CIMAT

Análisis de datos: Bank Marketing

Introducción a Ciencia de Datos

Autores:
María Alejandra Borrego Leal
Luz María Salazar Manjarrez

Profesor:
Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas A. C.

Septiembre 2025

1. Introducción

El presente trabajo analiza una base de datos proveniente de campañas de marketing directo realizadas por una institución bancaria portuguesa, cuyo propósito fue promocionar depósitos a plazo mediante llamadas telefónicas a clientes. El objetivo principal del estudio es desarrollar y comparar diferentes modelos de clasificación que permitan predecir si un cliente aceptará o no el producto ofrecido.

El conjunto de datos trabajado incluye variables predictoras de tipo numérica y categórica, además de la variable respuesta binaria que indica la suscripción del depósito. Un aspecto relevante en la preparación de los datos fue el manejo de valores desconocidos, especialmente en la variable **default**, donde el porcentaje de respuestas "unknown" fue considerablemente alto y se abordó mediante imputación proporcional, destacando la posibilidad de sesgo en este atributo. También, optamos por excluir la variable **duration** del modelado, al no ser realista como predictor antes de la llamada.

Para la fase de modelado se entrenaron y evaluaron distintos clasificadores: Naive Bayes, LDA, QDA, Fisher y k-NN. Se utilizaron métricas como exactitud, sensibilidad, precisión, F1-score, AUC y validación cruzada para valorar su desempeño.

Entre los resultados más importantes se destaca que: Naive Bayes obtuvo un desempeño aceptable con exactitud de 87 % y AUC de 0.775, útil como modelo base. LDA alcanzó alta exactitud (89.3 %) y buen equilibrio en precisión y F1, con AUC de 0.801. QDA mejoró la recuperación de la clase positiva y logró la mejor balanced accuracy (0.696), aunque con más falsos positivos. k-NN, pese a registrar alta exactitud (89.1 %), mostró el AUC más bajo (0.729), reflejando un sesgo hacia la clase mayoritaria. Fisher, con priores empíricas, alcanzó el mejor AUC (0.803), mostrando gran capacidad de discriminación en un escenario desbalanceado.

Todo este análisis parece indicarnos que en un problema con fuerte desbalance de clases, las métricas más informativas son aquellas que equilibran sensibilidad y especificidad (como AUC y balanced accuracy). De acuerdo con estos criterios, los clasificadores Fisher y QDA ofrecen un mejor rendimiento al identificar clientes potenciales, mientras que LDA destaca por su estabilidad y robustez global.

2. Exploración de datos

La base de datos analizada está relacionada con campañas de marketing directo realizadas en una institución bancaria portuguesa. Dichas campañas de marketing se llevarona cabo mediante llamadas telefónicas a clientes, con el propósito de promocionar un producto financiero: los depósitos a plazos. El objetivo principal de este análisis de datos es poder desarrollar un modelo de clasificación que nos permita predecir si el cliente al que se llamó suscribirá o no un depósito a plazo. Mencionamos que, a menudo, se requería más de un contacto con el mismo cliente para determinar si el producto (depósito a plazo bancario) sería suscrito ("sí") o no ("no").

El conjunto de datos recabados en esta campaña de marketing consta de 21 variables, incluida la variable respuesta (u objetivo). En esta lista de variables aparecen tanto númericas (continuas y discretas) como categóricas (dicotómicas, ordinales y nominales). Presentamos una breve descripción de estas variables:

Variable respuesta:

- **y**: tipo categórica (dicotómica). Indica si el cliente ha suscrito (o no) un depósito a plazo.

Variables predictoras:

- **age**: tipo numérica (entero). Indica la edad en años del cliente.

