

Proyecto 2: Clasificación Supervisada

Introducción a la Ciencia de Datos

Integrantes:	Sánchez, Hazel; Hernández, Debany ; Canché, Elías
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas (CIMAT)
Profesor:	Dr. Marco Antonio Aquino López

Resumen

1. Introducción

En el marketing digital y la gestión de relaciones con el cliente, la capacidad de predecir comportamientos del consumidor es una herramienta importante para las empresas. A diferencia del marketing convencional, el marketing digital permite una interacción directa y personalizada con los consumidores facilitando la promoción de productos y servicios, además de la recolección de datos sobre el comportamiento y las preferencias del público objetivo. En un entorno comercial cada vez más competitivo, esta capacidad de capturar y analizar información se traduce en ventajas de mercado.

La Ciencia de Datos, con técnicas de clasificación supervisada, proporciona herramientas para identificar patrones ocultos, predecir comportamientos futuros y optimizar recursos, también es posible obtener modelos predictivos que identifican clientes con mayor probabilidad de responder positivamente a una campaña, lo cual podría maximizar el retorno de inversión y fortalecer las estrategias de fidelización del cliente.

En este trabajo aplicamos diversos métodos de clasificación supervisada sobre la base de datos “Bank Marketing”, la cual documenta campañas de marketing directo realizadas por un banco portugués entre 2008 y 2013, cuyo objetivo fundamental fue identificar clientes propensos a suscribir depósitos a plazo.

Más allá de la implementación técnica, proponemos examinar cómo la naturaleza de los datos influye en el desempeño de cada clasificador. El análisis se estructura en tres fases: una exploración exhaustiva de las variables y sus relaciones, un procesamiento que incluye codificación de variables categóricas y manejo de desbalance de clases (en caso de encontrar), y una evaluación de los métodos mediante métricas múltiples que capturen distintos aspectos del desempeño predictivo.

2. Exploración inicial de los datos

2.1. *Bank Marketing Dataset.*

El **Bank Marketing Dataset** (`bank-full.csv`) fue desarrollado y publicado por el Instituto de Sistemas e Informática (INESC) de la Universidad de Lisboa, Portugal, como parte de un estudio longitudinal sobre efectividad de campañas de marketing en el sector bancario. Los datos fueron recolectados entre mayo de 2008 y noviembre de 2013 a través de campañas de marketing telefónico realizadas por una institución bancaria portuguesa y contiene información sobre clientes y el resultado de campañas de marketing telefónicas.

La estructura general de la base de datos consta de 45,211 observaciones, 16 variables predictoras y 1 variable de respuesta que indica si el cliente aceptó (`yes`) o no (`no`) contratar el depósito a plazo. En la tabla [??] se describen a detalle las variables, mientras que en las tablas [??] y [??] se muestra un resumen sobre la información de la base de datos respecto a cada variable.

El análisis exploratorio reveló problemas en las variables categóricas que requieren tratamiento específico durante el preprocesamiento. Se identificó alto desbalance en varias variables: la variable `default` presenta una concentración extrema del 98.9 % en la categoría “no”, lo que sugiere considerar su eliminación del modelo; `poutcome` muestra una distribución problemática con 85.3 % en “nonexistent” solo 0.2 % en “success”, recomendándose agrupar categorías; y `loan` presenta un desbalance moderado con 82.0 % en “no”, donde técnicas

Cuadro 1: Clasificación de Variables Predictoras del Bank Marketing Dataset

Variable	Tipo	Escala	Descripción y Características
age	Numérica	Continua	Edad del cliente (17-98 años). Distribución aprox. normal.
job	Categórica	Nominal	Ocupación 13 categorías.
marital	Categórica	Nominal	Estado civil: 'divorced', 'married', 'single', 'unknown'.
education	Categórica	Ordinal	Nivel educativo: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'.
default	Categórica	Binaria	Crédito en mora: 'no', 'yes', 'unknown'.
housing	Categórica	Binaria	Préstamo hipotecario: 'no', 'yes', 'unknown'.
loan	Categórica	Binaria	Préstamo personal: 'no', 'yes', 'unknown'.
contact	Categórica	Nominal	Tipo de contacto: 'cellular', 'telephone'.
month	Categórica	Ordinal	Mes último contacto: 'jan' a 'dec'.
day_of_week	Categórica	Ordinal	Día último contacto: 'mon' a 'fri'.
duration	Numérica	Continua	Duración de contacto (segundos).
campaign	Numérica	Discreta	Número de contactos en campaña actual.
pdays	Numérica	Discreta	Días desde último contacto.
previous	Numérica	Discreta	Número de contactos anteriores.
poutcome	Categórica	Nominal	Resultado campaña anterior: 'failure', 'nonexistent', 'success'.
emp.var.rate	Numérica	Continua	Tasa variación empleo (trimestral).
cons.price.idx	Numérica	Continua	Índice precios al consumidor (mensual).
cons.conf.idx	Numérica	Continua	Índice confianza del consumidor (mensual).
euribor3m	Numérica	Continua	Tasa Euribor 3 meses (diaria).
nr.employed	Numérica	Continua	Número de empleados (trimestral).

