Comparación de Técnicas de Clasificación para la Predicción de Suscripción de Depósitos a Plazo en Campañas de Telemarketing Bancario

Rodolfo Jesús Ramirez Lucario
Omar García Ramos
Eric Ernesto Moreles Abonce

Resumen

Utilizando el dataset *Bank Marketing* del repositorio UCI, se analizan 45,211 contactos telefónicos con 16 variables predictoras para predecir la suscripción de depósitos a plazo. Este estudio implementa y compara seis clasificadores supervisados—*Naive Bayes*, LDA, QDA, Fisher, y k-NN—enfrentando el desafío del desbalance de clases (12 % suscripciones vs 88 % no suscripciones). Los resultados revelan que el Discriminante Lineal de Fisher con ajuste de umbral alcanza el mejor balance entre métricas de evaluación, demostrando la importancia de adaptar los algoritmos a distribuciones desbalanceadas. El análisis proporciona criterios prácticos para la selección de modelos según diferentes objetivos de negocio en campañas de marketing bancario.

1 Introducción

La identificación automatizada de clientes con alta propensión a contratar productos financieros constituye un problema fundamental en la banca moderna. Mediante técnicas de clasificación supervisada, es posible optimizar la asignación de recursos en campañas de contacto telefónico, focalizando los esfuerzos en aquellos individuos con mayor probabilidad de suscripción. Estudios previos como el de Moro et al. (2014) demostraron la utilidad de técnicas de *data mining* en este dominio, logrando mejorar sustancialmente la eficiencia de campañas de marketing directo mediante modelos complejos como Redes Neuronales y SVM. Sin embargo, su aproximación requirió un extenso proceso de selección de características (de 150 a 22 variables) y modelos computacionalmente demandantes.

Mientras que Moro et al. se centraron en maximizar la precisión predictiva con modelos complejos, nuestro trabajo adopta un enfoque complementario al evaluar clasificadores paramétricos y no paramétricos de menor complejidad computacional, utilizando el conjunto reducido de 16 variables disponibles en el repositorio UCI.

Un desafío persistente en este tipo de problemas es el marcado desbalance de clases, donde típicamente un pequeño porcentaje de los contactos resultan en suscripciones exitosas. Esta característica distorsiona las métricas de evaluación convencionales y exige aproximaciones metodológicas específicas.

El presente trabajo aborda este desafío mediante la implementación comparativa de seis clasificadores supervisados: *Naive Bayes* como línea base, LDA, QDA, Discriminante Lineal de Fisher, y k-vecinos más cercanos. Esta progresión permite evaluar cómo diferentes supuestos teóricos se comportan ante el desbalance de clases, proporcionando insights valiosos para la selección de modelos según restricciones operativas específicas.

Cada algoritmo se entrena y evalúa considerando tanto métricas globales (exactitud, AUC) como específicas por clase (precisión, sensibilidad, F1-score), con especial atención al balance entre capturar la clase minoritaria y mantener la eficiencia operativa.

El objetivo principal es identificar la técnica más adecuada para asistir a gestores de campañas en la priorización de clientes, utilizando el dataset *Bank Marketing* como base empírica. Adicionalmente, se busca caracterizar cómo la naturaleza desbalanceada de los datos influye en el desempeño de diferentes familias de clasificadores y proponer estrategias de mitigación aplicables en contextos reales de marketing bancario.

2 Preprocesamiento de los Datos

En este trabajo se adoptó la versión estandarizada del dataset *Bank Marketing* (ID=222) del repositorio UCI, accedida mediante `ucimlrepo`. Mientras el estudio original de Moro et al. Moro et al. (2014) partió de 150 características requiriendo un complejo proceso de selección, nuestro enfoque demuestra que el conjunto pre-currado de 16 variables disponible en el repositorio abierto captura la información esencial para la predicción de suscripciones.