- **job**: tipo categórica (nominal). Indica el tipo de trabajo del cliente.
- **marital**: tipo categórica (nominal). Indica el estado civildel cliente.
- **education**: tipo categórica (ordinal). Indica el nivel de estudios del cliente.
- **default**: tipo categórica (dicotómica). Indica si el cliente se encuentra en inclupimiento de algún crédito bancario.
- **housing**: tipo categórica (dicotómica). Indica si el cliente tiene un crédito para vivienda.
- **loan**: tipo categórica (dicotómica). Indica si el cliente tiene un crédito personal.
- **contact**: tipo categórica (nominal). Indica el tipo de comunicación que se tuvo con el cliente.
- **month**: tipo categórica (nominal). Indica en qué mes se realizó la llamada.
- **day_of_week**: tipo categórica (nominal). Indica en qué día se realizó la llamada.
- **duration**: tipo numérica (entero). Indica el tiempo en segundos que duró la llamada.
- **campaing**: tipo numéritca (entero). Indica el número de contactos que se ha tenido con el cliente durante la actual camapaña.
- **pdays**: tipo numéritca (entero). Indica el número de días transcurridos desde el último contacto con el cliente en una campaña anterior.
- **previous**: tipo numértica (entero). Indica el número de contactos con el cliente antes de la actual campaña.
- **poutcome**: tipo categórica (nominal). Indica el resultado de la camapaña anterior.
- **emp.var.rate**: tipo numérica (decimal). Indica la tasa de variación del empleo, medida de forma trimestral.
- **cons.price.idx**: tipo numérica (decimal). Indica el índice de precios al consumidor.
- **cons.conf.idx**: tipo numérica (decimal). Corresponde al índice de confianza del consumidor.
- **euribor3m**: tipo numérica (decimal). Indica el tipo de interés a 3 meses del Euribor (European Interbank Offered Rate).
- **nr.employed**: tipo numérica (entero). Indica el número de empleados (en miles) como indicador del nivel de empleo.

Destacamos un punto importante en la variable **duration**: este atributo influye considerablemente en el resultado final (por ejemplo, si la duración es 0, entonces evidentemente el resultado es "no"). Sin embargo, la duración no se conoce hasta después de realizar la llamada. Además, una vez finalizada la llamada, el resultado ya es conocido. Por lo tanto, este dato solo debe incluirse para fines de comparación y debe omitirse si se desea un modelo predictivo realista. Consecuentemente, no hacemos uso de esta variable en los modelos de clasificación que trabajamos en este estudio.

En la Tabla 1 presentamos los valores que toma cada una de las variables categóricas descritas.

Variable	Valores
y	“yes”, “no”
job	“admin.”,“blue-collar”,“entrepreneur”,“housemaid”,“management”,“retired”, “self-employed”,“services”,“student”,“technician”,“unemployed”,“unknown”
marital	“divorced”, “married”, “single”, “unknown”
education	“basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”
default	“yes”, “no”
housing	“yes”, “no”
loan	“yes”, “no”
contact	“telephone”, “cellular”
month	“jan”, “feb”, “mar”, ... , “nov”, “dec”
day_of_week	“mon”, “tue”, “wed”, “thu”, “fri”
poutcome	“nonexistent”, “failure”, “success”

Tabla 1: Valores que puede tomar cada variable categórica.

En la base de datos no encontramos valores “vacíos”(Nan), lo que puede pensarse que la hace más fácil de manejar. Sin embargo, hay variables categóricas que presentan el valor “unknown”, estas son: **job**, **marital**, **education**, **default**, **housing** y **loan**. En la Tabla 2 podemos observar el porcentaje de “unknown” encontrado en cada una de estas variables

Variable	Porcentaje	Variable	Porcentaje
job	0.8 %	default	20.88 %
marital	0.19 %	housing	2.4 %
education	4.2 %	loan	2.4 %

Tabla 2: Porcentajes de valores “unknown” en cada variable.

Es importante que notemos que la mayoría de las variables presentadas en la Tabla 2 tienen un porcentaje muy bajo (algunas incluso menor al 1 %) de valores desconocidos, por lo que optamos simplemente por hacer una imputación aleatoria proporcional en ellas, considerando que esto no afectará de manera significativa el análisis de los datos. No obstante, también identificamos que en la variable **default**, el porcentaje de datos desconocidos, es considerablemente alto, y dada la naturaleza de la variable (es la respuesta a preguntarle al cliente si tiene algún incumplimiento bancario), podemos pensar que estos “datos faltantes”son del mecanismo MNAR (Missing Not At Random), pues las personas no suelen admitir estos detalles y prefieren omitir esta respuesta. Específicamente, encontramos que hay un total de 32588 “no”, 8597 “unknown” y solamente 3 “sí”, es decir, una desproporción enorme entre las respuestas. Esto hace que si intentamos realizar una imputación simple (ya sea moda, mediana, etc), una imputación proporcional aleatoria, o una imputación múltiple (que es la estrategia utilizada en el caso de sumir que los datos pertenecen al modelo MAR), estaríamos obteniendo prácticamente el mismo resultado, que es cambiar todos los “unknown” por “no”, pues la probabilidad de obtener “sí” sería prácticamente nula. Por esta razón decidimos trabajarla de igual manera que las demás variables categóricas, pero resaltanto que puede ser una opción bastante sesgada en este caso, pues de hecho es más lógico pensar que las respuestas desconocidas podrían ser más un “sí”que un “no”.