Cuadro 2: Estadísticas Descriptivas de Variables Numéricas

Variable	Media	Desv. Est.	Mínimo	Máximo	Asimetría
age	40.94	10.62	17	98	0.69
duration	258.28	259.28	0	4918	4.97
campaign	2.76	3.10	1	63	9.13
pdays	962.48	186.91	0	999	-4.50
previous	0.17	0.49	0	7	6.57
emp.var.rate	0.08	1.57	-3.4	1.4	-0.79
cons.price.idx	93.58	0.58	92.20	94.77	-0.37
cons.conf.idx	-40.50	4.63	-50.8	-26.9	0.57
euribor3m	3.62	1.73	0.63	5.04	-0.73
nr.employed	5167.03	72.25	4964	5228	-0.53

de ponderación podrían mejorar el modelado. Adicionalmente, el problema de valores desconocidos afecta significativamente a education (5.0 % unknown), donde la imputación por moda resulta apropiada; job (0.3 % unknown) podría resolverse mediante eliminación o imputación simple; y default (0.9 % unknown) donde considerar unknown como categoría separada preservaría información potencialmente valiosa. Estos hallazgos destacan la necesidad de estrategias de preprocesamiento diferenciadas para garantizar la calidad predictiva de los modelos.

Respecto a los **valores faltantes**, no se detectaron celdas con NaN. Sin embargo, algunas variables categóricas presentan la categoría unknown, la cual representa información faltante implícita (especialmente en job, marital, education y contact). Este aspecto deberá considerarse en la fase de preprocesamiento, ya sea tratándola como una categoría válida o mediante imputación o eliminación. En cuanto a los **duplicados**, no se encontraron registros repetidos.

Finalmente, la **variable de respuesta** y se encuentra desbalanceada: aproximadamente el **88.3 % de los registros corresponden a "no"** y solo el **11.7 % a "yes"**. Este desbalance es relevante, pues implica que las técnicas de modelado deberán considerar métodos para equilibrar la predicción, como métricas apropiadas (F1, recall) o técnicas de balanceo de clases.

Cuadro 3: Resumen Estadístico de Variables Categóricas

Variable	Nº Cat.	Moda	Distribución y Observaciones
job	12	admin. (22.9 %)	admin. (22.9 %), blue-collar (21.6 %), technician (15.7 %), services (11.5 %), management (10.9 %), retired (4.0 %), entrepreneur (3.6 %), self-employed (3.5 %), housemaid (2.8 %), unemployed (2.1 %), student (1.0 %), unknown (0.3 %)
marital	4	married (60.2 %)	married (60.2 %), single (28.5 %), divorced (11.1 %), unknown (0.2 %)
education	8	university.degree (22.8 %)	university.degree (22.8 %), high.school (20.7 %), basic.9y (14.6 %), professional.course (14.2 %), basic.4y (10.8 %), basic.6y (9.8 %), unknown (5.0 %), illiterate (0.2 %)
default	3	no (98.9 %)	no (98.9 %), unknown (0.9 %), yes (0.2 %) Alto desbalance - Considerar eliminar o agrupar
housing	3	yes (55.6 %)	yes (55.6 %), no (44.4 %), unknown (0.1 %)
loan	3	no (82.0 %)	no (82.0 %), yes (18.0 %), unknown (0.1 %)
contact	2	cellular (64.4 %)	cellular (64.4 %), telephone (35.6 %)
month	12	may (30.5 %)	may (30.5 %), jul (13.7 %), aug (13.1 %), jun (12.1 %), nov (9.6 %), apr (5.8 %), oct (4.9 %), mar (4.3 %), sep (3.2 %), dec (1.3 %), jan (0.9 %), feb (0.5 %) Estacionalidad - Picos en primavera/verano
day_of_week	5	thu (22.8 %)	thu (22.8 %), wed (22.3 %), tue (21.3 %), mon (20.7 %), fri (12.9 %)
poutcome	3	nonexistent (85.3 %)	nonexistent (85.3 %), failure (14.5 %), success (0.2 %) Desbalance extremo - Solo 0.2 % de éxitos previos