La implementación mediante script Python no es una limitación, sino una elección que garantiza la reproducibilidad exacta del análisis y la escalabilidad para futuras actualizaciones del dataset, el cual consistía de:

- **features.csv**: Contiene las 16 variables predictoras con 45,211 registros
- **targets.csv**: Contiene la variable objetivo binaria (suscripción: *yes*"/*no*")

La variable objetivo presenta un marcado desbalanceo: solo el 12.3 % de los contactos resultaron en suscripciones exitosas, replicando el patrón observado en el estudio original y permitiendo una comparación directa de metodologías en condiciones realistas de desbalance de clases.

2.1 Variables Predictoras

Las 16 variables predictoras del archivo *features.csv* se clasifican en cuatro categorías principales:

Cuadro 1: Clasificación de Variables Predictoras

Categoría	Variables	Descripción
Demográficas	age, job, marital, education	Características sociodemográficas del cliente
Financieras	default, balance, housing, loan	Situación financiera y productos contratados
Contexto de Contacto	contact, month, day_of_week, duration	Metadatos de la interacción actual
Historial de Marketing	campaign, pdays, previous, poutcome	Información de contactos previos

2.2 Decisiones de Preprocesamiento

El preprocesamiento se implementó mediante un script en Python, que aplica las siguientes transformaciones:

1. Se eliminaron las variables **contact** y **duration**. La primera presenta más de 13,000 valores faltantes, mientras que la segunda constituye un caso de *data leakage*[1] .

2. Se imputaron los valores faltantes de **job** y **education** asignando la categoría *unknown*, ya que la ausencia de información puede estar asociada a características no observadas de los clientes, y no resulta adecuado sustituirlos por una media o moda que distorsione la distribución real.

3. En la variable **poutcome**, los 36,959 valores faltantes (casi el 80 % del dataset) no se consideran aleatorios. Estos corresponden a clientes que nunca habían sido contactados en campañas previas, lo que constituye un caso de *Missing Not At Random* (MNAR). Dado que los valores observados —*success*, *failure* y *other*— resultan muy informativos sobre el historial del cliente, se optó por crear una categoría adicional (*no_contacted*) que preserva la naturaleza estructural de los faltantes. De este modo, el modelo puede diferenciar entre clientes con historial exitoso, fallido, desconocido o inexistente, evitando la pérdida de información que supondría imputar con la moda.

4. Las variables categóricas nominales (*job*, *marital*, *education*, *default*, *housing*, *loan*, *poutcome*) se codificaron con *One-Hot Encoding*, mientras que las variables ordinales (*month*, *day_of_week*) se codificaron con *Label Encoding*.

5. La variable objetivo se transformó a binaria: $y = 1$ para "*yes*" y $y = 0$ para "*no*".

6. Para reducir el desbalanceo de clases, se habilitó la opción de aplicar submuestreo moderado de la clase mayoritaria.

7. Finalmente, para algoritmos sensibles a la escala de las variables (como LDA y QDA), se aplicó un escalado estándar de las características numéricas.

[1] Uso de información en el proceso de entrenamiento del modelo que no se esperaría que estuviera disponible en el momento de la predicción

Estas decisiones aseguran que las variables predictoras utilizadas estén disponibles antes de la llamada, que los faltantes sean tratados de acuerdo con su mecanismo subyacente y que los modelos resultantes sean aplicables en escenarios reales de campañas de telemarketing.

3 Modelado

Para el modelado se entrenaron los clasificadores discutidos en clase: *Naive Bayes*, LDA, QDA, Fisher y k-vecinos más cercanos (*k-NN*). Cada clasificador se entrenó sobre los datos preprocesados, considerando el manejo de valores faltantes, variables categóricas y escalado apropiado para los clasificadores lineales y cuadráticos.

3.1 Naive Bayes

El clasificador *Naive Bayes* se incluyó como línea base, a pesar de que sus supuestos fundamentales no son razonables para este dataset:

- **Independencia condicional**: El supuesto de que las características son independientes dado la clase no se cumple en este dataset.

Se implementó utilizando la versión gaussiana de *scikit-learn*, entrenada con datos escalados para mantener consistencia con el preprocesamiento general, aunque este escalado no es requisito para el algoritmo.

A pesar de estas violaciones, *Naive Bayes* proporciona un punto de referencia útil.

