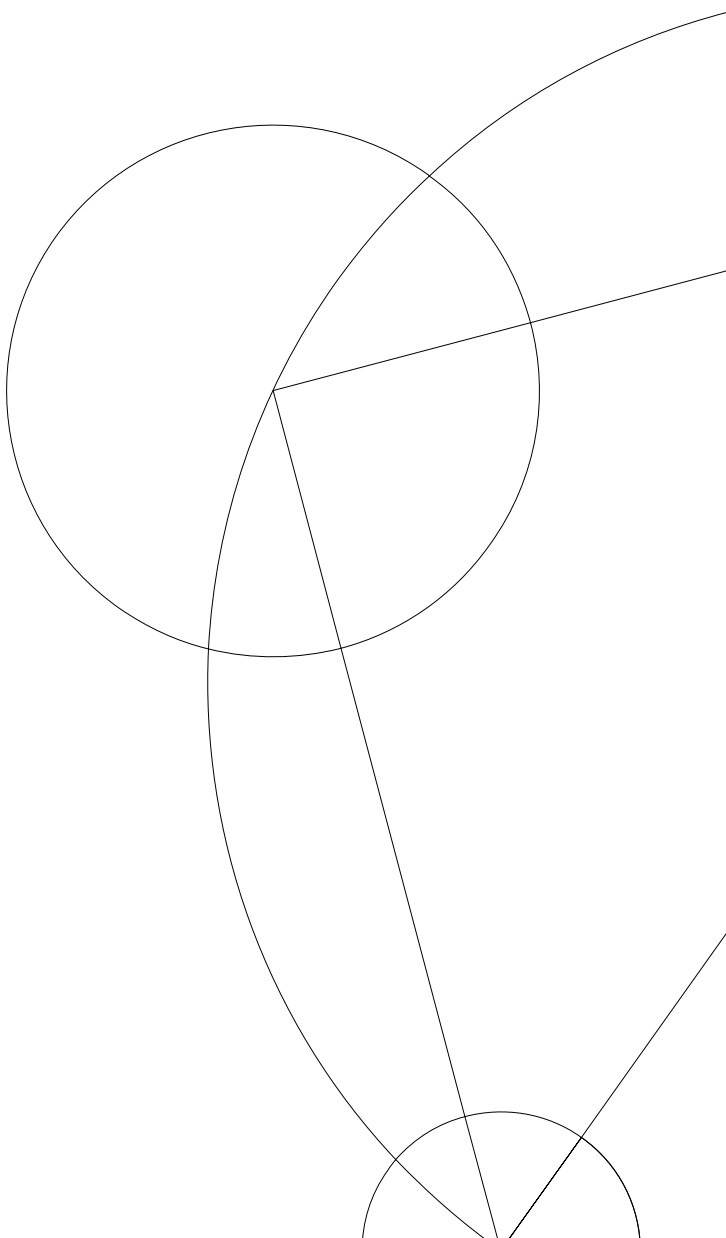
Estudio de simulación en clasificación supervisada

Alfredo Bistrain, Guillermo Aguilar, Oswaldo Bueno

Introducción a la Ciencia de Datos

Tarea 2



2 de octubre de 2025

Introducción

En este trabajo se estudia el comportamiento de diversos clasificadores supervisados en un entorno controlado, en el cual las distribuciones de las clases son conocidas y corresponden a distribuciones normales multivariadas.

El esquema metodológico consiste en generar datos sintéticos a partir de distribuciones normales multivariadas bajo distintos escenarios controlados, variando parámetros como las medias, matrices de covarianza y probabilidades a priori de las clases. Esto permite reproducir situaciones donde ciertos clasificadores resultan teóricamente óptimos (por ejemplo, LDA cuando las covarianzas son iguales, o QDA cuando difieren), así como casos de desbalance en las clases o fuerte correlación entre variables.

Dentro de los métodos considerados se incluyen:

- El clasificador óptimo de Bayes, utilizado como referencia teórica.
- El Análisis Discriminante Lineal (LDA) y el Análisis Discriminante Cuadrático (QDA).
- El método de los k vecinos más cercanos (k–NN).
- El clasificador de Fisher basado en una proyección unidimensional y un umbral de decisión.
- El método de Naive Bayes.

Para cada escenario y tamaño muestral n, se estiman los riesgos de clasificación mediante validación cruzada y se comparan con el riesgo verdadero calculado por Monte Carlo. El análisis incluye la construcción de tablas y gráficas que muestran:

1. El comportamiento del riesgo $L(g)$ en función de n.
2. La sensibilidad de k–NN respecto al número de vecinos k.
3. Las brechas $L(g) - L^*$ en relación al riesgo de Bayes.
4. La comparación entre riesgo estimado por validación cruzada y riesgo verdadero.