3. Preprocesamiento

El preprocesamiento tiene como finalidad transformar la base de datos **Bank Marketing** (`bank-full.csv`) en una matriz de diseño adecuado para el modelado supervisado. En particular, buscamos,

- i) representar de forma numérica las variables categóricas preservando su información,
- ii) homogenizar la escala de las variables numéricas para algoritmos sensibles a la magnitud,
- iii) separar correctamente predictores X y respuesta y evitando fugas de información.

Preparación y limpieza

Como paso inicial, se normalizó el formato de las variables categóricas (*trimming* de espacios y conversión a minúsculas) para evitar duplicidades por diferencias de capitalización o espacios. La variable de respuesta y se codificó como binaria (0 = *no*, 1 = *yes*).

Codificación de variables categóricas

Las variables cualitativas (`job`, `marital`, `education`, `contact`, `month`, `poutcome`, entre otras) se trataron como nominales y se transformaron mediante **One-Hot Encoding** (creación de indicadores binarios por categoría). Esta estrategia nos ayuda a:

- Evitar imponer un orden arbitrario sobre categorías no ordinales.
- Permite a los modelos lineales asignar efectos específicos por categoría.
- Facilita la interpretación al nivel de categoría (coeficientes por indicador).

Para garantizar una matriz de diseño de *rango completo*, se utilizó la opción `drop_first=True`, eliminando una categoría de referencia por variable (conocido como *dummy variable trap*). En términos prácticos, si `marital` tiene categorías `{married, single, divorced}`, tras la codificación con referencia en `married` se generan dos columnas (`marital_single, marital_divorced`); un registro con ambos indicadores en 0 se interpreta como la categoría de referencia (`married`).

Tratamiento de la categoría unknown. En este conjunto de datos, la ausencia de información mencionamos que no aparece como *NaN*, sino como la etiqueta `unknown` en varias variables categóricas. Para esto, se decidió **conservar `unknown` como categoría explícita** por tres motivos:

- i) su frecuencia es no despreciable y, por lo tanto, informativa;
- ii) evitar imputaciones potencialmente sesgadas al desconocer el mecanismo de ausencia;
- iii permitir que el modelo *aprenda* si la condición de desconocido posee valor predictivo propio.

Alternativamente, podría imputarse la moda o agruparse categorías raras en una clase `other`; estas variantes son útiles si se busca compactar dimensionalidad o si la dispersión de clases induce esparsidad excesiva.

Escalado de variables numéricas

Las variables numéricas (`age, balance, day, duration1, campaign, pdays, previous`, etc.) se estandarizaron con **StandardScaler**, es decir, cada columna x se transformó a:

$$z = \frac{x - \mu}{\sigma},$$

donde μ es la media y σ la desviación estándar de la variable. Esta normalización centra las variables en media 0 y varianza unitaria, lo que:

1. Evita que magnitudes grandes (`balance`) dominen sobre otras (`campaign`).
2. Mejora la estabilidad numérica y la convergencia en modelos como Regresión Logística y SVM.
3. Hace comparables los coeficientes (en modelos lineales) en términos de desviaciones estándar.

La interpretación de un valor estandarizado es directa: un $z = 1,2$ en `age` indica que el individuo está 1,2 desviaciones estándar *por encima* de la media de edad; un $z = -0,5$ en `balance` indica que está 0,5 desviaciones estándar *por debajo* de su media.

Separación de X e y y dimensiones resultantes

Tras la codificación y el escalado, se separaron los predictores en X y la respuesta en y (con $y \in \{0, 1\}$). La matriz X resultante contiene tanto las variables numéricas estandarizadas como los indicadores (*dummies*) generados por One-Hot Encoding. En nuestra ejecución, se obtuvieron dimensiones del tipo:

$$X \in \mathbb{R}^{n \times p}, \quad y \in \{0, 1\}^n,$$

donde n es el número de observaciones ($n = 45,211$) y p depende del número de categorías activas tras la codificación (puede superar varias decenas). El chequeo de `shape` corrobora la consistencia entre filas de X y longitud de y .