Finalmente, reemplazamos todos los valores desconocidos y realizamos la codificación one-hot a las variables categóricas (excepto la variable **education** y las binarias), para transformarlas a variables dummies y así convertirlas en un formato numérico que el algoritmo pueda usar. A las variables binarias les asiganmos 0 en los valores “no” y 1 en los valores “sí”. En el caso de la variable **education**, decidimos codificarla de manera diferente, pues al tratarse de niveles escolares, la podemos calificar como en una escala ordinal y así asignarle los valores del 0-6 correspondientemente a los niveles de

estudio, desde el menor hasta el más alto nivel.

3. Modelado y evaluación: Entrenamiento de clasificadores

A continuación presentamos los resultados de los entrenamientos y la validación de los siguientes clasificadores: Naive Bayes, LDA, QDA, Fisher y k-NN. El objetivo es comparar la capacidad de cada clasificador para predecir si un cliente realizará o no un depósito a plazo tras las llamadas de la institución bancaria. Para el entrenamiento de cada clasificador consideramos la misma muestra del 70 % de los datos seleccionados de manera aleatoria. Una vez terminado evaluamos cada uno calculando diversos estimadores de desempeño como son la exactitud, sensibilidad, precisión, F1, AUC y por supuesto presentamos la matriz de confusión.

3.1. Naive Bayes

Con el entrenamiento realizado al clasificador de Naive Bayes obtuvimos la matriz de confusión de la Figura 1 y los estimadores de desempeños de la Tabla 3

Estimador de desempeño	Valor
Exactitud	0.870
Sensibilidad	0.870
Precisión	0.872
F1-Score	0.871
ROC–AUC	0.775

Tabla 3: Estimadores de desempeño del clasificador Naive Bayes.

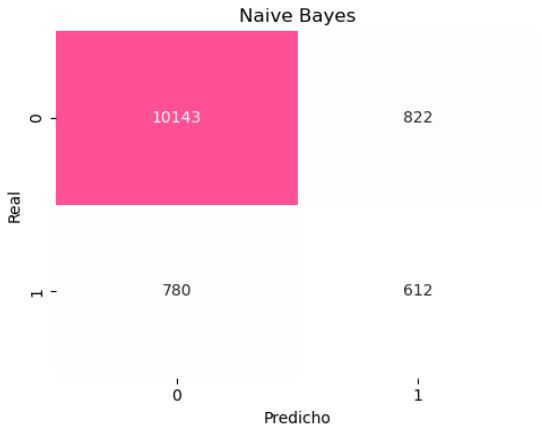

Figura 1: Matriz de Confusión del clasificador Naive Bayes.

En la matriz de confusión observamos que hay muchos verdaderos negativos, entonces el clasificador predice correctamente a la gran mayoría de clientes que no hacen depósitos pero aún confunde algunos casos al presentar 822 falsos positivos y 780 falsos negativos. Dado que la exactitud del modelo es de 0.87, entonces acierta en la clasificación del 87 % de los clientes (predice correctamente si harán o no un depósito a plazo), lo cual es un rendimiento aceptable considerando que el problema es binario y suele tener cierto desbalance de clases. Como tiene una sensibilidad de aproximadamente un 87 %, entonces el modelo identifica bien a los clientes que sí realizan depósitos pero no detecta a un 13 % de clientes interesados. Con una precisión de 0.87, podemos decir que de los clientes que el modelo predijo como que harán un depósito, el 87 % de verdad lo hará. Es decir que los falsos positivos son relativamente pocos. Del F1-Score tenemos que equilibrio entre precisión y sensibilidad es muy bueno, lo que confirma que el modelo no está sesgado hacia predecir siempre que "sí" o que "no".

Finalmente ROC-AUC = 0.775 indica una capacidad de discriminación moderada a buena: el clasificador ordena correctamente un par positivo/negativo aproximadamente el 77.5 % de las veces.

3.2. LDA

Con el entrenamiento realizado al clasificador LDA obtuvimos la matriz de confusión de la Figura 2 y los estimadores de desempeño de la Tabla 4.