3.2 K-NN

Para el clasificador *k-NN* se aplicó validación cruzada de 3 particiones para evaluar los hiperparámetros más relevantes:

- Número de vecinos $k \in \{1, \ldots, 20\}$.
- Esquema de pesos: *distance*.
- Métrica de distancia: Euclidiana.

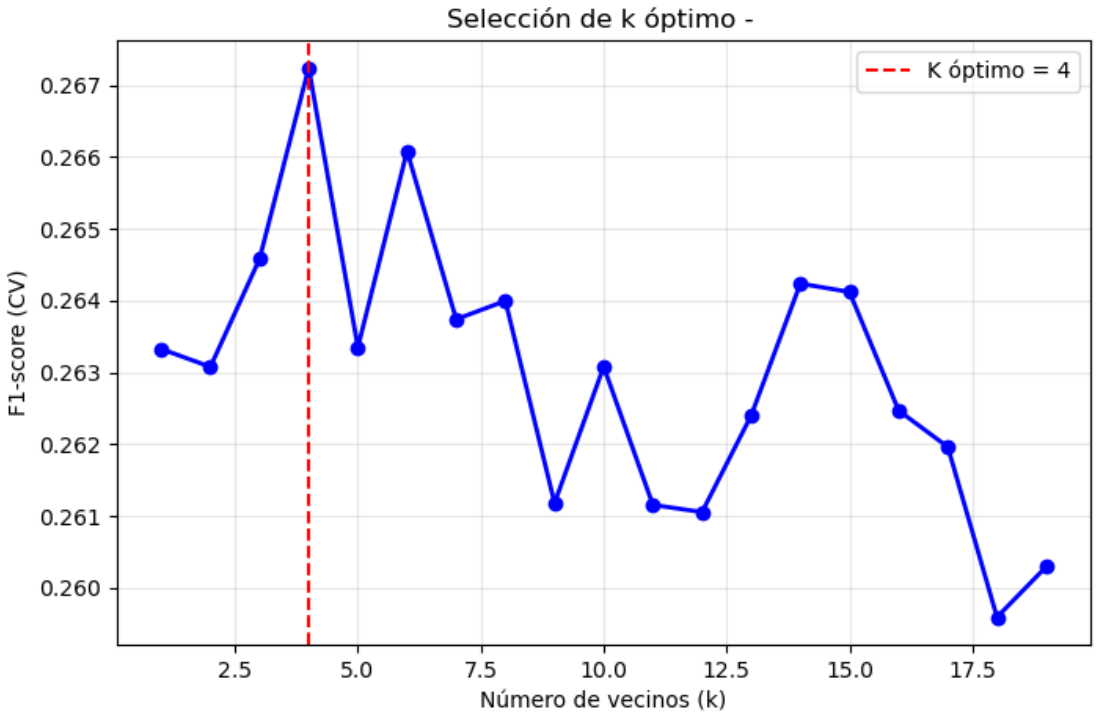


Figura 1: Elección del óptimo valor de k bajo el criterio de maximizar la métrica F1

3.3 LDA y QDA

Ambos clasificadores se entrenaron utilizando datos escalados mediante *StandardScaler* para cumplir con los supuestos de normalidad. El preprocesamiento incluyó codificación one-hot para variables categóricas y manejo específico de valores faltantes.

- **LDA**: Se utilizó la implementación de *scikit-learn* sin especificación de *priors*, permitiendo que el algoritmo estime las probabilidades a priori de las clases directamente desde los datos.
- **QDA**: Similar a LDA, pero modelando matrices de covarianza específicas para cada clase, capturando así relaciones no lineales entre características.

3.4 Discriminante Lineal de Fisher

El *Discriminante Lineal de Fisher* (FLD) busca proyectar datos de un espacio de dimensión p sobre una única dirección $a \in R^p$, de manera que la separación entre clases sea máxima en dicha proyección.