Este enfoque permite evaluar empíricamente la consistencia de los métodos, visualizar sus limitaciones en diferentes condiciones y reforzar la conexión entre teoría estadística y práctica computacional. En particular, se destacan los escenarios donde LDA o QDA coinciden con el clasificador de Bayes, así como la variabilidad del k–NN en función de los parámetros. De esta manera, el estudio no solo ilustra la aplicabilidad de estos algoritmos, sino que también sirve como base didáctica para comprender los principios fundamentales de la clasificación supervisada.

Implementación

Clasificación Bayesiana con densidades Gaussianas

Una de las cosas necesarias para llevar a cabo la implementación es el riesgo verdadero usando el error de Bayes y se optó por dar una expresión analítica. Sea $X \in \mathbb{R}^p$ una variable aleatoria gaussiana con media μ_k y matriz de covarianza Σ_k. Su densidad de probabilidad es

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\!\left(-\tfrac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\right).$$

El clasificador de Bayes asigna un punto x a la clase k que maximiza el producto $\pi_k f_k(x)$, donde π_k es la probabilidad a priori de la clase. Para simplificar los cálculos se toma el logaritmo natural, definiendo la función discriminante

$$\delta_k(x) = \log \pi_k + \log f_k(x).$$

Al sustituir la forma explícita de $f_k(x)$, se obtiene

$$\delta_k(x) = \log \pi_k - \tfrac{p}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma_k| - \tfrac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k).$$

El término $-\frac{p}{2}\log(2\pi)$ es común a todas las clases y no influye en la comparación, por lo que puede omitirse. Así, la regla de decisión depende únicamente de

$$\delta_k(x) = \log \pi_k - \tfrac{1}{2}\log|\Sigma_k| - \tfrac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k).$$

En código, el término cuadrático se calcula como

```
quad = np.einsum('ij,jk,ik->i', XC, invS, XC)
return np.log(pi) - 0.5*logdet - 0.5*quad.
```

El clasificador resultante es

$$g^*(x) = \arg\max_k \ \delta_k(x).$$

Para aplicar el operador arg máx se evalúa para cada observación de manera individual: dado un punto x, se calculan $\delta_0(x)$ y $\delta_1(x)$ (o, en general, $\delta_1(x), \ldots, \delta_K(x)$) y se asigna x a la clase correspondiente al valor máximo. Así cada observación queda clasificada según la función discriminante que resulte mayor.

Clasificador de Fisher (proyección 1D con umbral)

El clasificador de Fisher fue propuesto en [Fisher, 1936] y discutido más recientemente en [Hastie et al., 2009]. Consiste en proyectar los datos $x \in \mathbb{R}^p$ sobre una dirección $w \in \mathbb{R}^p$ que maximiza la separabilidad entre clases. La proyección se define como:

$$z = w^\top x.$$

La dirección de Fisher se obtiene resolviendo

$$w \propto S_W^{-1}(\mu_1 - \mu_0),$$

donde $S_W = \Sigma_0 + \Sigma_1$ es la matriz de dispersión intra–clases, y μ_k, Σ_k son la media y covarianza de cada clase $k = 0, 1$.

Una vez proyectados los datos, se calculan las medias

$$m_0 = w^\top \mu_0, \qquad m_1 = w^\top \mu_1,$$

y la varianza común proyectada s^2. El umbral t de clasificación se fija como

$$t = \frac{m_0 + m_1}{2}, \quad \text{si } \pi_0 = \pi_1,$$

y en el caso general con probabilidades a priori (π_0, π_1):

$$t = \frac{m_0 + m_1}{2} + \frac{s^2}{m_1 - m_0} \log\left(\frac{\pi_0}{\pi_1}\right).$$

La regla de decisión es entonces

$$g(x) = \begin{cases} 0, & \text{si } w^\top x \le t, \\ 1, & \text{si } w^\top x > t. \end{cases}$$

Clasificador de Fisher (proyección 1D con umbral)

La implementación en `Python` se llevó a cabo en dos pasos:

1. Se definió una función `fit_fisher_1d` que, a partir de un conjunto de entrenamiento (X, y), estima w, las medias proyectadas m_0, m_1, la varianza s^2 y calcula el umbral t.
2. Se implementó una función `predict_fisher_1d` que asigna la clase 0 o 1 según la regla de decisión anterior.

Para integrarlo al flujo de validación cruzada junto con LDA, QDA y k–NN, se definió un envoltorio tipo estimador de `scikit-learn`:

```
class Fisher1DClassifier(BaseEstimator, ClassifierMixin):
    def __init__(self, pi0=0.5, pi1=0.5): ...
    def fit(self, X, y): ...
    def predict(self, X): ...
```

De esta manera, el clasificador de Fisher puede ser evaluado en el mismo esquema de simulación Monte Carlo y validación cruzada que los demás métodos.