Estimador de desempeño	Valor
Exactitud	0.893
Sensibilidad	0.893
Precisión	0.879
F1-Score	0.884
ROC–AUC	0.801

Tabla 4: Estimadores de desempeño del clasificador LDA.

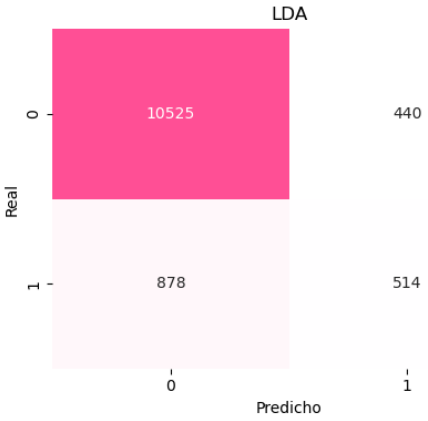

Figura 2: Matriz de Confusión del clasificador LDA.

En la matriz de confusión se observa un alto número de verdaderos negativos (10 525) y verdaderos positivos moderados (514), frente a 440 falsos positivos y 878 falsos negativos. Esto implica que LDA identifica muy bien la clase negativa (alta especificidad), alcanzando una exactitud global del 0.893. No obstante, la recuperación de la clase positiva es más limitada (sensibilidad positiva baja), algo esperable por el desbalance de clases. Los valores de precisión (0.879) y F1-Score (0.884) ponderados muestran un equilibrio razonable entre aciertos y errores, manteniendo un desempeño robusto en el conjunto de prueba.

Finalmente ROC-AUC = 0.801 indica una capacidad de discriminación buena: LDA ordena correctamente pares positivo/negativo el 80.1 % de las veces, independientemente del umbral de decisión. Este valor, mayor al de Naive Bayes (0.775), sugiere que la frontera lineal de LDA separa mejor ambas clases.

3.3. QDA

Con el entrenamiento realizado al clasificador QDA obtuvimos la matriz de confusión de la Figura 3 y los estimadores de desempeño de la Tabla 5.

Estimador de desempeño	Valor
Exactitud	0.879
Sensibilidad	0.879
Precisión	0.879
F1-Score	0.879
ROC–AUC	0.792

Tabla 5: Estimadores de desempeño del clasificador QDA.

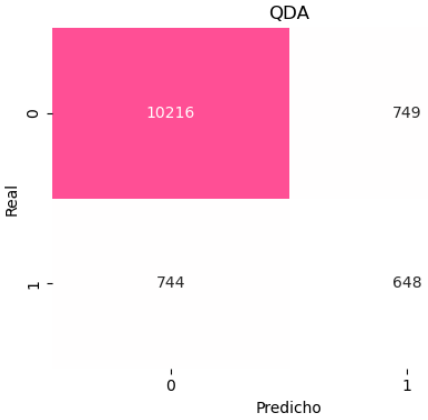

Figura 3: Matriz de Confusión del clasificador QDA.

En la matriz de confusión se observan 10, 216 verdaderos negativos y 648 verdaderos positivos, frente a 749 falsos positivos y 744 falsos negativos. El conjunto tiene una clase positiva de aproximadamente el 11.3 % y una clase negativa de aproximadamente el 88.7 %, por ello, las métricas ponderadas

por soporte tienden a estar dominadas por el desempeño en la clase negativa.

QDA capta no linealidades (fronteras cuadráticas) y, comparado con clasificadores lineales típicos, mejora la recuperación de la clase positiva a costa de aumentar los falsos positivos. Lo observado es consistente con priorizar mayor cobertura de clientes potencialmente interesados aun con más alarmas falsas. Si el objetivo operativo es no perder posibles suscriptores, este movimiento puede ser deseable; si el costo de contactar falsos positivos es alto, quizá no lo sea.

Un ROC-AUC = 0.792 indica una buena capacidad de discriminación: QDA ordena correctamente pares positivo/negativo en torno al 79.2 % de las veces, de forma independiente al umbral. El valor queda por debajo de LDA (0.801) pero por encima de Naive Bayes (0.775).

3.4. k-NN

Con el entrenamiento realizado al clasificador k-NN $(k = 5)$ obtuvimos la matriz de confusión de la Figura 4 y los estimadores de desempeño de la Tabla 6.

Estimador de desempeño	Valor
Exactitud	0.891
Sensibilidad	0.891
Precisión	0.871
F1-Score	0.876
ROC–AUC	0.729

Tabla 6: Estimadores de desempeño del clasificador k-NN.