Sean $m_0, m_1 \in R^p$ los vectores de medias de cada clase y $S_0, S_1 \in R^{p\times p}$ sus matrices de covarianza respectivas. Se define la matriz de dispersión intra-clase como

$$S_w = S_0 + S_1.$$

El criterio de Fisher se expresa mediante el cociente de Rayleigh

$$J(a) = \frac{(a^\top (m_1 - m_0))^2}{a^\top S_w a},$$

el cual cuantifica la razón entre la varianza inter-clase (separación entre las medias proyectadas) y la varianza intra-clase (dispersión de los datos proyectados).

El objetivo consiste en encontrar la dirección a que maximiza $J(a)$, lo que equivale a resolver el problema de optimización

$$a^* = \arg\max_a J(a).$$

La solución óptima viene dada por:

$$a^* = S_w^{-1}(m_1 - m_0),$$

donde S_w^{-1} denota la inversa de la matriz de dispersión intra-clase.

Una vez obtenida la dirección óptima a^*, podemos proyectar cualquier vector $x \in R^p$ en el espacio unidimensional mediante la transformación:

$$z = a^{*\top} x.$$

Para realizar la clasificación, se establece un umbral t que separa las clases en el espacio proyectado.(Véase Silva (2019)) Dado un vector de entrada x:

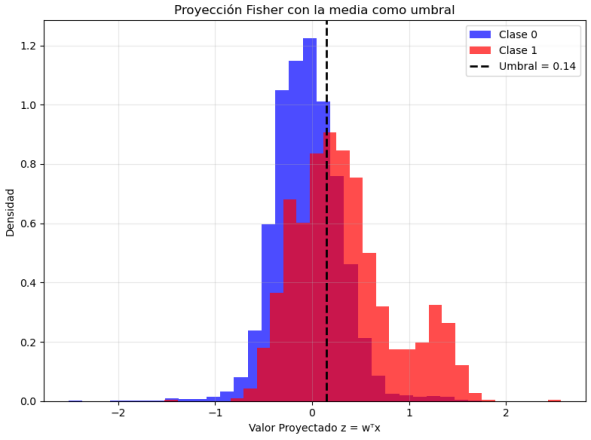
- Si $z \geq t$, entonces x pertenece a la clase C_1 (clase 1)
- Si $z < t$, entonces x pertenece a la clase C_0 (clase 0)

En esta implementación, el umbral t se calcula como el punto medio entre las medias proyectadas de ambas clases:

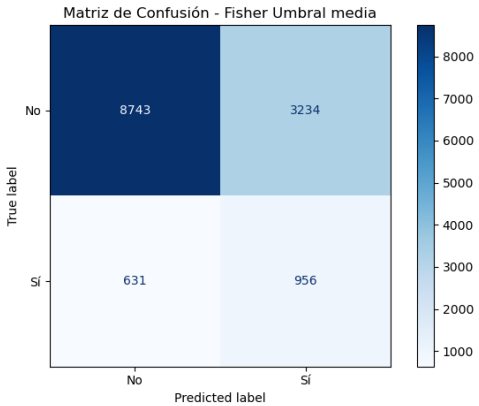
$$t = \frac{\bar{z}_0 + \bar{z}_1}{2},$$

donde $\bar{z}_0$ y $\bar{z}_1$ representan las medias de las proyecciones de las clases 0 y 1, respectivamente.

Los resultados de la aplicación del Discriminante Lineal de Fisher se muestran en la Figura 2:



(a) Distribución de proyecciones con umbral óptimo



(b) Matriz de confusión del clasificador

Figura 2: Resultados del Discriminante Lineal de Fisher

En la Figura 2(a) se observa la distribución de las proyecciones unidimensionales de los vectores de prueba. Las barras azules representan la clase 0, mientras que las rojas corresponden a la clase 1. La línea vertical discontinua indica el umbral t calculado, el cual separa ambas distribuciones. La superposición entre las distribuciones es considerable.

La Figura 2(b) presenta la matriz de confusión del clasificador, donde se puede observar que la concentración de valores en la primera fila de la matriz corresponde a que la distribución de casos cuya etiqueta verdadera es ”no”predomina en el dataset.