¹La variable `duration` se registra al final de la llamada y puede inducir fuga de información si se usa para predecir el éxito de la campaña antes o durante la llamada. Por ello, aunque se reporta en la exploración, suele *excluirse* como predictor en modelos destinados a la toma de decisión previa.

Buenas prácticas y control de fuga de información

Para una evaluación honesta del desempeño, el ajuste (*fit*) de transformaciones debe realizarse *exclusivamente* sobre el conjunto de entrenamiento y luego aplicarse (*transform*) al de validación/prueba. Esto incluye el cálculo de medias y desviaciones del escalado así como el mapeo de categorías en One-Hot. En la implementación, esto se logra de forma segura mediante *pipelines* que encadenan *encoder* y *scaler* con el estimador final. Asimismo, se recomienda evaluar la exclusión de *duration* para escenarios de predicción en tiempo real y considerar técnicas para manejar el desbalance de clases (métricas como F1/recall, *class weights* o remuestreo).

4. Modelado

5. Análisis computacional mediante simulaciones (Segunda Parte de la Tarea)

Introducción

El propósito de esta segunda parte es estudiar, en un entorno controlado donde la distribución verdadera de las clases son conocidas, el comportamiento de los clasificadores vistos en el curso frente al clasificador óptimo (de Bayes).

Para este propósito, se diseñan y ejecutan simulaciones con datos sintéticos generados a partir de distribuciones normales multivariadas (en dos dimensiones), comparamos el riesgo de clasificación de cada método tanto contra el riesgo óptimo como contra estimadores obtenidos por métodos de validación, para estas comparaciones variamos parámetros como los tamaños de muestras y valores de k (en k -NN). Además, para simplificar el cálculo del error de Bayes teórico y compararlo con los otros errores, se utilizan muestras proporcionales y con covarianzas iguales para las dos clases.

Comportamiento de clasificadores

A continuación se muestran los resultados obtenidos en las simulaciones realizadas. Consideraremos los siguientes casos para cada simulación:

- **Clases separadas:** Mismas covarianzas para ambas clases.
- **Clases sobreuestas:** Distintas covarianzas para ambas clases.
- **Datos balanceados:** Mismos tamaños de muestras para ambas clases.
 - Mismas covarianzas
 - Distintas covarianzas
- **Datos desbalanceados:** Tamaños de muestras distintos (proporciones de 0.2 y 0.8).
 - Mismas covarianzas
 - Distintas covarianzas

Clases separadas

Los clasificadores con las clases separadas se comportaron como la Figura ??.

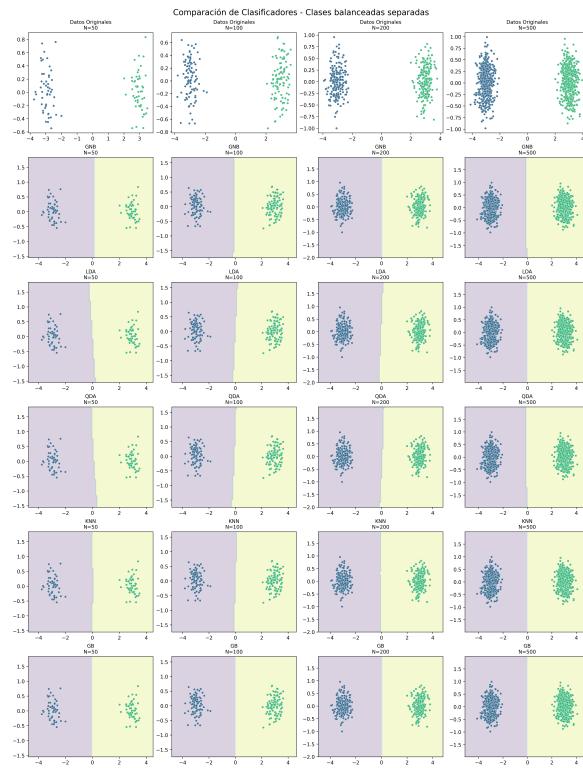


Figura 1: Comportamiento de clasificadores con datos balanceados y clases separadas.

Clases sobreuestas

Los clasificadores con las clases sobreuestas se comportaron como la Figura ??.