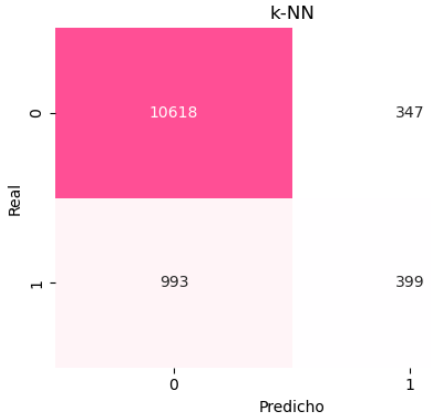

Figura 4: Matriz de Confusión del clasificador k-NN.

A partir de la matriz de confusión se observan 10 618 verdaderos negativos y 399 verdaderos positivos, frente a 347 falsos positivos y 993 falsos negativos. Estos valores muestran que k-NN es conservador para predecir la clase positiva: logra una alta especificidad (pocos falsos positivos), pero baja sensibilidad (muchos falsos negativos). La exactitud 0.891 y el F1 ponderado 0.876 lucen altos en parte por el desbalance del conjunto (la clase negativa es mayoritaria) y reflejan principalmente el buen desempeño en la clase negativa.

El ROC-AUC = 0.729 indica una discriminación moderada: el modelo ordena correctamente pares positivo/negativo en torno al 72.9 % de las veces, independiente del umbral.

3.5. Fisher (con priores empíricas)

Cómo construimos el clasificador de Fisher (binario). Dado un conjunto (X, y) con $y \in \{0, 1\}$:

1. Separar por clase y calcular las medias: $\mu_0,\ \mu_1$.
2. Estimar la dispersión *dentro de clase*: $S_W = S_0 + S_1$, donde S_k es la covarianza de la clase k.
3. Dirección discriminante (Fisher/FLD):

$$a \ = \ S_W^{-1}(\mu_1 - \mu_0).$$

4. Proyectar cada punto: $z = a^\top x$. Denotar $z_k = a^\top \mu_k$.

5. Elegir el umbral t y clasificar: $\hat{y} = 1_{\{z>t\}}$. Con *priores* (π_0, π_1) y varianza proyectada $\sigma^2 = a^\top \hat{\Sigma}\, a$:

$$t \;=\; \frac{z_0 + z_1}{2} \;+\; \frac{\sigma^2}{z_1 - z_0}\, \log \frac{\pi_0}{\pi_1}.$$

¿Por qué usamos priores empíricas? En nuestros datos, la clase positiva es minoritaria. Bajo pérdida 0–1, el clasificador de Bayes usa las priores reales de cada clase. Cuando éstas no se conocen, un estimador natural es la prior empírica ($\pi_k = n_k/n$, proporción de ejemplos de la clase k).

Con el entrenamiento realizado al clasificador de Fisher (umbral ajustado con priores empíricas) obtuvimos la matriz de confusión de la Figura 5 y los estimadores de desempeño de la Tabla 7.

Estimador de desempeño	Valor
Exactitud	0.889
Sensibilidad	0.889
Precisión	0.882
F1-Score	0.885
ROC–AUC	0.803

Tabla 7: Estimadores de desempeño del clasificador Fisher.

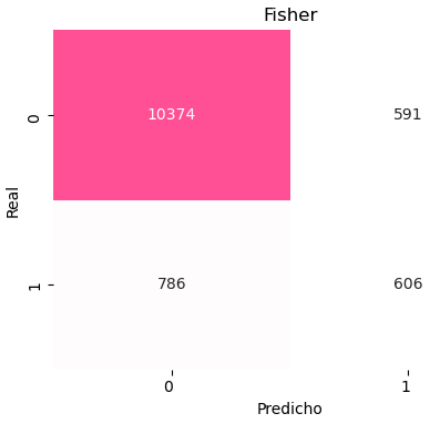

Figura 5: Matriz de Confusión del clasificador Fisher.

La matriz de confusión reporta 10 374 verdaderos negativos y 606 verdaderos positivos, frente a 591 falsos positivos y 786 falsos negativos. Esto muestra un modelo muy preciso identificando la clase negativa (alta especificidad), pero con recuperación moderada de la clase positiva, patrón coherente con el desbalance de clases. EL ROC-AUC = 0.803 indica una buena capacidad de discriminación independiente del umbral, y se sitúa ligeramente por encima de LDA (0.801).