Diseño experimental

Se generó un vector Y de tamaño n con entradas binarias de acuerdo con probabilidades a priori (π_0, π_1) fijadas en cada escenario. Condicionado en la clase, se simularon muestras i.i.d. según

$$X \mid Y = 0 \sim \mathcal{N}_p(\mu_0, \Sigma_0), \qquad X \mid Y = 1 \sim \mathcal{N}_p(\mu_1, \Sigma_1).$$

Sobre estos datos se aplicaron los clasificadores LDA, QDA y k-NN, y se estimó su riesgo de clasificación para compararlo contra el riesgo del clasificador óptimo de Bayes. El estudio se realizó variando el tamaño muestral $n \in \{50, 100, 200, 500\}$ y el número de vecinos en k-NN, $k \in \{1, 3, 5, 11, 21\}$. Para cada combinación de parámetros se efectuaron $R = 20$ réplicas independientes, reportando media $\pm$ desviación estándar del riesgo.

Bajo estas condiciones se consideraron cinco escenarios de simulación:

1. **LDA óptimo:** $\pi_0 = \pi_1 = 0.5$ y $\Sigma_0 = \Sigma_1$. En este caso la frontera de decisión es lineal y queremos mostrar que LDA coincide con el clasificador de Bayes.
2. **QDA óptimo:** $\pi_0 = \pi_1 = 0.5$ y $\Sigma_0 \neq \Sigma_1$. La frontera es cuadrática y QDA se busca ejemplificar que coincide con el clasificador de Bayes.
3. **Desbalance de clases:** $\Sigma_0 = \Sigma_1$ pero $\pi_0 = 0.8,\ \pi_1 = 0.2$. Este escenario evalúa la sensibilidad de los métodos ante el desbalance en las clases.
4. **Alta correlación:** se consideran covarianzas con correlaciones altas y similares en ambas clases, con priori balanceadas $(\pi_0 = \pi_1 = 0.5)$.
5. **Medias cercanas:** se fijan $\pi_0 = \pi_1 = 0.5$, medias próximas $\mu_0 = (0, 0)$ y $\mu_1 = (0.5, 0.5)$, junto con matrices de covarianza altamente correlacionadas. Este caso genera un traslape considerable y representa un escenario difícil de clasificación.

Resultados de la modelación

A continuación se presenta la tabla de resultados de los modelos LDA, QDA, Naive Bayes y Fisher en el caso balanceado con $\Sigma_0 = \Sigma_1$. El riesgo de Bayes está dado por aproximadamente $L_*(Bayes) \approx 0.0386$. En este caso se toman

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}, \quad \Sigma_0 = \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Los resultados son los siguientes.

n	LDA	QDA	Naive Bayes	Fisher
50	0.047 ± 0.033	0.051 ± 0.041	0.050 ± 0.036	0.049 ± 0.038
100	0.045 ± 0.024	0.044 ± 0.020	0.043 ± 0.020	0.042 ± 0.020
200	0.042 ± 0.015	0.044 ± 0.014	0.043 ± 0.016	0.043 ± 0.015
500	0.038 ± 0.009	0.038 ± 0.008	0.039 ± 0.010	0.039 ± 0.009

Cuadro 1: Riesgos estimados en el caso balanceado con $\Sigma_0 = \Sigma_1$

Debido al gran número de casos de los parámetros n y k, los resultados del modelo k-NN no se presentan en tabla, pero sí se presentan de manera gráfica. A continuación se presentan gráficas en las cuales se podrán comparar los riesgos estimados de cada modelo con respecto al riesgo de Bayes