Ajuste del Umbral de Decisión para Datos Desbalanceados

Dado el desbalance inherente en el conjunto de datos, se implementó una variante del método de Fisher que optimiza el umbral de clasificación maximizando el $F1$-score en lugar de utilizar el punto medio entre las medias proyectadas.

Sea $z = a^{*\top} x$ la proyección de un dato x en la dirección óptima a^*. El clasificador estándar utiliza el umbral:

$$t_{\text{std}} = \frac{\bar{z}_0 + \bar{z}_1}{2},$$

donde $\bar{z}_0, \bar{z}_1$ son las medias proyectadas de cada clase.

En la versión balanceada, se busca el umbral t^* que maximiza el $F1$-score:

$$t^* = \arg\max_t F1(t).$$

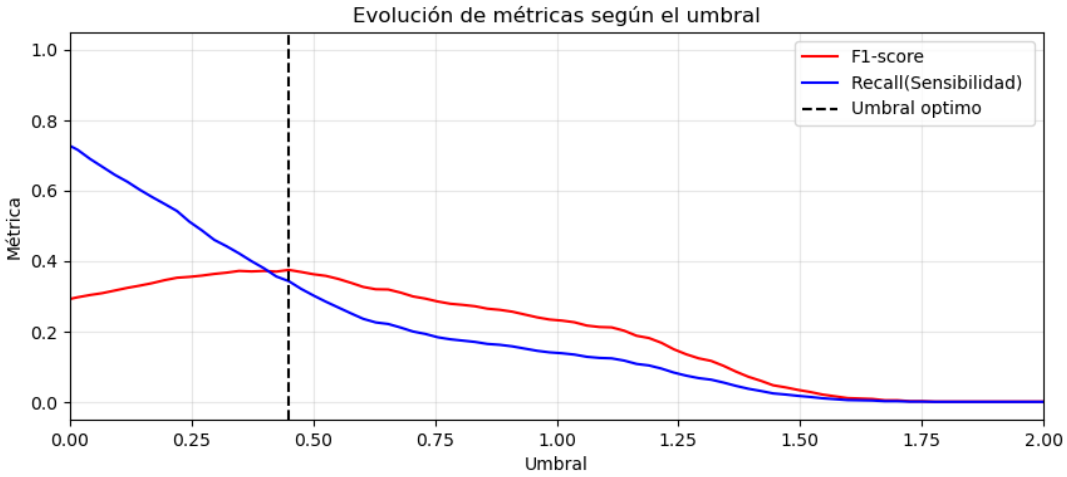
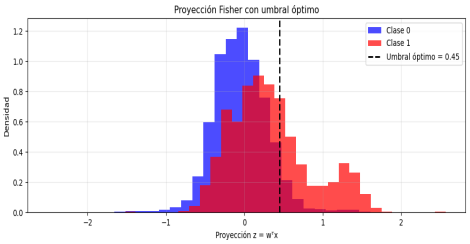
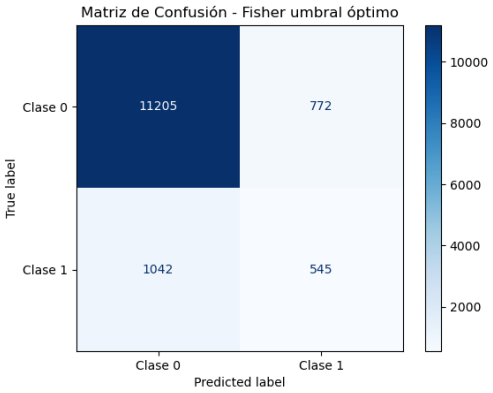


Figura 3: Elección del umbral óptimo por medio del criterio $F1$ y la sensibilidad

La Figura 3 revela el *trade-off* entre precisión y exhaustividad. De esta manera tenemos que bajo este umbral la clasificación es la siguiente:



(a) Distribución de proyecciones con umbral óptimo



(b) Matriz de confusión del clasificador

Figura 4: Resultados del Discriminante Lineal de Fisher con un umbral optimizado

4 Métricas de Desempeño

En esta sección se presentan los resultados de los diferentes clasificadores implementados, evaluados mediante métricas robustas que consideran el desbalance inherente del dataset.