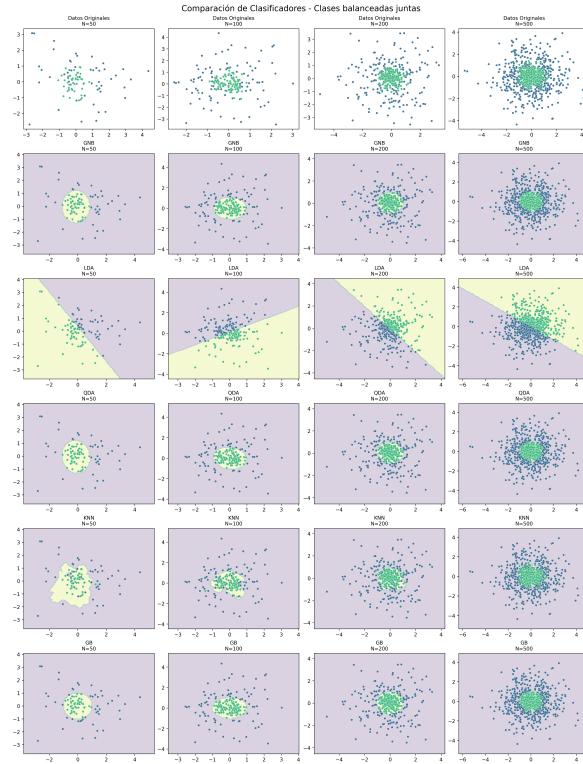


Figura 2: Comportamiento de clasificadores con datos balanceados y clases separadas.

Datos Balanceados

A partir de este punto, las simulaciones se realizan con las medias $\mu_0 = (-1, 0)$ y $\mu_1 = (0, 1)$ para sus respectivas clases, a menos que se indique lo contrario.

Para datos generados con la misma covarianza se observan las siguientes comparaciones en la Figura ??

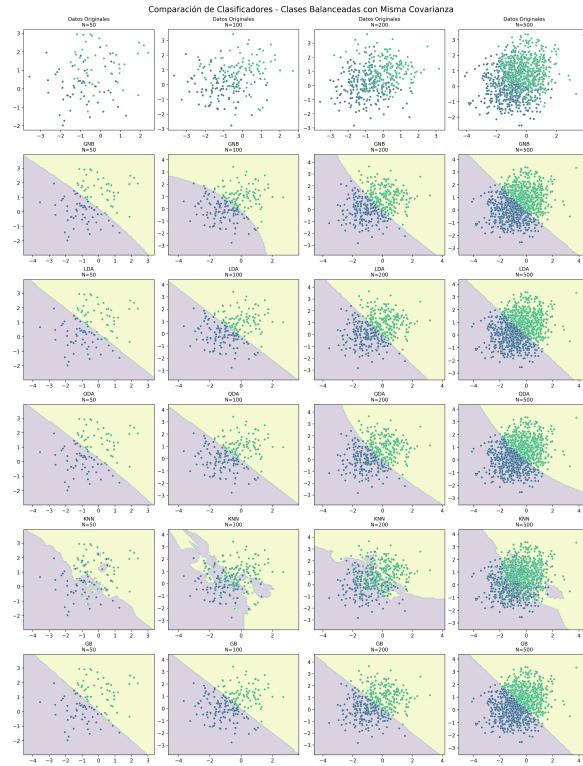


Figura 3: Comportamiento de clasificadores con datos balanceados y covarianzas iguales.

Para datos generados con distintas covarianzas se observan las siguientes comparaciones en la Figura ??