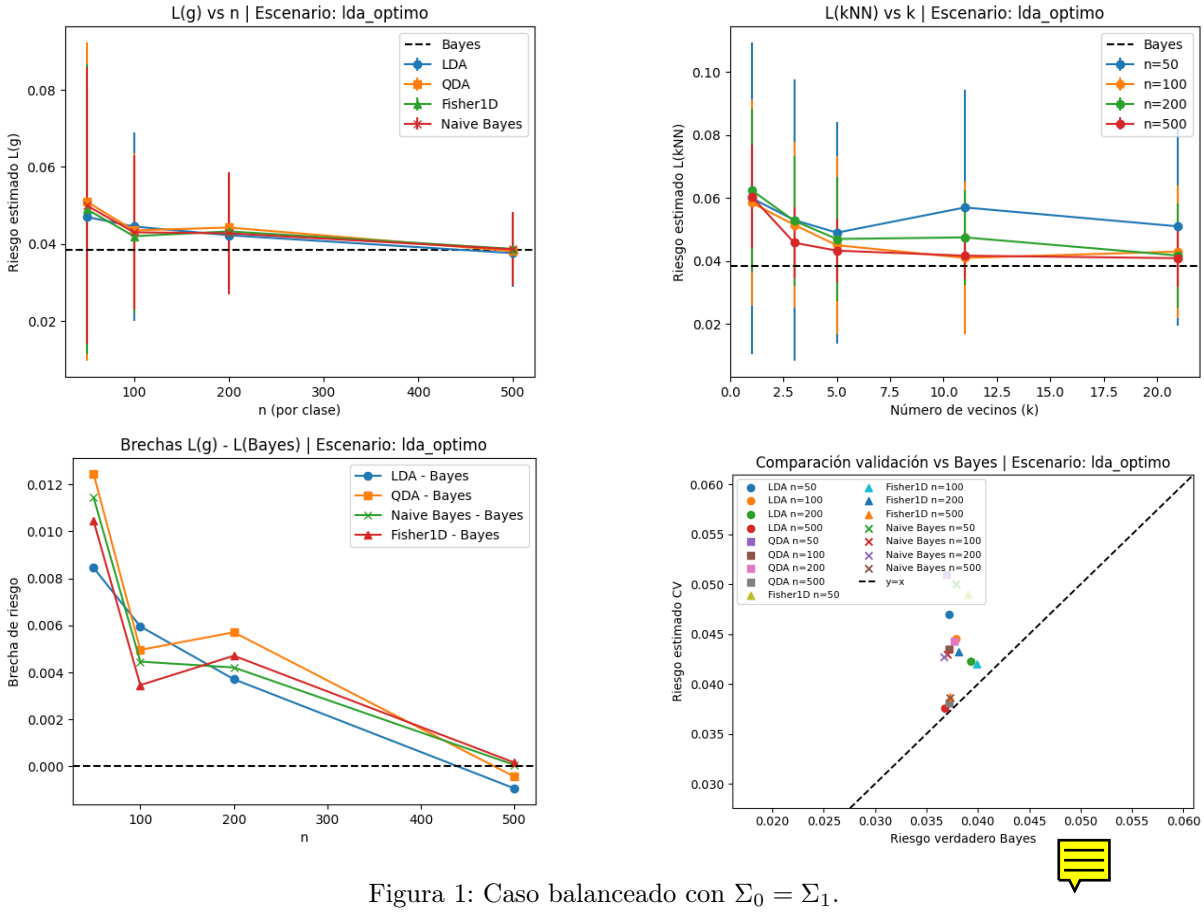


Figura 1: Caso balanceado con $\Sigma_0 = \Sigma_1$.

En lo anterior puede verse que el modelo LDA disminuye el riesgo más que los otros modelos, pero todos tienen un muy buen desempeño. Esto puede estar influenciado por el hecho de tener covarianzas iguales y medias distintas relativamente alejadas, de manera que la clasificación no es difícil para ninguno de los modelos.

Los métodos de k-NN muestran alta varianza para k pequeños (especialmente $k = 1$ y aproximadamente $k = 11$), y un mal desempeño, pero presentan mejor estabilidad para valores moderados como $k = 11$ o $k = 21$, aunque aparentemente, sin alcanzar la optimalidad de LDA o QDA.

Para el caso balanceado con $\Sigma_0 \neq \Sigma_1$, en el cual el clasificador óptimo es el QDA, se tiene $L_*(Bayes) \approx 0.2358$. En este caso se toman

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1.2 \\ 1.2 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1.2 \end{pmatrix}$$

Los resultados son los siguientes.

n	LDA	QDA	Naive Bayes	Fisher
50	0.244 ± 0.073	0.256 ± 0.079	0.239 ± 0.064	0.239 ± 0.077
100	0.230 ± 0.049	0.232 ± 0.051	0.227 ± 0.050	0.234 ± 0.058
200	0.258 ± 0.027	0.253 ± 0.029	0.257 ± 0.028	0.258 ± 0.028
500	0.246 ± 0.017	0.245 ± 0.014	0.243 ± 0.016	0.244 ± 0.017

Cuadro 2: Riesgos estimados en el caso balanceado con $\Sigma_0 \neq \Sigma_1$

Las gráficas de los riesgos estimados se presentan a continuación.