3.6. Validación cruzada

Modelo	Acc (± sd)	BAcc	F1_w	ROC–AUC
Naive Bayes	0.869 ± 0.002	0.682	0.870	0.767
LDA	0.890 ± 0.002	0.666	0.883	0.789
QDA	0.877 ± 0.003	0.696	0.877	0.786
k-NN (k=5)	0.890 ± 0.004	0.628	0.876	0.729
Fisher	0.886 ± 0.003	0.688	0.883	0.792

Tabla 8: Validación cruzada 5-fold: promedio de métricas.

- **Exactitud.** LDA y k-NN obtienen la mayor Acc (0.890), pero la desviación estándar es mayor en k-NN (0.004), lo que sugiere un desempeño menos estable entre pliegues.
- **Balanced accuracy (BAcc).** Con clases desbalanceadas, BAcc es más informativa: QDA (0.696) y Fisher (0.688) equilibran mejor TPR/TNR que LDA (0.666) y, sobre todo, k-NN (0.628).
- **F1 ponderado.** LDA y Fisher empatan en el mejor F1_w (0.883), señal de buen rendimiento global ponderado por el soporte de cada clase.

- **ROC–AUC.** La capacidad de discriminación es mayor en Fisher (0.792), seguida muy de cerca por LDA (0.789) y QDA (0.786); Naive Bayes queda detrás (0.767) y k-NN es el más bajo (0.729).
- Si priorizas discriminación y balance entre clases en un escenario desbalanceado, Fisher (con priores empíricas) y QDA son opciones sólidas; LDA ofrece métricas globales fuertes y buena estabilidad, mientras que k-NN, pese a su alta Acc, muestra peor BAcc y AUC, por lo que es menos adecuado si importa detectar la clase minoritaria.

4. Conclusiones

En un problema desbalanceado y con mezcla de variables numéricas y muchas dummies, la exactitud resultó poco informativa por estar dominada por la clase mayoritaria. Métricas más robustas (ROC-AUC y *balanced accuracy*, BAcc) clarifican la comparación. Así, aunque LDA y k-NN logran altas Acc, Fisher y QDA son preferibles cuando importa recuperar mejor la clase positiva sin sacrificar demasiado la especificidad.

El desbalance que hay penaliza modelos que fijan umbrales "neutros". Con umbral de Fisher ajustado por priores empíricas mejoró la discriminación frente a su versión de punto medio. Otra de las cosas a considerar en cuanto a la naturaleza de los datos es su alta dimensionalidad y dummies, afectando por ejemplo a k-NN (distancias menos informativas), lo que explica su AUC más bajo pese a buena Acc.

Aportes de cada técnica

- **Naive Bayes (NB):** Rápido, sencillo y con interpretabilidad aditiva (evidencias por atributo). Útil como línea base; sin embargo, la suposición de independencia lo deja por detrás en AUC.
- **LDA:** Frontera lineal robusta, interpretable por coeficientes y estable en CV. Excelente compromiso entre Acc/F1 y AUC.
- **QDA:** Permite fronteras no lineales (covarianzas por clase). Logra la mejor BAcc (0.696) y buen AUC (0.792), a costa de más falsos positivos y mayor varianza si no se regula.
- **Fisher:** Al mover el umbral con priores empíricas se alineó mejor con el desbalance y alcanzó el mayor AUC (0.803). Aporta claridad geométrica y una vía directa para control de umbrales.
- **k-NN:** No paramétrico y simple, pero sensible a la escala y a la dimensión. Aunque su Acc es alta, su BAcc/AUC (0.628/0.729) muestran que favorece la clase mayoritaria; requiere escalado, elección cuidadosa de k y, de ser posible, reducción de dimensión.

En síntesis, en este conjunto desbalanceado y de alta dimensión, los modelos lineales bien especificados y con umbral adecuado (Fisher/LDA) ofrecen la mejor discriminación y estabilidad global; QDA aporta flexibilidad útil cuando existen no linealidades, y NB/k-NN cumplen roles de línea base e intuición, pero quedan por detrás cuando se prioriza recuperar la clase positiva.

Referencias

[Brodersen, 2010] Brodersen, K. H., O. C. S. S. K. E. . B. J. M. (2010). The balanced accuracy and its posterior distribution.

[Moro, 2014] Moro, S., R. P. . C. P. (2014). Bank marketing [dataset]. uci machine learning repository.