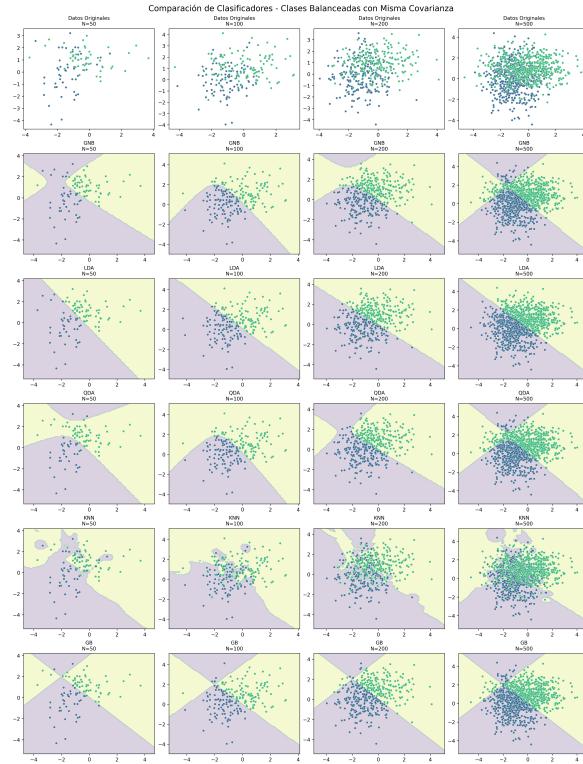


Figura 4: Comportamiento de clasificadores con datos balanceados y covarianzas distintas.

Datos Desbalanceados

Para datos generados con la misma covarianza se observan las siguientes comparaciones en la Figura ??

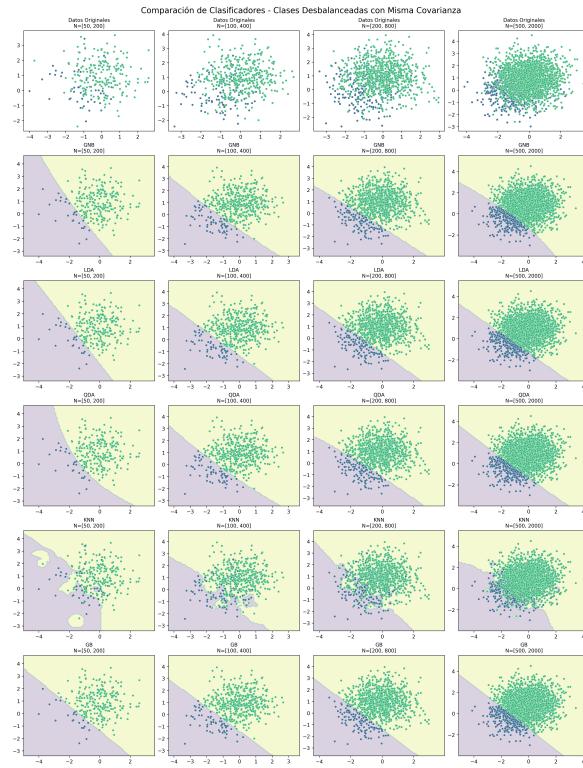


Figura 5: Comportamiento de clasificadores con datos desbalanceados y covarianzas iguales.

Para datos generados con distintas covarianzas se observan las siguientes comparaciones en la Figura ??

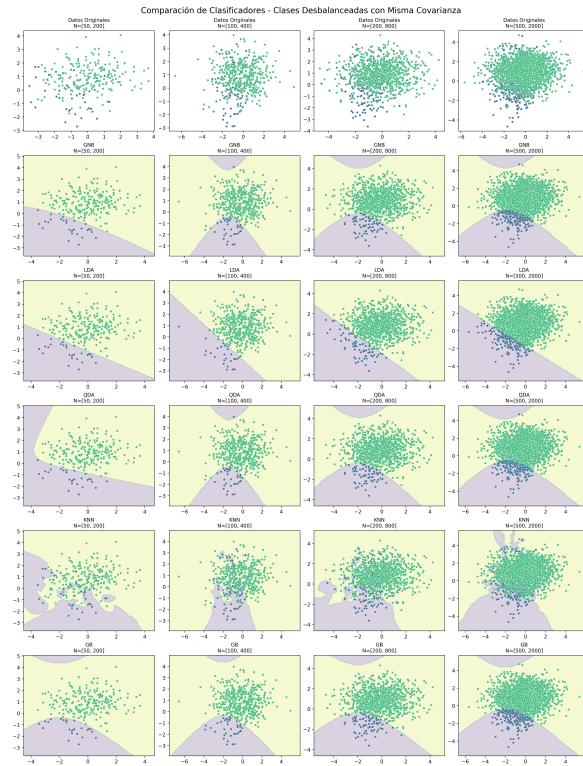


Figura 6: Comportamiento de clasificadores con datos desbalanceados y covarianzas distintas.

Variando K en K -NN

Ademas, en el algoritmo de k -NN se puede variar el valor de k para obtener diferentes resultados. De mismo modo, consideraremos clases balanceadas y clases desbalanceadas, ambas con las mismas covarianzas.