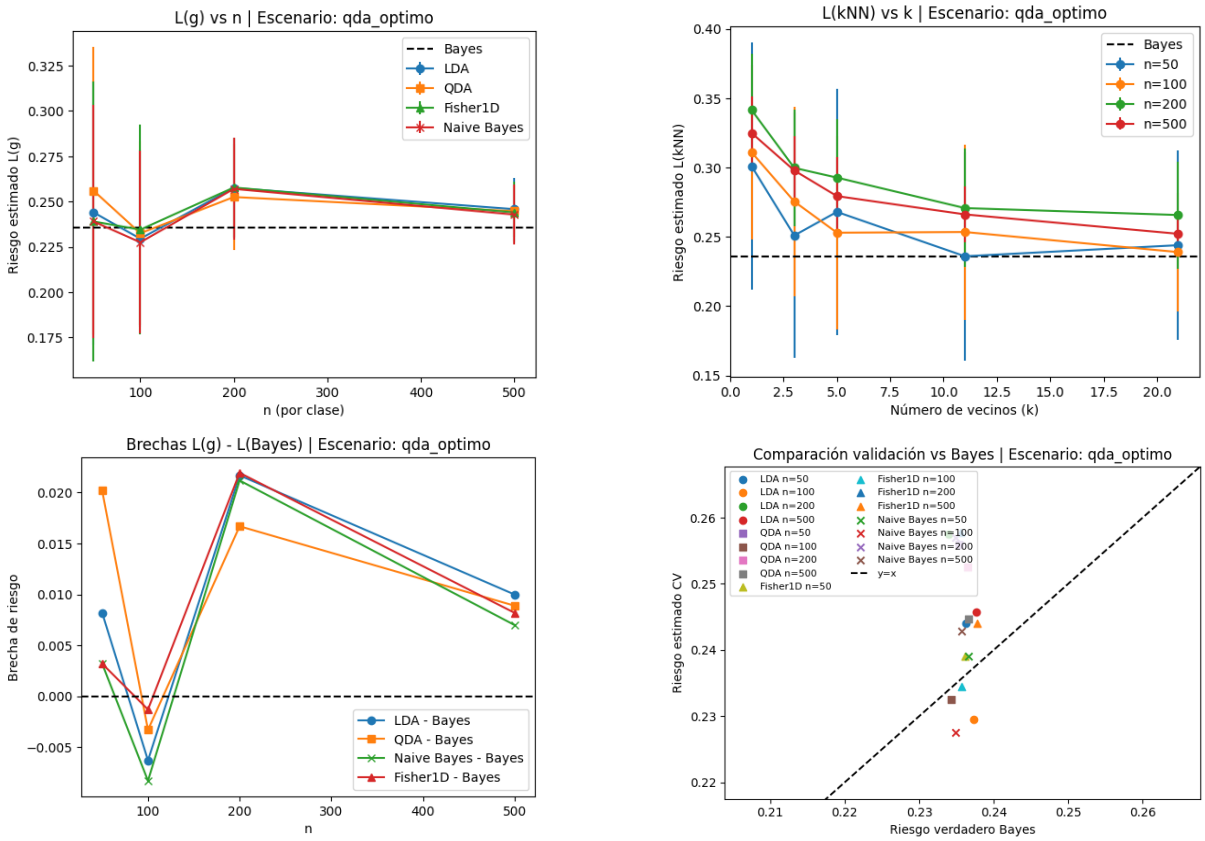


Figura 2: Caso balanceado con $\Sigma_0 \neq \Sigma_1$.

En este escenario observemos como el modelo QDA no es aquel que resulta tener el mejor desempeño, aunque no por mucho. Además, no se aproxima tanto al riesgo de Bayes a pesar de ser el clasificador óptimo. Esto puede deberse al hecho de que las medias son cercanas y la matriz de covarianzas Σ_1 toma valores "grandes" que permiten que los datos se mezclen.

Resulta interesante notar como para $n = 100$, el riesgo del modelo k-NN se aproxima bastante bien al riesgo de Bayes, pero incrementa para tamaños de muestra más grandes; el tan buen desempeño del modelo k-NN bajo $n = 100$ puede deberse a simple aleatoriedad, pues al incrementar el tamaño muestral, el clasificador debería tener un mejor desempeño.

Para el caso desbalanceado con $\Sigma_0 = \Sigma_1$, en el cual el clasificador óptimo es el LDA, se tiene $L_*(Bayes) \approx 0.0583$. En este caso se toman

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma_0 = \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

Los resultados son los siguientes.

n	LDA	QDA	Naive Bayes	Fisher
50	0.070 ± 0.032	0.073 ± 0.034	0.066 ± 0.028	0.064 ± 0.029
100	0.054 ± 0.038	0.058 ± 0.033	0.057 ± 0.031	0.058 ± 0.037
200	0.057 ± 0.016	0.060 ± 0.016	0.058 ± 0.017	0.062 ± 0.015
500	0.055 ± 0.009	0.056 ± 0.010	0.055 ± 0.010	0.056 ± 0.009

Cuadro 3: Riesgos estimados en el caso desbalanceado con $\Sigma_0 = \Sigma_1$

Las gráficas de los riesgos estimados se presentan a continuación.