Cuadro 2: Métricas de desempeño de los clasificadores implementados

Clasificador	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-Score**	**AUC**
LDA	0.8938	0.6637	0.1865	0.2912	0.7191
k-NN (k=4)	0.8758	0.4350	0.2067	0.2802	0.6524
Fisher (Ajustado)	0.8663	0.4138	0.3434	0.3753	0.7221
QDA	0.8224	0.3058	0.4077	0.3494	0.7039
Naive Bayes	0.8047	0.2771	0.4159	0.3326	0.6937
Fisher (Standard)	0.7151	0.2282	0.6024	0.3310	0.7221

Cuadro 3: Especificidad y análisis de matrices de confusión

Clasificador	**Especificidad**	**VP**	**FP**	**FN**	**VN**
LDA	0.9875	296	150	1291	11827
k-NN (k=4)	0.9644	328	426	1259	11551
Fisher (Ajustado)	0.9355	545	772	1042	11205
QDA	0.8773	647	1469	940	10508
Naive Bayes	0.8563	660	1722	927	10255
Fisher (Standard)	0.7300	956	3234	631	8743

Nota: VP: Verdaderos Positivos, FP: Falsos Positivos, FN: Falsos Negativos, VN: Verdaderos Negativos.

En base a la evidencia proporcionada por los datos y presentadas en el Cuadro 3 , Cuadro 2, se pueden observar los siguientes patrones:

- **LDA** logra la mayor exactitud (0.8938) pero con sensibilidad extremadamente baja (0.1865), indicando una marcada tendencia a clasificar instancias como clase mayoritaria.
- **Fisher Standard** muestra el patrón opuesto: alta sensibilidad (0.6024) pero precisión muy baja (0.2282), reflejando sobre-detección de la clase minoritaria.
- **Fisher con umbral ajustado** encuentra el mejor balance: F1-score más alto (0.3753) manteniendo exactitud razonable (0.8663).

Al comparar los clasificadores lineales y cuadráticos:

- **LDA vs QDA**: QDA mejora significativamente la sensibilidad (0.4077 vs 0.1865) a costa de exactitud, demostrando mayor flexibilidad para capturar fronteras no lineales en el espacio de características.

Considerando la relación entre exactitud y F1-score, se observa una dicotomía entre métricas globales y específicas por clase:

- **LDA** posee la mayor exactitud pero el peor F1-score, demostrando categóricamente que la exactitud puede ser una métrica engañosa en contextos desbalanceados.
- **Fisher ajustado** tiene la tercera exactitud general pero el mejor F1-score, posicionándose como el clasificador más equilibrado para el problema.

Teniendo en cuenta el contexto del problema, donde los agentes de telemercadeo deben evitar que los clientes se sientan acosados por llamadas excesivas, se recomiendan los siguientes clasificadores según diferentes objetivos estratégicos:

- **Para maximizar detección de clientes potenciales:**
  - *Fisher Standard*: Si el costo de falsos negativos es muy alto y se prioriza la cobertura.
  - *Fisher Ajustado*: Para balance óptimo entre precisión y exhaustividad.
- **Para eficiencia operativa:**
  - *LDA*: Si se prioriza minimizar contactos innecesarios (alta precisión) y costos operativos.

No obstante, si se debe escoger un único clasificador, se recomienda **Fisher Linear Discriminant con umbral ajustado**, pues emerge como la elección óptima al ofrecer el mejor compromiso entre todas las métricas relevantes y demostrar empíricamente la importancia de adaptar los clasificadores a la distribución real de los datos.

Su superioridad en F1-score (0.3753) combinada con un AUC competitivo (0.7221) y sensibilidad adecuada (0.3434) lo posiciona como el clasificador más adecuado para el problema de marketing bancario bajo estudio, donde tanto la identificación correcta de clientes potenciales como la eficiencia en el uso de recursos son consideraciones críticas.

Referencias

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Silva, T. S. (2019). An illustrative introduction to fisher's linear discriminant. Accessed: 2024.