Para clases balanceadas se observan las siguientes comparaciones en la Figura ??.

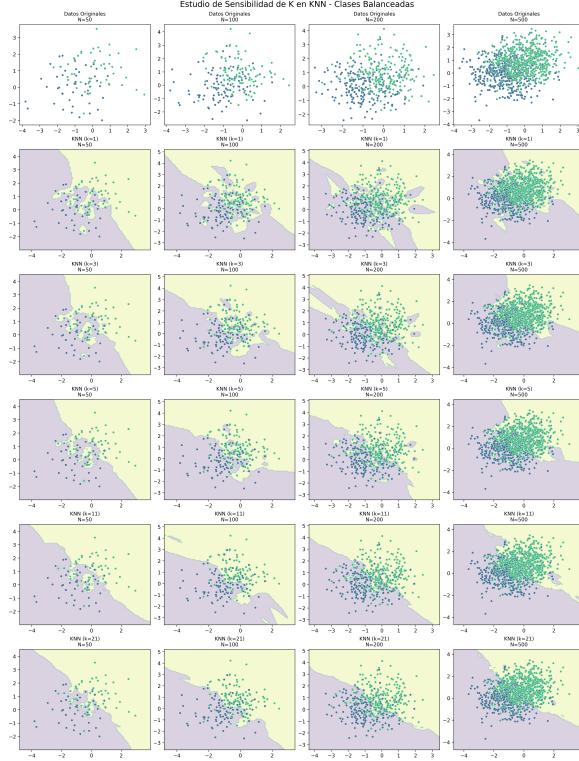


Figura 7: Comportamiento de clasificadores con datos desbalanceados y covarianzas distintas.

Para clases desbalanceadas se observan las siguientes comparaciones en la Figura ??.



Figura 8: Comportamiento de clasificadores con datos desbalanceados y covarianzas distintas.

Analisis de riesgo

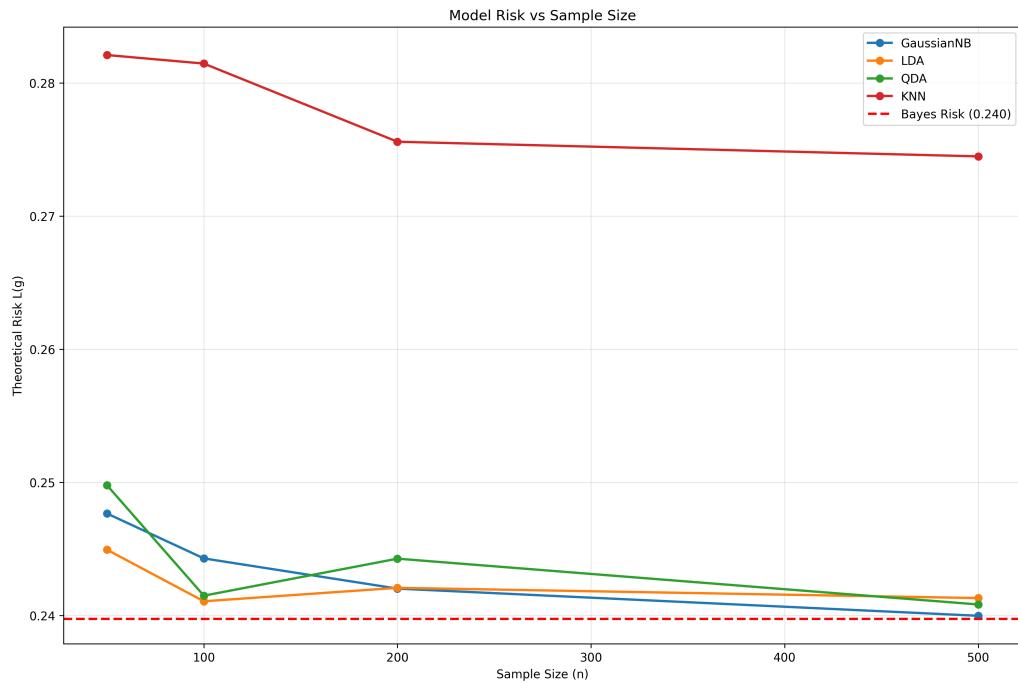


Figura 9:

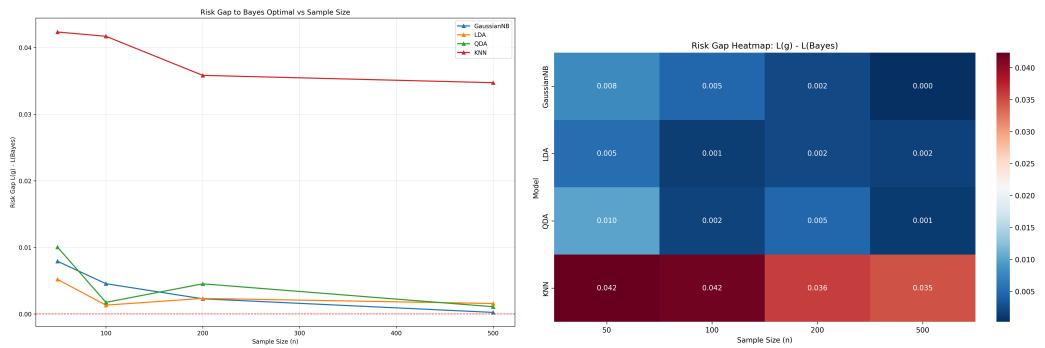


Figura 10: Risk Gaps $L(g) - L(\text{Bayes})$

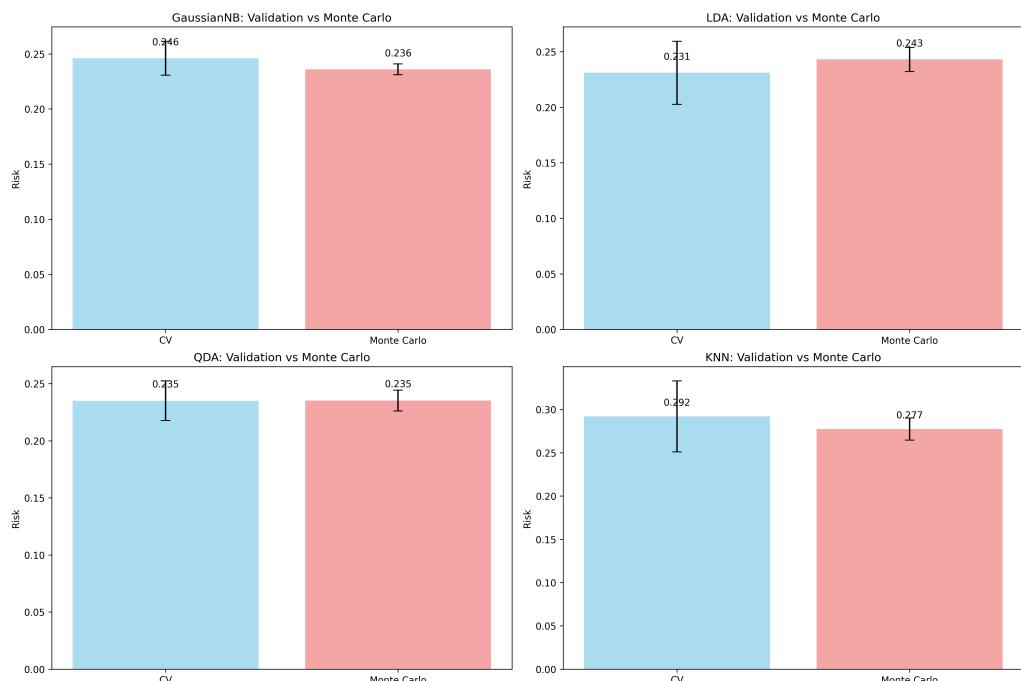


Figura 11: Monte Carlo vs Cross Validation

Cuadro 4: Resumen: Riesgos Promedio y Desviaciones Estándar

Modelo	Riesgo Promedio	Desv Estándar	Gap vs Bayes
GaussianNB	0.2435	0.0108	0.0037
LDA	0.2423	0.0103	0.0026
QDA	0.2441	0.0098	0.0043
KNN	0.2784	0.0161	0.0386
BAYES	0.2398	-	0

Cuadro 5: Comparación: Cross-Validation vs Monte Carlo

Modelo	CV Risk	CV Std	MC Risk	MC Std
GaussianNB	0.2460	0.0155	0.2359	0.0049
LDA	0.2310	0.0284	0.2430	0.0108
QDA	0.2350	0.0173	0.2351	0.0091
KNN	0.2920	0.0410	0.2774	0.0128