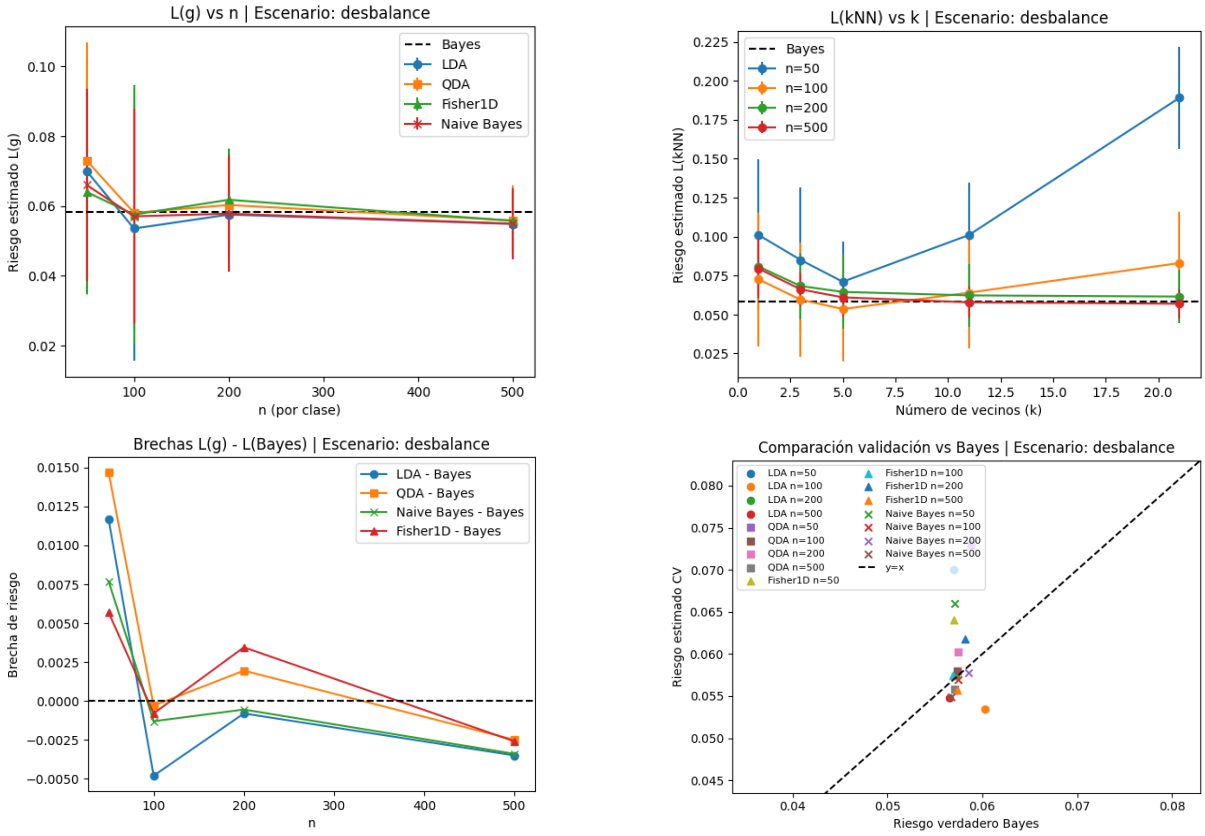


Figura 3: Caso desbalanceado con $\Sigma_0 = \Sigma_1$.

En este escenario puede observarse que a medida que el tamaño de muestra aumenta, los clasificadores empiezan a disminuir su riesgo incluso por debajo del riesgo de Bayes, y además tienen un comportamiento muy similar para $n = 200$ y $n = 500$. Dicho comportamiento debe provenir de haber tomado las covarianzas con valores pequeños y las medias ligeramente distanciadas.

En el caso del modelo k-NN nótese que el riesgo incrementó a medida que el parámetro k aumentaba su valor, y esto se debe justamente al hecho de tener clases desbalanceadas. Al revisar a los k vecinos más cercanos, por el hecho de tener una clase significativamente mayoritaria, dicha clase empezará a dominar las votaciones. En este tipo de situaciones es más adecuado considerar clasificación k-NN ponderada por distancia, así se atenúa un poco el efecto del desbalance.

Para el caso de correlación fuerte, en el cual se consideran

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix},$$

con clases balanceadas, donde el riesgo de Bayes que se obtuvo es $L_*(Bayes) \approx 0.1913$, los resultados son los siguientes.

n	LDA	QDA	Naive Bayes	Fisher
50	0.200 ± 0.048	0.192 ± 0.056	0.204 ± 0.049	0.206 ± 0.057
100	0.218 ± 0.038	0.205 ± 0.046	0.214 ± 0.032	0.218 ± 0.042
200	0.221 ± 0.033	0.200 ± 0.026	0.221 ± 0.032	0.221 ± 0.032
500	0.215 ± 0.017	0.194 ± 0.022	0.212 ± 0.016	0.214 ± 0.016

Cuadro 4: Riesgos estimados en el caso de correlación fuerte

Las gráficas de los riesgos estimados se presentan a continuación.

