostrar que "funciona", sino responder honestamente:

- ¿Dónde falla?
- ¿En qué casos no es fiable?
- ¿Es aceptable ese nivel de error para este problema?

Un modelo puede tener buenas métricas y aun así **no ser adecuado** para su uso real. La evaluación es el puente entre el modelo matemático y la toma de decisiones responsable.

10. Visualización y diagnóstico

Una vez entrenado y evaluado el modelo con métricas numéricas, el siguiente paso es **mirarlo**. Las visualizaciones permiten entender qué está haciendo el modelo, detectar problemas que no se ven en los números y explicar los resultados a otras personas. En muchos casos, un gráfico bien elegido es más informativo que varias métricas.

La visualización no es solo comunicación, también es una herramienta de diagnóstico.

Distribución de las probabilidades predichas

Un primer gráfico fundamental es la distribución de las probabilidades que produce el modelo.

Normalmente se representan:

- Las probabilidades predichas para la clase positiva
- Separadas por clase real (positivos reales y negativos reales)

Este gráfico permite ver:

- Si el modelo asigna probabilidades altas a los positivos y bajas a los negativos
- Si hay mucha superposición entre ambas distribuciones

Un buen modelo tiende a producir:

- Probabilidades cercanas a 1 para positivos reales
- Probabilidades cercanas a 0 para negativos reales

Si ambas distribuciones se mezclan mucho, el modelo tiene poca capacidad de discriminación, aunque la accuracy pueda parecer aceptable.

Separación entre clases

Cuando el número de variables es pequeño (una o dos), se pueden hacer gráficos que muestren la **frontera de decisión** del modelo.

Estos gráficos ayudan a entender:

- Cómo separa el modelo el espacio de datos
- Qué regiones del espacio se consideran positivas o negativas
- Dónde se concentran los errores

Incluso en dimensiones altas, proyecciones parciales o gráficos por pares pueden revelar:

- Variables con poco poder discriminativo
- Regiones problemáticas del espacio de datos

Curva ROC como herramienta visual

La curva ROC no solo es una métrica, también es una visualización diagnóstica.

Permite observar:

- Cómo cambia el rendimiento al variar el umbral
- Si pequeñas mejoras en recall implican grandes costes en falsos positivos
- Qué zonas de la curva son operativamente aceptables

Comparar curvas ROC de distintos modelos suele ser más informativo que comparar una sola métrica puntual.

Visualización de coeficientes e importancia de variables

En regresión logística, los coeficientes son una fuente directa de información.

Visualizarlos permite:

- Identificar qué variables empujan hacia la clase positiva o negativa
- Comparar magnitudes relativas
- Detectar coeficientes exageradamente grandes, posibles síntomas de multicolinealidad

Es importante recordar que:

- La magnitud depende de la escala de la variable
- Por eso es preferible visualizar coeficientes tras el escalado

Detección de problemas comunes mediante gráficos

Las visualizaciones ayudan a detectar problemas como:

- **Sobreajuste**
  Muy buena separación en entrenamiento, mala en validación
- **Infraajuste**
  Distribuciones de probabilidades muy concentradas alrededor de 0.5
- **Desbalance de clases**
  Probabilidades sistemáticamente bajas para la clase minoritaria
- **Variables mal codificadas o mal escaladas**
  Coeficientes extremos o comportamiento errático

Visualización como herramienta de explicación

Un aspecto clave del trabajo es poder **explicar el modelo**.

Gráficos útiles para esto:

- Histogramas de probabilidades predichas
- Curva ROC con puntos de umbral marcados
- Barras de coeficientes ordenadas por magnitud

Estos gráficos permiten justificar decisiones como:

- Por qué se eligió un determinado umbral
- Por qué ciertas variables se consideran importantes
- Por qué el modelo no es perfecto pero sí útil

Idea central

Las métricas resumen el rendimiento, pero las visualizaciones muestran el comportamiento.

Un buen análisis no se limita a decir "el AUC es alto", sino que muestra:

- Cómo se distribuyen las probabilidades
- Dónde se equivoca el modelo
- Qué patrones está captando

11. Pensamiento crítico y toma de decisiones

¿El modelo responde al objetivo inicial?

La primera pregunta que hay que hacerse es si el modelo cumple el propósito para el que fue creado.

Algunas cuestiones clave:

- ¿Predice lo que realmente queríamos predecir?
- ¿La variable objetivo está bien definida?
- ¿Las probabilidades generadas tienen sentido en el contexto real?

Un modelo puede tener buenas métricas y aun así no responder al problema correcto. Por ejemplo, predecir "riesgo de abandono" sin que esa predicción se traduzca en una acción concreta puede ser inútil en la práctica.

Limitaciones del modelo

La regresión logística tiene limitaciones estructurales que hay que reconocer:

- Asume una relación aproximadamente lineal entre las variables y el log-odds
- No captura bien interacciones complejas si no se modelan explícitamente
- Es sensible a outliers y a multicolinealidad
- Su rendimiento depende mucho de la calidad del dataset

Reconocer estas limitaciones no debilita el análisis, lo fortalece.

Sesgos en el dataset

Un modelo aprende únicamente de los datos que se le proporcionan. Si los datos están sesgados, el modelo también lo estará.

Ejemplos comunes:

- Muestras no representativas
- Variables que reflejan desigualdades estructurales
- Etiquetas ruidosas o mal definidas

Es fundamental preguntarse:

- ¿De dónde vienen los datos?
- ¿A quién representan y a quién no?
- ¿Qué casos están infrarrepresentados?

Un modelo aparentemente "preciso" puede estar reforzando sesgos existentes.

Sobreajuste e infraajuste

Dos problemas clásicos del aprendizaje automático:

- **Sobreajuste**
  El modelo se adapta demasiado a los datos de entrenamiento y pierde capacidad de generalización.
- **Infraajuste**
  El modelo es demasiado simple y no capta los patrones relevantes.

La regresión logística suele ser menos propensa al sobreajuste que modelos más complejos, pero no está exenta, especialmente con muchas variables o pocas observaciones.

Uso responsable del modelo

Antes de confiar en un modelo para decisiones reales, hay que reflexionar sobre su impacto.

Preguntas necesarias:

- ¿Qué consecuencias tiene un error del modelo?
- ¿Quién asume el riesgo?
- ¿Existe supervisión humana?

En contextos sensibles, el modelo debe servir como **apoyo a la decisión**, no como sustituto automático del juicio humano.

Cuándo no usar regresión logística

La regresión logística no siempre es la mejor opción.

No es adecuada cuando:

- La relación entre variables es altamente no lineal
- El problema requiere modelar interacciones complejas sin ingeniería de variables
- Se dispone de grandes volúmenes de datos con patrones complejos

En estos casos, otros modelos pueden ser más apropiados, aunque menos interpretables.

Conclusión final

Un buen trabajo con regresión logística no se mide solo por métricas altas, sino por:

- La claridad del planteamiento
- La coherencia de las decisiones
- La honestidad sobre las limitaciones
- La capacidad de interpretar y justificar resultados

La regresión logística es valiosa no solo porque predice, sino porque obliga a pensar en términos de probabilidad, incertidumbre y consecuencias.