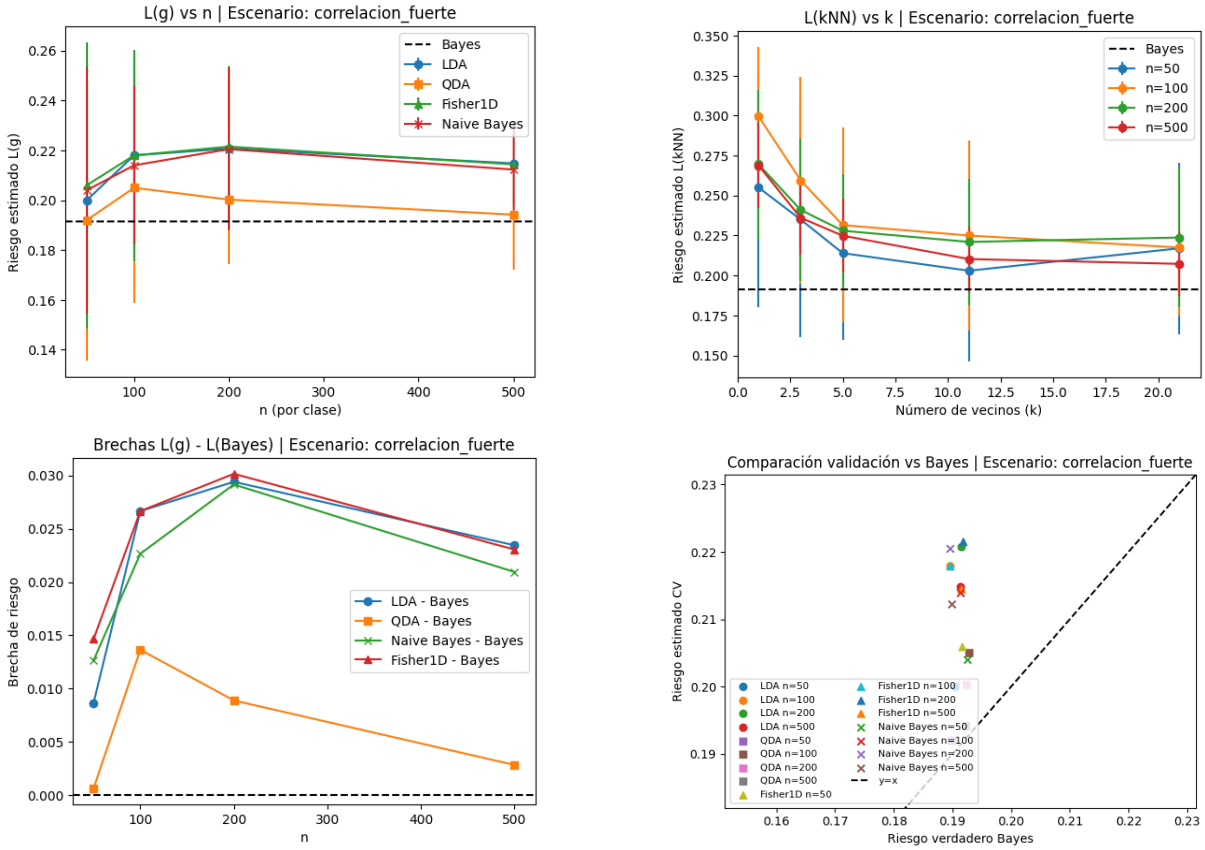


Figura 4: Caso balanceado con correlación fuerte.

En el escenario de correlación fuerte puede observarse que el modelo QDA tiene un desempeño notablemente mejor que los demás modelos, esto pues las matrices de covarianzas son distintas y se sabe que en dicho caso el clasificador óptimo es el QDA. Más aún, al tener medias relativamente cercanas es mayor el efecto en los riesgos de los clasificadores y es más notable la optimalidad del QDA.

Finalmente se considera un escenario de medias más cercanas que en el caso anterior, con clases balanceadas. Los parámetros considerados son,

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}.$$

Los resultados en este escenario son los siguientes.

n	LDA	QDA	Naive Bayes	Fisher
50	0.385 ± 0.080	0.355 ± 0.097	0.399 ± 0.086	0.389 ± 0.083
100	0.384 ± 0.057	0.352 ± 0.068	0.390 ± 0.052	0.385 ± 0.060
200	0.418 ± 0.049	0.352 ± 0.035	0.409 ± 0.042	0.410 ± 0.040
500	0.401 ± 0.013	0.348 ± 0.024	0.401 ± 0.016	0.397 ± 0.012

Cuadro 5: Riesgos estimados en el caso de medias cercanas

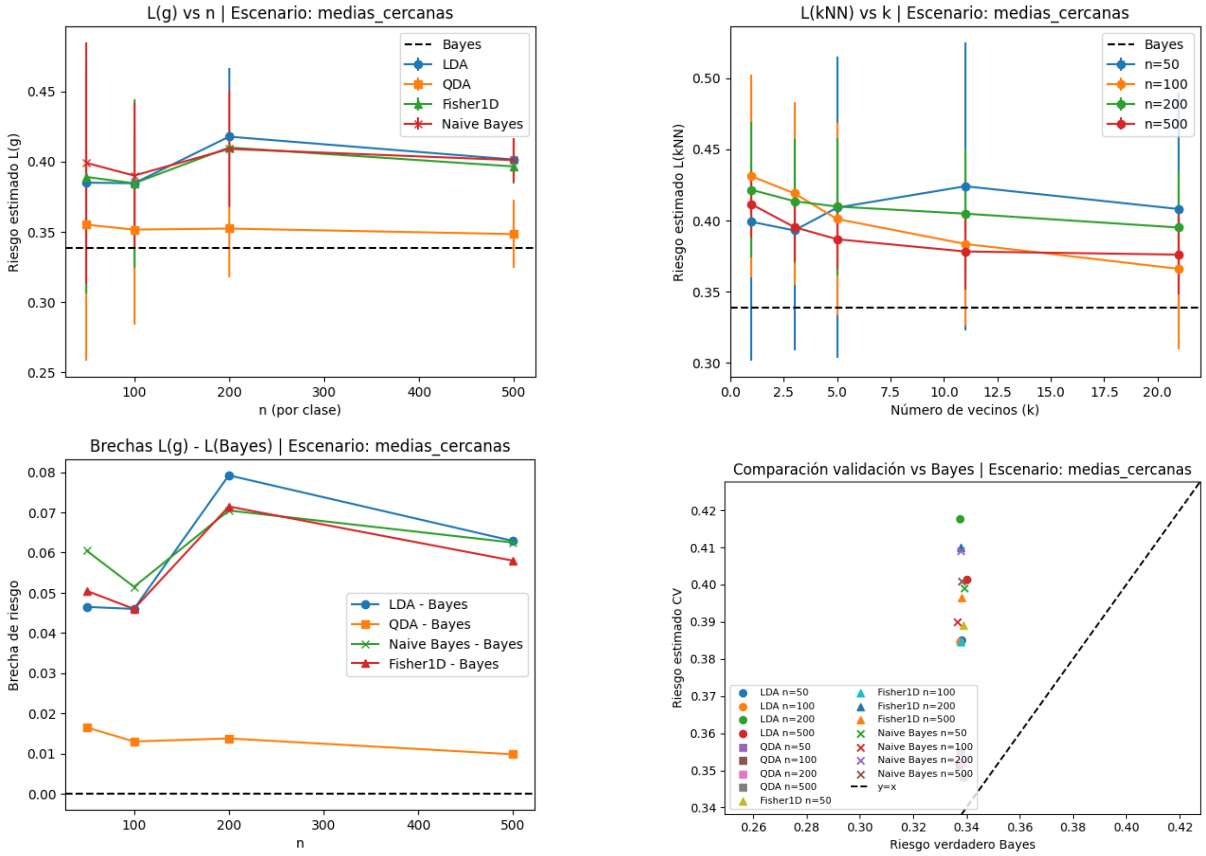


Figura 5: Caso medias cercanas.

En este último escenario puede observarse un comportamiento muy similar al caso anterior, ya que en ambas situaciones las medias se tomaron cercanas (aunque en este caso se acentuó más esta cercanía), y al tener matrices de covarianzas distintas se vuelve más visible la optimalidad del modelo QDA con respecto a los demás.

Conclusiones

El estudio de simulación permitió contrastar distintos clasificadores supervisados bajo escenarios controlados y conociendo de antemano el clasificador de Bayes como referencia óptima. A partir de los resultados pueden destacarse varias observaciones relevantes:

- LDA es muy competitivo bajo homocedasticidad. Esto a su vez asegura la validez teórica de este método en tales condiciones.
- QDA es preferible cuando los datos se encuentran muy mezclados, mientras que en condiciones sencillas tiene un desempeño tan bueno como los demás.
- k-NN ofrece flexibilidad y no depende de supuestos paramétricos, pero su desempeño varía fuertemente con k y n. La elección del parámetro k debe hacerse cuidadosamente teniendo en cuenta si existe desbalance entre las clases; considerar la ponderación por distancia es crucial también.
- Bajo condiciones dificiles (es decir, medias cercanas y covarianzas distintas, de manera que los datos se encuentran significativamente mezclados) Naive Bayes y Fisher mostraron un desempeño notablemente peor al clasificador óptimo, pero se comportaron de manera muy similar. En casos de condiciones fáciles ocurre lo mismo, ya que se muestran desempeños igual de buenos que los demás clasificadores.
- En presencia de clases desbalanceadas, tanto LDA como QDA requieren el uso explícito de probabilidades a priori para evitar un sesgo hacia la clase mayoritaria. Este efecto es menos crítico en k–NN, aunque también tiende a degradar su precisión.

- Los escenarios de alta correlación y medias cercanas ilustran las limitaciones inherentes de todos los métodos, pues la superposición entre distribuciones produce un error mínimo (Bayes) relativamente alto. En tales casos, incluso los clasificadores óptimos no logran riesgos bajos.

En conjunto, los experimentos reafirman que no existe un clasificador superior, sino que la elección depende de las condiciones de los datos. LDA y QDA funcionan como métodos paramétricos muy eficientes cuando los supuestos son razonables, mientras que k–NN ofrece mayor flexibilidad a costa de una mayor sensibilidad al tamaño de muestra y a la selección de hiperparámetros. Comparar los riesgos estimados mediante validación cruzada contra el riesgo verdadero de Bayes brinda una perspectiva clara sobre la eficiencia y las limitaciones de cada algoritmo en distintos contextos.

Referencias

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition.