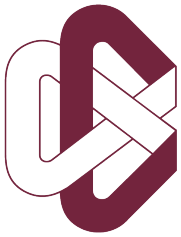


CIMAT

Análisis computacional mediante simulaciones

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 objetivo de este reporte es analizar, dentro de un entorno controlado, el desempeño de los diferentes métodos de clasificación estudiados en el curso frente al clasificador óptimo (de Bayes). Para ello, planteamos la comparación entre clasificadores tales como Naive Bayes, Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrático (QDA), k-vecinos más cercanos (k-NN) y el método de Fisher, frente al clasificador óptimo de Bayes, que constituye el punto de referencia teórico ideal.

Dado que trabajamos en un escenario simulado tenemos la ventaja de que la distribución verdadera de los datos es conocida y, por tanto, es posible calcular el riesgo de clasificación de manera exacta. Para esto, generamos datos sintéticos a partir de distribuciones normales multivariadas bajo distintos parámetros de medias y covarianzas, de manera que se puedan representar situaciones de separación más o menos compleja entre clases. De esta forma, no solo se podremos contrastar el rendimiento de los clasificadores en condiciones ideales, sino también podemos observar cómo varían sus resultados según el grado de dificultad del problema.

Centramos el análisis en medir el riesgo verdadero de cada método frente al riesgo mínimo alcanzable por el clasificador de Bayes, lo que permitirá evaluar la cercanía de cada técnica al óptimo.

Así, tenemos una perspectiva tanto teórica como empírica lo cual nos permite tener una visión más completa del comportamiento relativo de cada clasificador y aunado a esto podemos discutir ventajas, limitaciones y condiciones en las que cada uno resulta más adecuado.

2. Planteamiento.

La idea general de este estudio es considerar un problema de clasificación binaria con dos clases $\{0, 1\}$) y probabilidades a priori de manera que podamos explorar tanto situaciones balanceadas como desbalanceadas. Modelamos cada clase mediante una distribución normal multivariada con parámetros de media y covarianza elegidos para analizar distintos niveles de dificultad en la separación de clases:

- casos sencillos en los que las clases están claramente separadas,
- casos intermedios en los que la separación entre clases no es tan evidente,
- casos difíciles en los que las clases se traslapan entre ellas.

Dentro de este marco tomamos en cuenta dos escenarios principales:

- Covarianzas iguales: la frontera de decisión óptima es lineal y coincide con LDA.
- Covarianzas distintas: la frontera de Bayes es cuadrática, y en este caso QDA representa la solución óptima..

En cada escenario comparamos los riesgos de distintos clasificadores (Naive Bayes, LDA, QDA, Fisher y k-NN) frente al riesgo de Bayes. Para obtener resultados estables, generamos múltiples muestras simuladas de distintos tamaños y repetimos el procedimiento 20 veces. Con ello calculamos promedios y desviaciones estándar de los riesgos, y graficamos tanto los valores obtenidos como las brechas respecto al clasificador óptimo. Este enfoque permite observar no solo la convergencia de cada método hacia el riesgo óptimo, sino también la velocidad con que lo logran y el efecto de parámetros como el número de vecinos en k-NN.

3. Resultados de la simulación.

A continuación presentamos los resultados de las simulaciones para los distintos escenarios planteados. Cabe señalar que únicamente en el primer escenario (balanceado con covarianzas iguales) mostramos los tres casos de separación entre clases (fácil, intermedio y difícil) con el fin de ilustrar cómo este factor influye en los riesgos de los clasificadores. En los demás escenarios optamos por considerar únicamente el caso intermedio de separación, con el propósito de evitar que el análisis resultara excesivamente extenso o repetitivo y facilitar así la comprensión de los resultados.

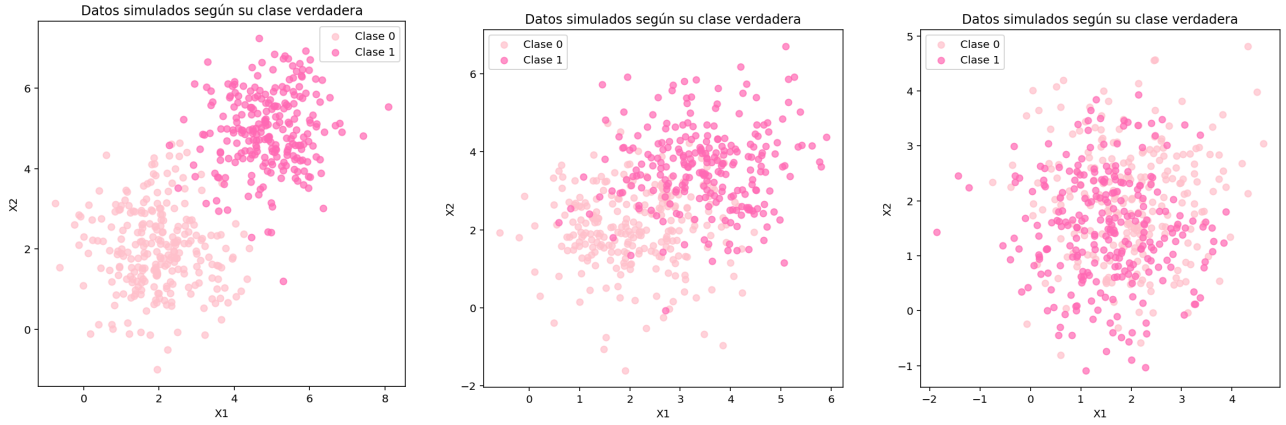
Caso balanceado.

- **Covarianzas iguales.**

  En este caso, consideramos covarianzas iguales y generamos tres escenarios de dificultad usando la matriz identidad como covarianza común y variando las medias:

  - Fácil: $\mu_0 = (2, 2)$ y $\mu_1 = (5, 5)$
  - Medio: $\mu_0 = (2, 2)$ y $\mu_1 = (3.5, 3.5)$.
  - Difícil: $\mu_0 = (2, 2)$ y $\mu_1 = (1.5, 1.5)$

  Presentamos en la Figura 1 las gráficas de los datos simulados según su correspondiente clase para apreciar visualmente los distintos grados de separación en cada caso.



(a) Grado de separación fácil. (b) Grado de separación medio. (c) Grado de separación difícil.

Figura 1: Datos simulados con diferentes grados de separación.

En la Figura 2 observamos la gráfica de los riesgos de los clasificadores de Naive Bayes, LDA, QDA y Fisher contra el error óptimo de Bayes (el cual es aproximadamente 0.016947), en el caso en que los datos tienen un grado de separación sencillo. En ella podemos observar como, en todos los tamaños de muestra, los riesgos de LDA, QDA y Naive Bayes se asemejan bastante al error de Bayes.

En la Figura 3 podemos ver las brechas existentes entre el riesgo de cada clasificador y el riesgo de Bayes, para cada tamaño de muestra. Como esperábamos, en este escenario (fácil agrupamiento) los riesgos de LDA, QDA y Naive Bayes se acercan bastante al riesgo de Bayes. Sin embargo, los resultados sugieren que en la práctica QDA y Naive Bayes no solo alcanzan un rendimiento similar, sino que incluso se aproximan más al óptimo que LDA, lo cual podría deberse a la variabilidad de la simulación y a las condiciones específicas de los parámetros elegidos.

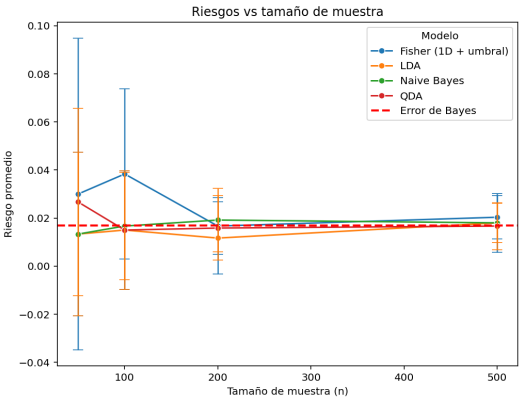


Figura 2: Gráfica de riesgos de clasificadores vs Bayes. Caso fácil.

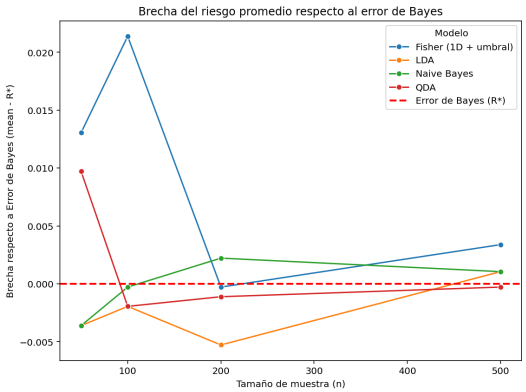


Figura 3: Gráfica de brechas entre riesgos de clasificadores vs Bayes. Caso fácil.

En las Figuras 4 y 5 mostramos los riesgos promedios y las brechas entre estos riesgos y el riesgo de Bayes del clasificador k-NN para cada tamaño muestral considerado antes y para distintos valores de k. En este casom los resultados mostraron mayor sensibilidad al valor de k. Para valores pequeños, el clasificador exhibe alta varianza y errores más altos, mientras que para valores más grandes los riesgos se estabilizan y se acercan al error de Bayes. Aunque las gráficas de brechas parecen mostrar grandes oscilaciones, la escala es en realidad muy pequeña, lo que indica que k-NN también es competitivo en escenarios sencillos.

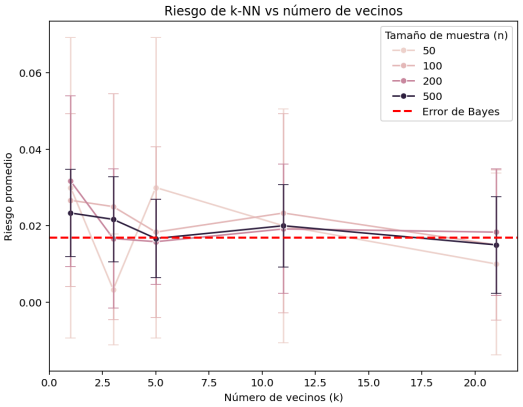


Figura 4: Gráfica de riesgos de clasificadores vs Bayes. Caso fácil.

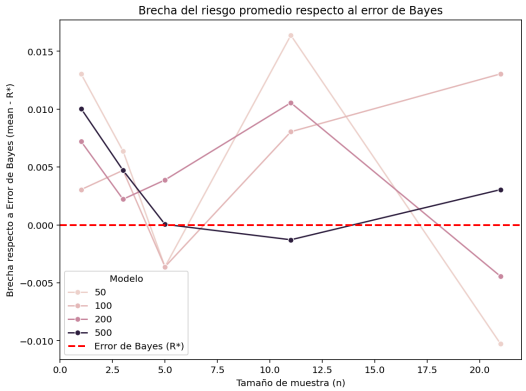


Figura 5: Gráfica de brechas entre riesgos de clasificadores vs Bayes. Caso fácil.

En la Tabla 1 presentamos las tablas con los promedios $\pm$ las desviaciones estándar de cada clasificador para cada tamaño de muestra. En la Tabla 6a notamos que en su mayoría los valores se acercan mucho al error óptimo, pudiéndose deber al hecho de que la separación entre clases es de un grado sencillo y todos los clasificadores hacen su trabajo de buena manera. En la Tabla 6b podemos ver justamente lo que apreciamos en la Figura 4, que conforme crece el número de vecinos considerados, el riesgo del clasificador se acerca más al riesgo de Bayes.

Modelo	mean ± std	n
Fisher (1D + umbral)	0.030 ± 0.064	50
Fisher (1D + umbral)	0.038 ± 0.035	100
Fisher (1D + umbral)	0.016 ± 0.011	200
Fisher (1D + umbral)	0.020 ± 0.009	500
LDA	0.013 ± 0.033	50
LDA	0.015 ± 0.024	100
LDA	0.011 ± 0.015	200
LDA	0.018 ± 0.012	500
Naive Bayes	0.013 ± 0.033	50
Naive Bayes	0.016 ± 0.022	100
Naive Bayes	0.019 ± 0.013	200
Naive Bayes	0.018 ± 0.008	500
QDA	0.026 ± 0.038	50
QDA	0.015 ± 0.024	100
QDA	0.015 ± 0.013	200
QDA	0.016 ± 0.009	500

(a) Resultados de los modelos con sus medias, desviaciones estándar y tamaño de muestra.

Modelo	n	Media ± std
k-NN (k=1)	50	0.030 ± 0.039
k-NN (k=1)	100	0.020 ± 0.026
k-NN (k=1)	200	0.024 ± 0.023
k-NN (k=1)	500	0.027 ± 0.011
k-NN (k=3)	50	0.023 ± 0.043
k-NN (k=3)	100	0.021 ± 0.024
k-NN (k=3)	200	0.019 ± 0.024
k-NN (k=3)	500	0.021 ± 0.008
k-NN (k=5)	50	0.013 ± 0.033
k-NN (k=5)	100	0.013 ± 0.019
k-NN (k=5)	200	0.020 ± 0.019
k-NN (k=5)	500	0.017 ± 0.012
k-NN (k=11)	50	0.033 ± 0.044
k-NN (k=11)	100	0.025 ± 0.023
k-NN (k=11)	200	0.027 ± 0.019
k-NN (k=11)	500	0.015 ± 0.008
k-NN (k=21)	50	0.006 ± 0.020
k-NN (k=21)	100	0.030 ± 0.029
k-NN (k=21)	200	0.012 ± 0.013
k-NN (k=21)	500	0.020 ± 0.009

(b) Resultados del clasificador k-NN para distintos valores de k y tamaños de muestra.

Tabla 1: Comparación de promedios de riesgos por modelo.

En la Figura 6 observamos la gráfica de los riesgos de los clasificadores de Naive Bayes, LDA, QDA y Fisher contra el error óptimo de Bayes (el cual es aproximadamente 0.144), en el caso en que los datos tienen un grado de separación intermedio. En ella podemos observar como los clasificadores LDA y QDA parecen ser los más cercanos al error de Bayes.

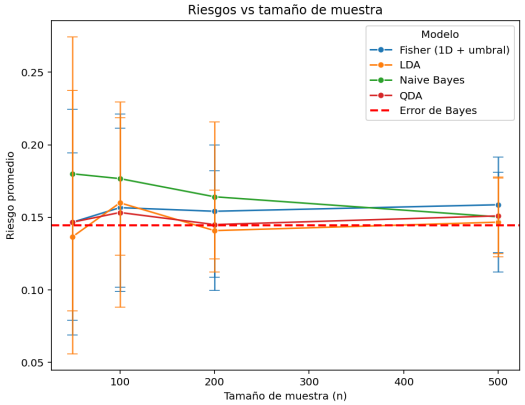


Figura 6: Gráfica de riesgos de clasificadores vs Bayes. Caso medio.

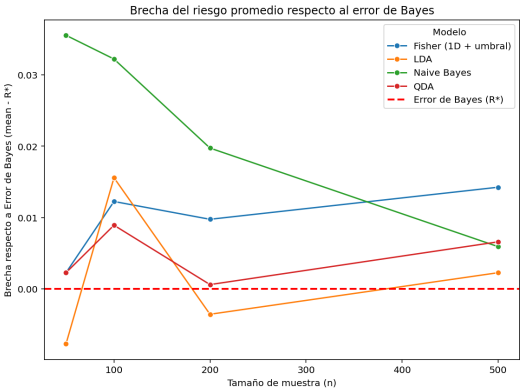


Figura 7: Gráfica de brechas entre riesgos de clasificadores vs Bayes. Caso medio.

En la Figura 7 podemos ver las brechas existentes entre el riesgo de cada clasificador y el riesgo de Bayes, para cada tamaño de muestra. En esta podemos ver como Naive Bayes es que más difiere del error de Bayes y como QDA parece ser más parecido en los primeros tamaños de muestra, mientras LDA es el más cercano en los tamaños más grandes.

En las Figuras 8 y 9 mostramos los riesgos promedios y las brechas entre estos riesgos y el riesgo de Bayes del clasificador k-NN para cada tamaño muestral considerado antes y para distintos valores de k. Vemos como para k pequeños la diferencia de este clasificador con el de Bayes es relativamente considerable, pero para los valores más grandes las líneas se acercan más a la correspondiente al error de Bayes. En la Figura 9 vemos como para la muestra más grande, la brecha entre el riesgo del clasificador y el óptimo va tendiendo a 0 conforme crece el valor de k, mientras que para los otros tamaños de muestra, las brechas parecen oscilar mucho.

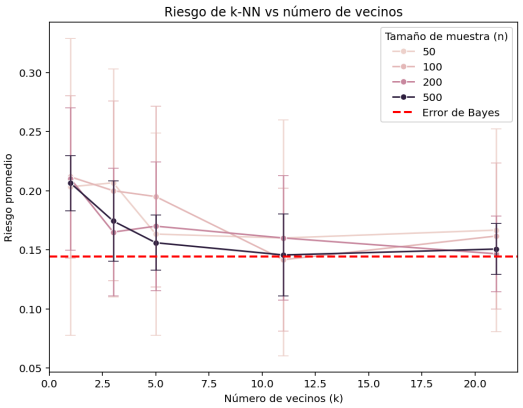


Figura 8: Gráfica de riesgos de k-NN vs Bayes. Caso medio.

Figura 9: Gráfica de brechas entre riesgos de k-NN vs Bayes. Caso medio.

Modelo	**mean ± std**	**n**
Fisher (1D + umbral)	0.147 ± 0.078	50
Fisher (1D + umbral)	0.157 ± 0.055	100
Fisher (1D + umbral)	0.154 ± 0.046	200
Fisher (1D + umbral)	0.159 ± 0.033	500
LDA	0.137 ± 0.058	50
LDA	0.160 ± 0.061	100
LDA	0.141 ± 0.041	200
LDA	0.147 ± 0.034	500
Naive Bayes	0.180 ± 0.095	50
Naive Bayes	0.177 ± 0.053	100
Naive Bayes	0.164 ± 0.052	200
Naive Bayes	0.150 ± 0.028	500
QDA	0.147 ± 0.091	50
QDA	0.153 ± 0.065	100
QDA	0.145 ± 0.024	200
QDA	0.151 ± 0.026	500

(a) Resultados de los modelos con sus medias, desviaciones estándar y tamaño de muestra.

Modelo	**n**	**Media ± std**
k-NN (k=1)	50	0.200 ± 0.112
k-NN (k=1)	100	0.197 ± 0.074
k-NN (k=1)	200	0.202 ± 0.056
k-NN (k=1)	500	0.212 ± 0.023
k-NN (k=3)	50	0.150 ± 0.101
k-NN (k=3)	100	0.200 ± 0.076
k-NN (k=3)	200	0.198 ± 0.051
k-NN (k=3)	500	0.174 ± 0.025
k-NN (k=5)	50	0.210 ± 0.074
k-NN (k=5)	100	0.178 ± 0.075
k-NN (k=5)	200	0.176 ± 0.057
k-NN (k=5)	500	0.168 ± 0.036
k-NN (k=11)	50	0.150 ± 0.070
k-NN (k=11)	100	0.197 ± 0.070
k-NN (k=11)	200	0.150 ± 0.044
k-NN (k=11)	500	0.150 ± 0.031
k-NN (k=21)	50	0.167 ± 0.095
k-NN (k=21)	100	0.147 ± 0.051
k-NN (k=21)	200	0.146 ± 0.042
k-NN (k=21)	500	0.146 ± 0.035

(b) Resultados del clasificador k-NN para distintos valores de k y tamaños de muestra.

Tabla 2: Comparación de promedios de riesgos por modelo.

En la Tabla 2 presentamos las tablas con los promedios ± las desviaciones estándar de cada clasificador para cada tamaño de muestra, para ver en resultados numéricos lo que ya apreciamos

visualmente.

En la Figura 10 observamos la gráfica de los riesgos de los clasificadores de Naive Bayes, LDA, QDA y Fisher contra el error óptimo de Bayes (el cual es aproximadamente 0.3618), en el caso en que los datos tienen un grado de separación Difícil. En ella podemos observar como el clasificador de Fisher parece ser el que da riesgos más alejados al óptimo.

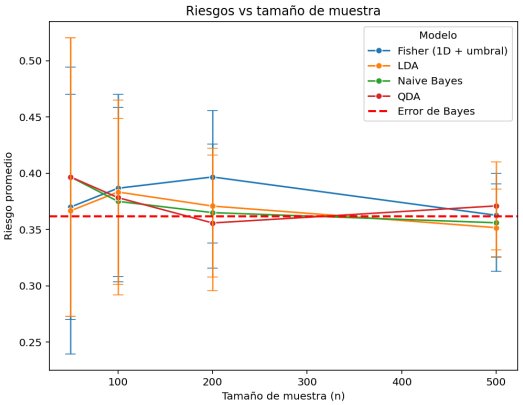


Figura 10: Gráfica de riesgos de clasificadores vs Bayes. Caso difícil.

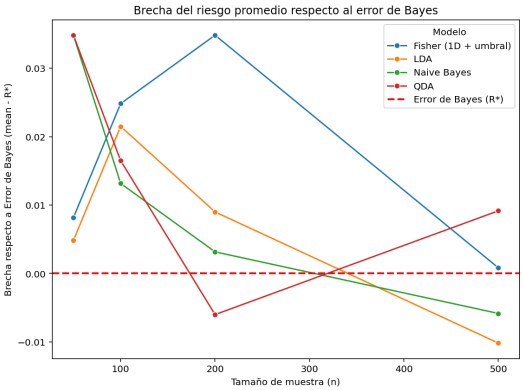


Figura 11: Gráfica de brechas entre riesgos de clasificadores vs Bayes. Caso difícil.

En la Figura 11 podemos ver las brechas existentes entre el riesgo de cada clasificador y el riesgo de Bayes, para cada tamaño de muestra. En esta podemos ver como los riesgos de LDA parecen ser los que menos difieren del óptimo en la mayoría de los tamaños muestrales.

En las Figuras 12 y 13 mostramos los riesgos promedios y las brechas entre estos riesgos y el riesgo de Bayes del clasificador k-NN para cada tamaño muestral considerado antes y para distintos valores de k. Vemos como para todos los valores de k el nivel de riesgo no varía de manera muy significativa para los dos tamaños de muestra más grandes. En la Figura 9 vemos como para la muestra de 500 datos, la brecha entre el riesgo del clasificador y el óptimo va decreciendo aunque no termina de pegarse mucho al 0.

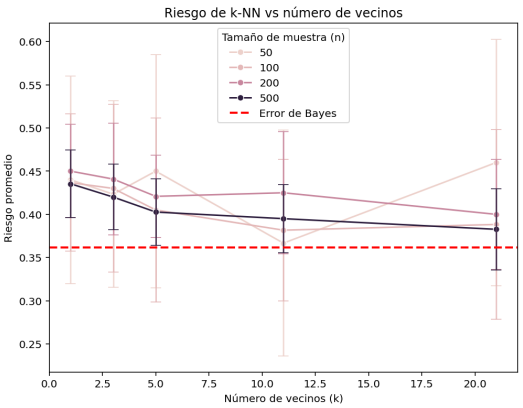


Figura 12: Gráfica de riesgos de k-NN vs Bayes. Caso difícil.

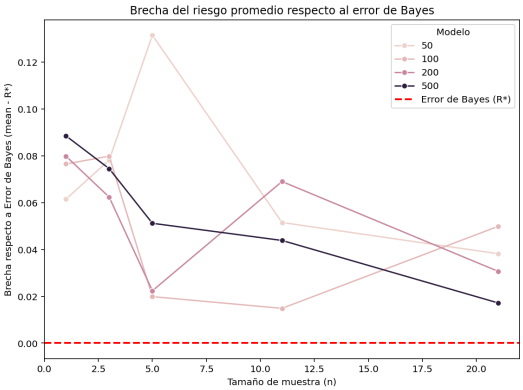


Figura 13: Gráfica de brechas entre riesgos de k-NN vs Bayes. Caso difícil.

En la Tabla 3 presentamos las tablas con los promedios $\pm$ las desviaciones estándar de cada clasificador para cada tamaño de muestra, para ver en resultados numéricos lo que ya apreciamos visualmente.

Modelo	mean ± std	n
Fisher (1D + umbral)	0.376 ± 0.115	50
Fisher (1D + umbral)	0.321 ± 0.123	100
Fisher (1D + umbral)	0.373 ± 0.055	200
Fisher (1D + umbral)	0.376 ± 0.045	500
LDA	0.360 ± 0.127	50
LDA	0.378 ± 0.084	100
LDA	0.385 ± 0.071	200
LDA	0.360 ± 0.044	500
Naive Bayes	0.440 ± 0.112	50
Naive Bayes	0.375 ± 0.066	100
Naive Bayes	0.375 ± 0.065	200
Naive Bayes	0.360 ± 0.037	500
QDA	0.373 ± 0.110	50
QDA	0.396 ± 0.094	100
QDA	0.373 ± 0.057	200
QDA	0.353 ± 0.037	500

(a) Resultados de los modelos con sus medias, desviaciones estándar y tamaño de muestra.

Modelo	n	Media ± std
k-NN (k=1)	50	0.423 ± 0.130
k-NN (k=1)	100	0.438 ± 0.082
k-NN (k=1)	200	0.442 ± 0.060
k-NN (k=1)	500	0.450 ± 0.043
k-NN (k=3)	50	0.440 ± 0.131
k-NN (k=3)	100	0.442 ± 0.079
k-NN (k=3)	200	0.424 ± 0.049
k-NN (k=3)	500	0.436 ± 0.052
k-NN (k=5)	50	0.493 ± 0.114
k-NN (k=5)	100	0.382 ± 0.088
k-NN (k=5)	200	0.384 ± 0.064
k-NN (k=5)	500	0.413 ± 0.040
k-NN (k=11)	50	0.413 ± 0.147
k-NN (k=11)	100	0.377 ± 0.098
k-NN (k=11)	200	0.431 ± 0.057
k-NN (k=11)	500	0.406 ± 0.040
k-NN (k=21)	50	0.400 ± 0.099
k-NN (k=21)	100	0.412 ± 0.107
k-NN (k=21)	200	0.393 ± 0.050
k-NN (k=21)	500	0.379 ± 0.044

(b) Resultados del clasificador k-NN para distintos valores de k y tamaños de muestra.

Tabla 3: Comparación de promedios de riesgos por modelo.

Una vez analizados los tres escenarios planteados (fácil, intermedio y difícil) observamos un patrón claro: conforme aumentaba la dificultad de separación entre clases, los riesgos de los clasificadores tendieron a alejarse más del riesgo óptimo de Bayes. Aunque las gráficas parecen mostrar comportamientos muy similares entre métodos, la diferencia real se manifestaba en la escala de los riesgos, la cual crecía a medida que el problema se volvía más complejo.

En el caso de LDA, esperábamos, basado en la teoría, que coincidiera con el clasificador de Bayes (pues tenemos covarianzas iguales). Sin embargo, los resultados no siempre reflejaron esta coincidencia, lo cual pensamos que se atribuye a la variabilidad propia de la simulación y a los tamaños muestrales considerados. Aun así, LDA fue consistente apareciendo siempre como uno de los métodos más cercanos al óptimo en los tres niveles de dificultad. Por su parte, el clasificador k-NN mostró un comportamiento dependiente del número de vecinos considerados. Al aumentar el número de vecinos, los riesgos se estabilizaron y se acercaron más al riesgo de Bayes, mostrando un mejor desempeño relativo.
En resumen, dados los resultados podemos concluir que la complejidad del problema tiene un impacto directo en la capacidad de los clasificadores de aproximarse al óptimo, y que tanto LDA como k-NN (con un número adecuado de vecinos) resultan ser opciones competitivas a lo largo de los distintos escenarios evaluados.

- **Covarianzas distintas.**

  En este caso, consideramos las matrices de covarianzas $\Sigma_0 \neq \Sigma_1$ y las medias μ_0, μ_1 de manera que el la separación de las clases presentara un nivel intermedio. Los parámetros que utilizamos para las simulaciones en este escenario fueron:

$$\Sigma_0 = I_2,\ \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},\quad \mu_0 = (1.5, 1.5),\ \mu_1 = (3, 3).$$

En la Figura 14 observamos la gráfica de los riesgos de los clasificadores de Naive Bayes, LDA, QDA y Fisher contra el error óptimo de Bayes (el cual es aproximadamente 0.1611). En ella podemos observar como después de un tamaño muestral de 200 todos los clasificadores (salvo Fisher) se van acercando más al error de Bayes.

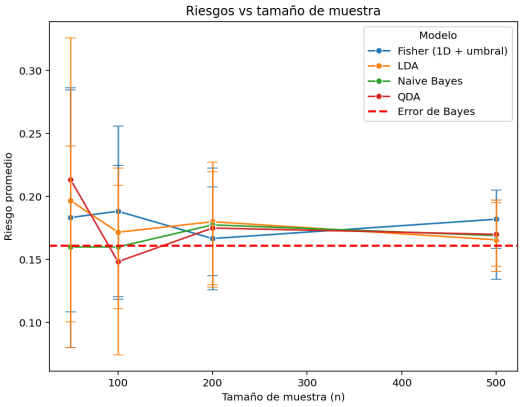


Figura 14: Gráfica de riesgos de clasificadores vs Bayes.

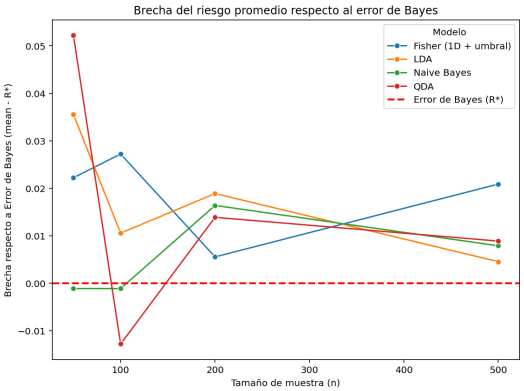


Figura 15: Gráfica de brechas entre riesgos de clasificadores vs Bayes.

En la Figura 15 vemos las brechas existentes entre el riesgo de cada clasificador y el riesgo de Bayes y notamos como el riesgo de QDA es nunca es el que difiere menos de Bayes, lo cual es extraño pues, teóricamente, cuando las covarianzas son distintas se espera que QDA conincida con Bayes. Aunque bien, esto puede ser debido a la variabilidad en la simulación o los específicos parámetros que escogimos.

En las Figuras 16 y 17 mostramos los promedios y las brechas entre los riesgos y el riesgo de Bayes del clasificador k-NN para cada tamaño muestral considerado antes y para distintos valores de k. Observamos como para todas las k los riesgos se están acercando al óptimo y las brechas entre ellos no son tan grandes considerando la escala a la que se encuentran.

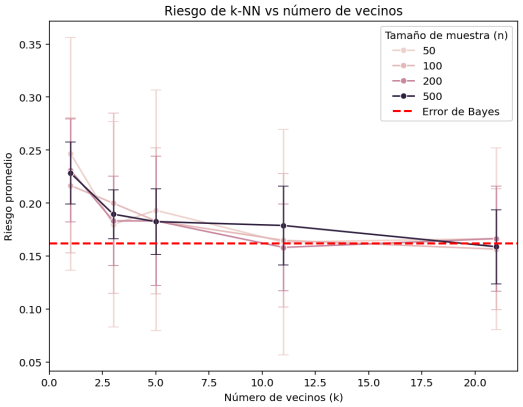


Figura 16: Gráfica de riesgos de k-NN vs Bayes.

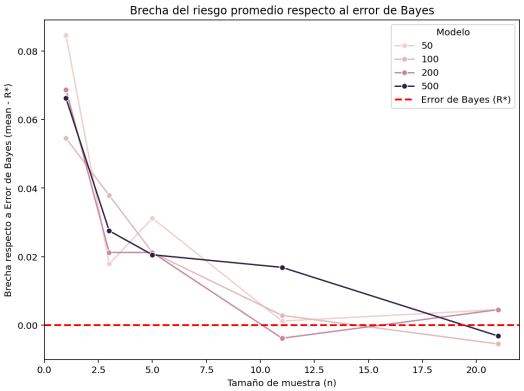


Figura 17: Gráfica de brechas entre riesgos de k-NN vs Bayes.

En la Tabla 4 presentamos las tablas con los promedios $\pm$ las desviaciones estándar de cada clasificador para cada tamaño de muestra, para ver en resultados numéricos lo que ya apreciamos visualmente.

Modelo	n	mean ± std
Fisher (1D + umbral)	50	0.183 ± 0.103
Fisher (1D + umbral)	100	0.188 ± 0.068
Fisher (1D + umbral)	200	0.167 ± 0.041
Fisher (1D + umbral)	500	0.182 ± 0.023
LDA	50	0.197 ± 0.088
LDA	100	0.172 ± 0.053
LDA	200	0.180 ± 0.043
LDA	500	0.166 ± 0.032
Naive Bayes	50	0.160 ± 0.080
Naive Bayes	100	0.160 ± 0.049
Naive Bayes	200	0.178 ± 0.050
Naive Bayes	500	0.169 ± 0.028
QDA	50	0.213 ± 0.113
QDA	100	0.148 ± 0.074
QDA	200	0.175 ± 0.045
QDA	500	0.170 ± 0.025

(a) Resultados de los modelos con sus medias, desviaciones estándar y tamaño de muestra.

Modelo	n	Media ± std
k-NN (k=1)	50	0.247 ± 0.110
k-NN (k=1)	100	0.217 ± 0.064
k-NN (k=1)	200	0.231 ± 0.049
k-NN (k=1)	500	0.228 ± 0.029
k-NN (k=3)	50	0.180 ± 0.097
k-NN (k=3)	100	0.200 ± 0.085
k-NN (k=3)	200	0.183 ± 0.042
k-NN (k=3)	500	0.190 ± 0.023
k-NN (k=5)	50	0.193 ± 0.113
k-NN (k=5)	100	0.183 ± 0.069
k-NN (k=5)	200	0.183 ± 0.061
k-NN (k=5)	500	0.183 ± 0.031
k-NN (k=11)	50	0.163 ± 0.106
k-NN (k=11)	100	0.165 ± 0.063
k-NN (k=11)	200	0.158 ± 0.041
k-NN (k=11)	500	0.179 ± 0.037
k-NN (k=21)	50	0.167 ± 0.086
k-NN (k=21)	100	0.157 ± 0.057
k-NN (k=21)	200	0.167 ± 0.049
k-NN (k=21)	500	0.159 ± 0.035

(b) Resultados del clasificador k-NN para distintos valores de k y tamaños de muestra.

Tabla 4: Comparación de promedios de riesgos por modelo.

Caso desbalanceado.

- **Covarianzas iguales.**

  En este caso, consideramos las distribuciones a priori $\pi_0 = 0.8$ y $\pi_1 = 0.2$, para tener un escenario desbalanceado, y la identidad como matriz de covarianzas. Las medias consideradas son iguales a las del caso balanceado con covarianzas distintas.

  En la Figura 18 observamos la gráfica de los riesgos de los clasificadores de Naive Bayes, LDA, QDA y Fisher contra el error óptimo de Bayes (el cual es aproximadamente 0.1029). Vemos como los riesgos de Naive Bayes y LDA son sumamente parecidos entre los tamaños de muestra más grandes, y como es Naive Bayes el que parece coinicidir más con el óptimo a tráves de todo el dominio de tamaños muestrales. En la Figura 19 podemos ver las brechas existentes entre el riesgo que nos muestran justo los comportamientos que acabamos de describir.

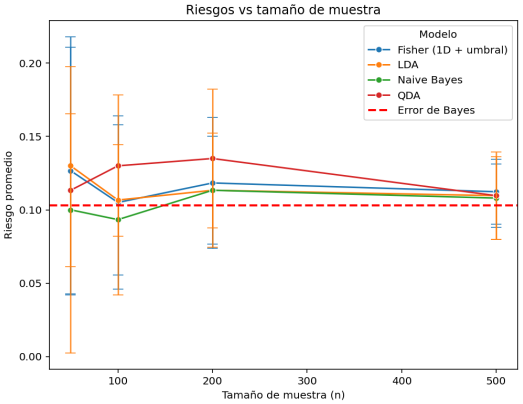


Figura 18: Gráfica de riesgos de clasificadores vs Bayes.

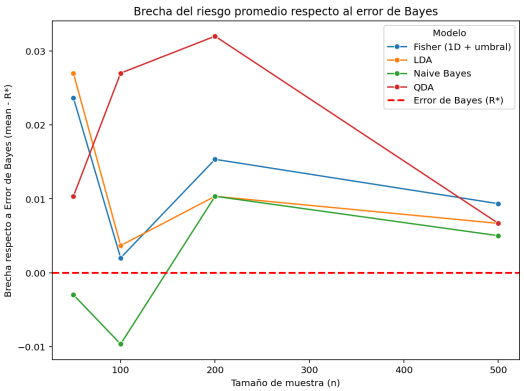


Figura 19: Gráfica de brechas entre riesgos de clasificadores vs Bayes.

En las Figuras 20 y 21 mostramos los riesgos promedios y las brechas entre estos riesgos y el riesgo de Bayes del clasificador k-NN para cada tamaño muestral considerado antes y para distintos valores de k. Observamos como para los tamaños de muestra más pequeños los riesgos van aumentando conforme aumenta también el valor número de vecinos considerados y en la gráfica de brechas justamente observamos cómo es que estas van creciendo significativamente, por lo que parece ser que los riesgos coinciden más con el óptimo para valores pequeños de k en el caso de las muestras pequeñas y para valores más grandes de k en el caso de las muestras más grandes, particularmente, para $k = 21$ notamos como la brecha entre el riesgo promedio del clasificador y el de Bayes es casi nula.

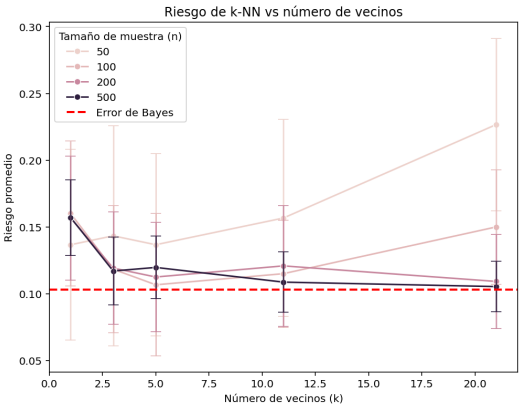


Figura 20: Gráfica de riesgos de clasificadores vs Bayes.

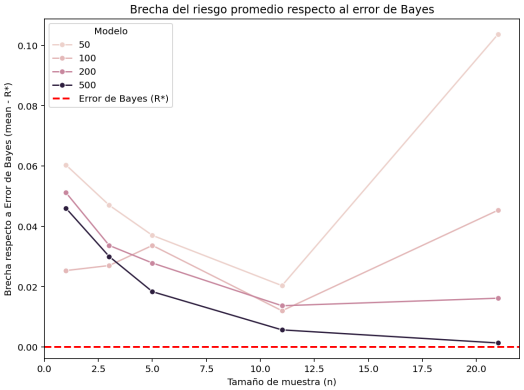


Figura 21: Gráfica de brechas entre riesgos de clasificadores vs Bayes.

En la Tabla 6 presentamos las tablas con los promedios $\pm$ las desviaciones estándar de cada clasificador para cada tamaño de muestra, para ver en resultados numéricos lo que ya apreciamos visualmente.

Modelo	n	mean ± std
Fisher (1D + umbral)	50	0.127 ± 0.084
Fisher (1D + umbral)	100	0.105 ± 0.059
Fisher (1D + umbral)	200	0.118 ± 0.045
Fisher (1D + umbral)	500	0.112 ± 0.022
LDA	50	0.130 ± 0.088
LDA	100	0.107 ± 0.051
LDA	200	0.113 ± 0.037
LDA	500	0.110 ± 0.022
Naive Bayes	50	0.100 ± 0.098
Naive Bayes	100	0.093 ± 0.051
Naive Bayes	200	0.113 ± 0.039
Naive Bayes	500	0.108 ± 0.028
QDA	50	0.113 ± 0.052
QDA	100	0.130 ± 0.048
QDA	200	0.135 ± 0.047
QDA	500	0.110 ± 0.030

(a) Resultados de los modelos con sus medias, desviaciones estándar y tamaño de muestra.

Modelo	n	Media ± std
k-NN (k=1)	50	0.163 ± 0.077
k-NN (k=1)	100	0.128 ± 0.037
k-NN (k=1)	200	0.154 ± 0.058
k-NN (k=1)	500	0.149 ± 0.029
k-NN (k=3)	50	0.150 ± 0.081
k-NN (k=3)	100	0.130 ± 0.052
k-NN (k=3)	200	0.137 ± 0.039
k-NN (k=3)	500	0.133 ± 0.026
k-NN (k=5)	50	0.140 ± 0.073
k-NN (k=5)	100	0.137 ± 0.048
k-NN (k=5)	200	0.131 ± 0.038
k-NN (k=5)	500	0.121 ± 0.025
k-NN (k=11)	50	0.123 ± 0.057
k-NN (k=11)	100	0.115 ± 0.049
k-NN (k=11)	200	0.117 ± 0.028
k-NN (k=11)	500	0.109 ± 0.028
k-NN (k=21)	50	0.207 ± 0.063
k-NN (k=21)	100	0.148 ± 0.044
k-NN (k=21)	200	0.119 ± 0.035
k-NN (k=21)	500	0.104 ± 0.020

(b) Resultados del clasificador k-NN para distintos valores de k y tamaños de muestra.

Tabla 5: Comparación de promedios de riesgos por modelo.

- **Covarianzas distintas.**

  En este caso, consideramos las matrices de covarianzas $\Sigma_0 \neq \Sigma_1$ y las medias μ_0, μ_1 mencionadas en el caso balanceado con covarianzas disntintas.

  En la Figura 22 observamos la gráfica de los riesgos de los clasificadores de Naive Bayes, LDA, QDA y Fisher contra el error óptimo de Bayes (el cual es aproximadamente 0.105) y en la Figura 23 podemos ver las brechas existentes entre el riesgo. Lo apreciable en estas figuras es ver que los riesgos promedio de QDA son los más próximos al óptimo.

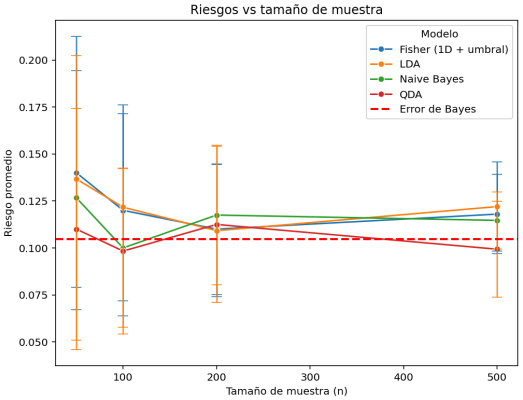


Figura 22: Gráfica de riesgos de clasificadores vs Bayes.

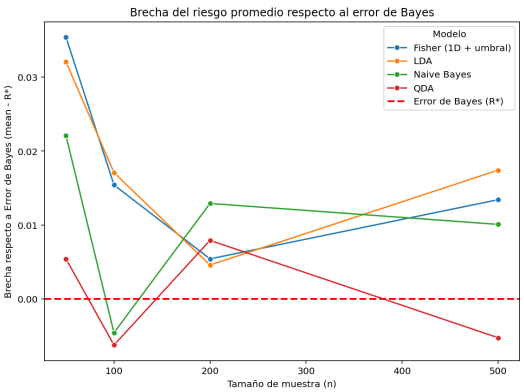


Figura 23: Gráfica de brechas entre riesgos de clasificadores vs Bayes.

En las Figuras 24 y 25 vemos los riesgos promedios y las brechas entre estos riesgos y el riesgo de Bayes del clasificador k-NN para cada tamaño muestral considerado antes y para distintos

valores de k. Notamos en ambas gráficas que los riesgos para casi todos los casos no parecen coincidir mucho con el óptimo.

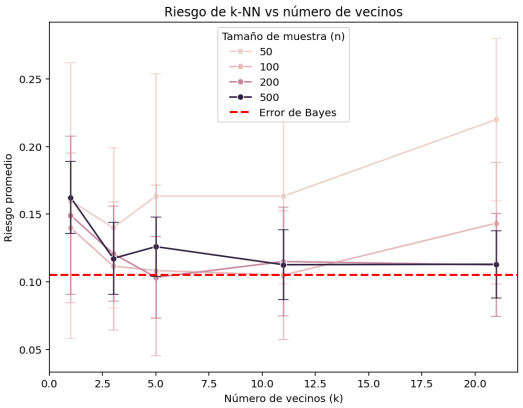


Figura 24: Gráfica de riesgos de clasificadores vs Bayes.

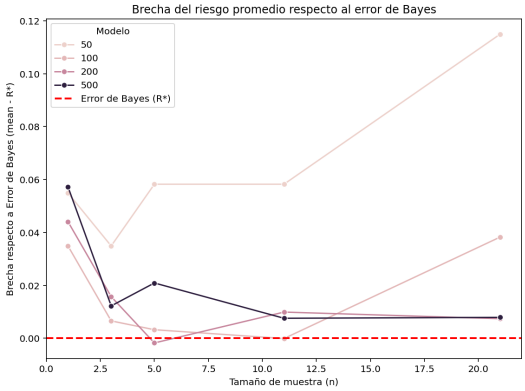


Figura 25: Gráfica de brechas entre riesgos de clasificadores vs Bayes.

En la Tabla 6 presentamos las tablas con los promedios ± las desviaciones estándar de cada clasificador para cada tamaño de muestra, para ver en resultados numéricos lo que ya apreciamos visualmente. Notamos que en la Tabla 6b la mayoría de los valores de las medias se ven considerablemente mayores al valor óptimo, y son los valores finales los que parecen acercarse más, tal como mirábamos en la gráfica.

Modelo	**n**	**mean ± std**
Fisher (1D + umbral)	50	0.127 ± 0.084
Fisher (1D + umbral)	100	0.105 ± 0.059
Fisher (1D + umbral)	200	0.118 ± 0.045
Fisher (1D + umbral)	500	0.112 ± 0.022
LDA	50	0.130 ± 0.088
LDA	100	0.107 ± 0.051
LDA	200	0.113 ± 0.037
LDA	500	0.110 ± 0.022
Naive Bayes	50	0.100 ± 0.098
Naive Bayes	100	0.093 ± 0.051
Naive Bayes	200	0.113 ± 0.039
Naive Bayes	500	0.108 ± 0.028
QDA	50	0.113 ± 0.052
QDA	100	0.130 ± 0.048
QDA	200	0.135 ± 0.047
QDA	500	0.110 ± 0.030

(a) Resultados de los modelos con sus medias, desviaciones estándar y tamaño de muestra.

Modelo	**n**	**Media ± std**
k-NN (k=1)	50	0.163 ± 0.077
k-NN (k=1)	100	0.128 ± 0.037
k-NN (k=1)	200	0.154 ± 0.058
k-NN (k=1)	500	0.149 ± 0.029
k-NN (k=3)	50	0.150 ± 0.081
k-NN (k=3)	100	0.130 ± 0.052
k-NN (k=3)	200	0.137 ± 0.039
k-NN (k=3)	500	0.133 ± 0.026
k-NN (k=5)	50	0.140 ± 0.073
k-NN (k=5)	100	0.137 ± 0.048
k-NN (k=5)	200	0.131 ± 0.038
k-NN (k=5)	500	0.121 ± 0.025
k-NN (k=11)	50	0.123 ± 0.057
k-NN (k=11)	100	0.115 ± 0.049
k-NN (k=11)	200	0.117 ± 0.028
k-NN (k=11)	500	0.109 ± 0.028
k-NN (k=21)	50	0.207 ± 0.063
k-NN (k=21)	100	0.148 ± 0.044
k-NN (k=21)	200	0.119 ± 0.035
k-NN (k=21)	500	0.104 ± 0.020

(b) Resultados del clasificador k-NN para distintos valores de k y tamaños de muestra.

Tabla 6: Comparación de promedios de riesgos por modelo.

4. Riesgo verdadero vs. validación

4.1. LDA

4.1.1. Covarianzas iguales, $\pi_0 = 0.5$ (balance).

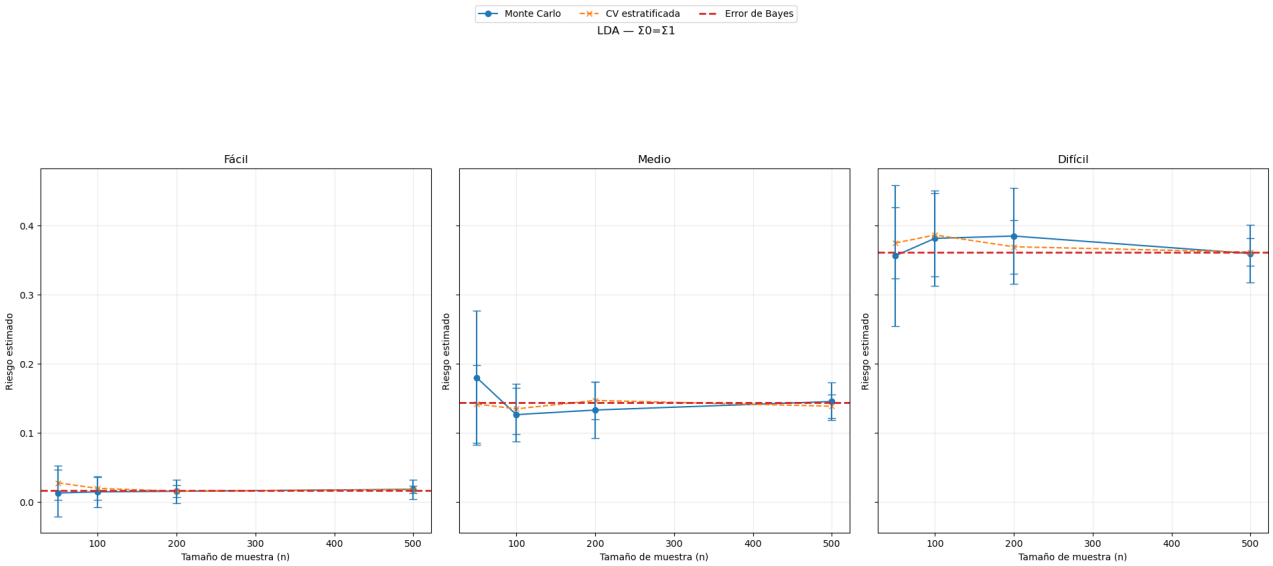


Figura 26: LDA: Riesgo verdadero vs validación para diferentes dificultades, en el caso de covarianzas iguales y con balance

En las tres dificultades se observa la convergencia esperada de LDA hacia R^* (Error de Bayes). En el caso fácil la brecha $L_{\text{MC}} - R^*$ es prácticamente nula desde $n \approx 100$; en medio aún hay una pequeña sobreestimación para n pequeños que se reduce con el tamaño muestral; en difícil la curva también desciende hacia R^* aunque con mayor varianza inicial debido a la mayor superposición de clases. Las curvas de L_{CV} quedan muy cercanas a L_{MC}.

4.1.2. Covarianzas iguales, $\pi_0 = 0.8$ (desbalance).

El cambio en los priors en la Figura 27 desplaza el umbral favoreciendo la clase mayoritaria. El nivel de R^* cambia acorde al desbalance, pero la jerarquía se mantiene: LDA sigue aproximando muy bien a Bayes en fácil y medio, y en difícil la brecha persiste para n pequeños y se atenúa al crecer n. La estratificación en la validación estabiliza la estimación en presencia del desbalance y produce curvas L_{CV} casi superpuestas con L_{MC}.

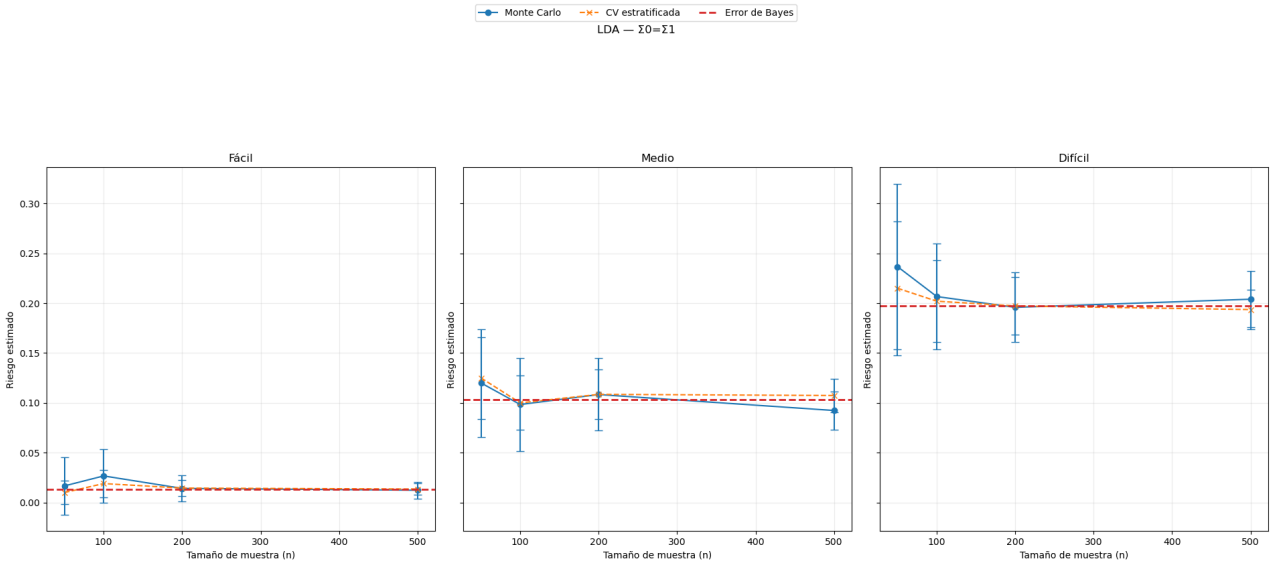


Figura 27: LDA: Riesgo verdadero vs validación para diferentes dificultades, en el caso de covarianzas iguales y con desbalance

4.1.3. Covarianzas distintas, $\pi_0 = 0.5$ (balance).

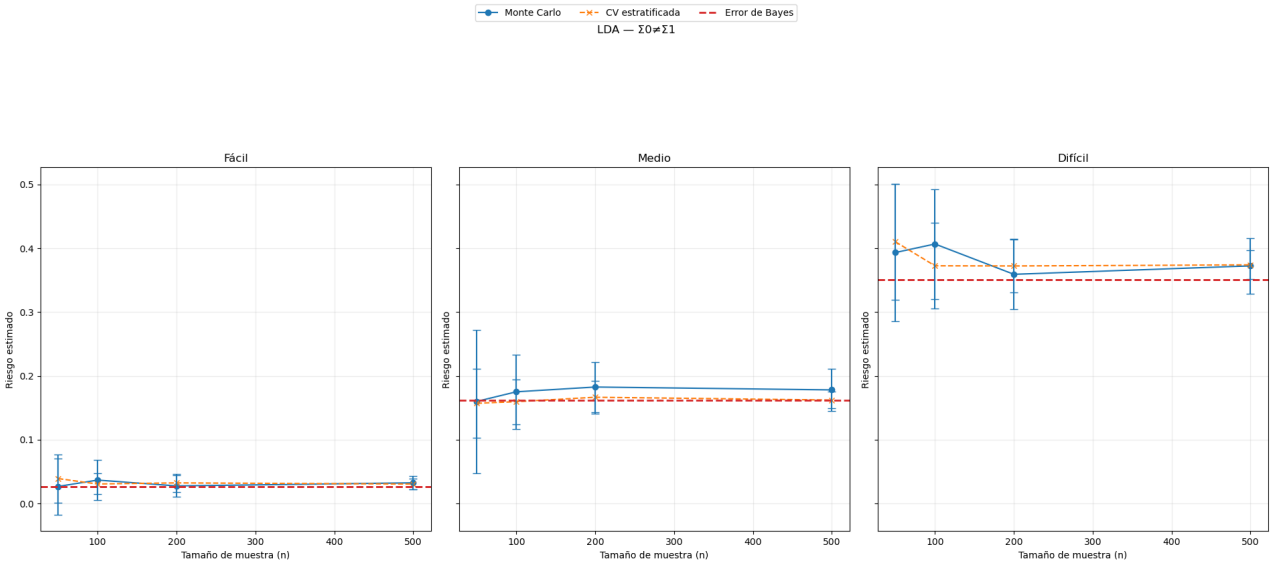


Figura 28: LDA: Riesgo verdadero vs validación para diferentes dificultades, en el caso de covarianzas distintas y con balance

Aquí LDA está mal especificado: la frontera óptima de Bayes es cuadrática. Las gráficas muestran una brecha asintótica positiva entre LDA y R^* que no desaparece al aumentar n; el modelo converge hacia su mejor desempeño lineal pero por encima del óptimo. El efecto es más notorio conforme aumenta la dificultad (mayor solapamiento), donde la diferencia $L_{\text{MC}} - R^*$ se mantiene visible aun con $n = 500$. L_{CV} reproduce el mismo patrón con ligera elevación respecto a L_{MC} por el sesgo propio de la validación cruzada.

4.1.4. Covarianzas distintas, $\pi_0 = 0.8$ (desbalance).

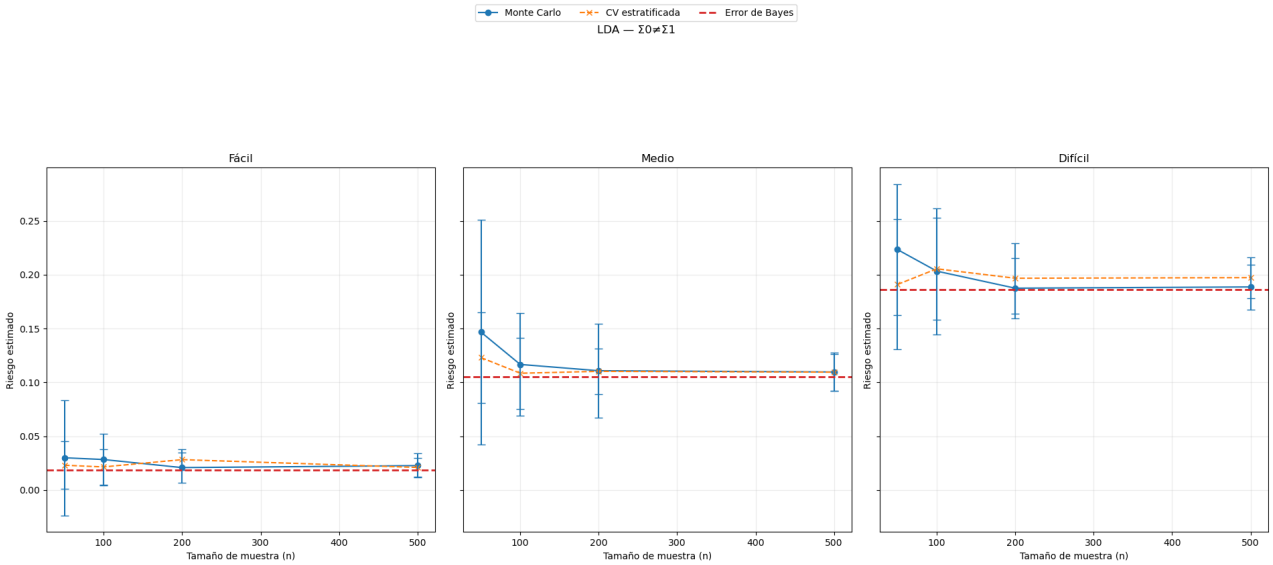


Figura 29: LDA: Riesgo verdadero vs validación para diferentes dificultades, en el caso de covarianzas distintas y con desbalance

Se combinan ambos efectos: desbalance y mala especificación. El prior asimétrico ajusta el umbral y puede reducir la pérdida esperada al penalizar menos los errores sobre la clase minoritaria, pero la restricción de linealidad impide alcanzar R^*. En fácil la separación geométrica hace que la brecha sea pequeña; en medio y sobre todo en difícil la brecha persiste con n grande, evidenciando el límite estructural de LDA bajo $\Sigma_0 \neq \Sigma_1$. De nuevo, L_{CV} es coherente con L_{MC} y presenta menor variabilidad para n pequeños gracias a la estratificación.

4.2. QDA

4.2.1. Covarianzas iguales, $\pi_0 = 0.5$ (balance).

En este escenario la frontera de Bayes es lineal (LDA es el clasificador óptimo). QDA, al estimar dos matrices de covarianza separadas, está sobreparametrizado: paga un mayor costo de varianza sin obtener ventaja de sesgo. Las curvas en la Figura 30 muestran exactamente este comportamiento:

- **Fácil:** L_{MC} y L_{CV} descienden rápidamente y se alinean con R^* desde $n \approx 100$. La brecha respecto a R^* es prácticamente nula para n moderado, aunque la dispersión inicial es algo mayor que la de un clasificador lineal.
- **Medio:** persiste una pequeña diferencia para tamaños muestrales chicos, consecuencia de la varianza en la estimación de las dos Σ_k. A medida que crece n, QDA se aproxima a R^*.
- **Difícil:** la superposición entre clases amplifica la varianza: las barras de error son mayores y L_{CV} queda levemente por encima de L_{MC} (sesgo propio de la CV). Aun así, ambas curvas tienden hacia R^* al aumentar n.

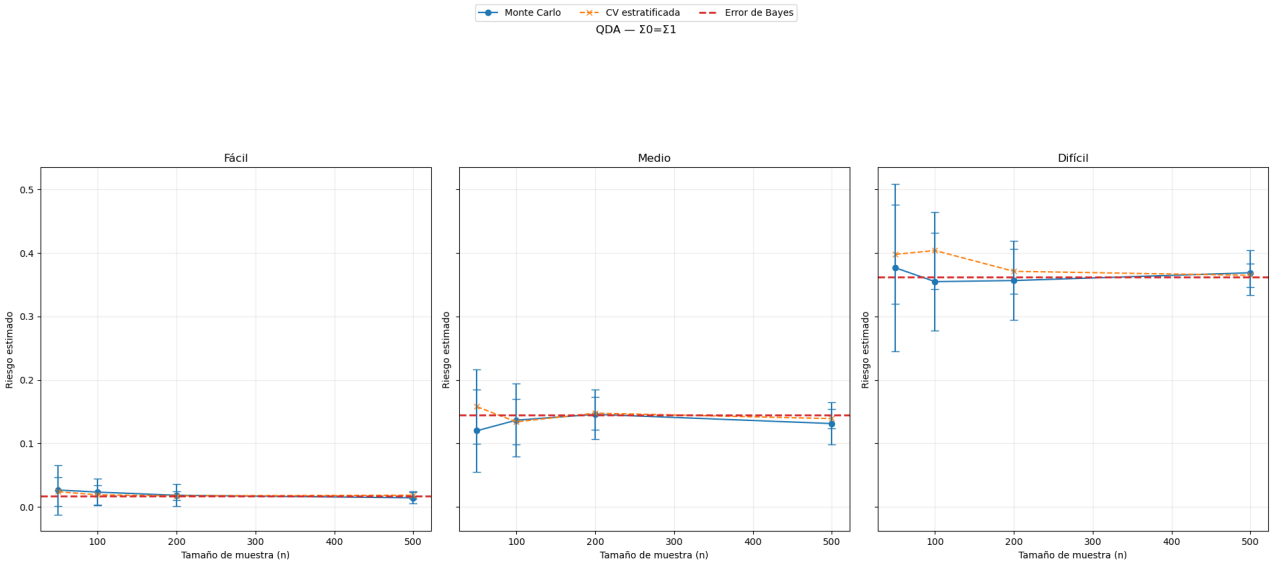


Figura 30: QDA: riesgo verdadero vs. validación para diferentes dificultades, en el caso de covarianzas iguales y balance.

4.2.2. Covarianzas iguales, $\pi_0 = 0.8$ (desbalance).

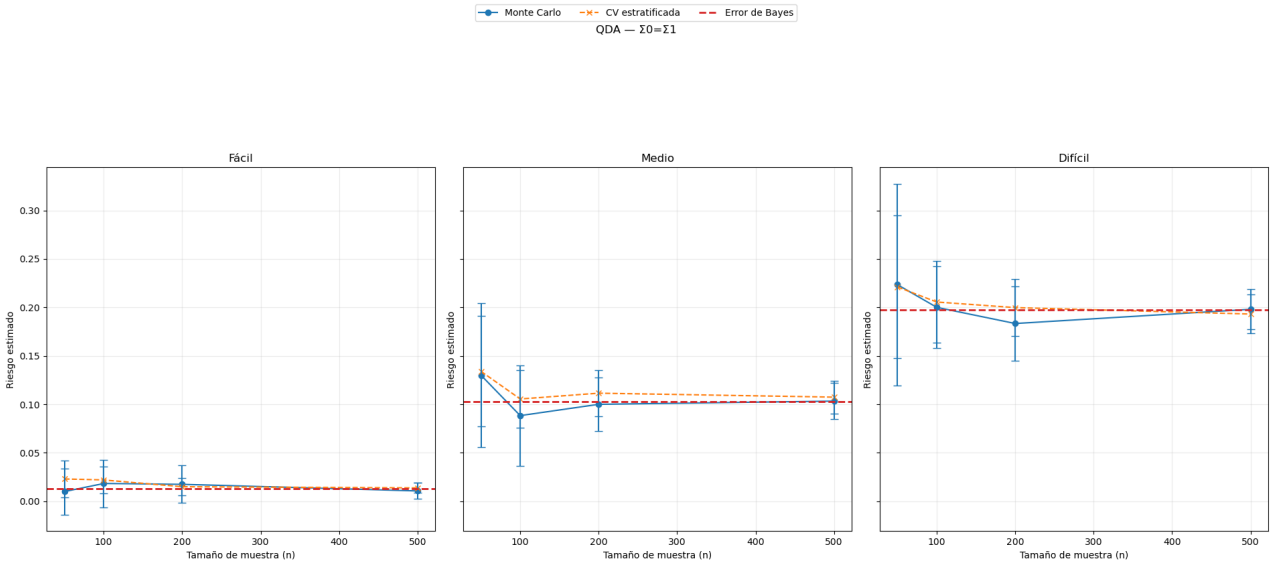


Figura 31: QDA: riesgo verdadero vs. validación para diferentes dificultades, covarianzas iguales y desbalance $\pi_0 = 0.8$.

El desbalance ($\pi_0 = 0.8$) mueve el umbral hacia la clase mayoritaria y cambia el nivel de R^*, pero no altera la conclusión de especificación.

- **Fácil.** L_{MC} y L_{CV} quedan muy próximos a R^* ya con $n \approx 100$; la brecha es mínima y la variabilidad pequeña. La CV estratificada presenta el sesgo positivo esperado (ligeramente por encima de MC), pero con menor dispersión.

- **Medio.** Para n pequeño se observa sobreestimación debida a la varianza en $\widehat{\Sigma}_0, \widehat{\Sigma}_1$; al crecer n ambas curvas descienden y se alinean con R^*. La estratificación en la CV es clave para mantener la proporción 0.8/0.2 en cada pliegue y estabilizar la estimación bajo desbalance.
- **Difícil.** La superposición entre clases eleva tanto R^* como la varianza inicial: MC muestra barras más amplias y valores algo alejados del óptimo cuando n es bajo. A partir de $n \approx 200$ las estimaciones de MC y CV convergen y se sitúan alrededor de R^*, aunque QDA sigue siendo más variable que un clasificador lineal en este caso.

4.2.3. Covarianzas distintas, $\pi_0 = 0.5$ (balance).

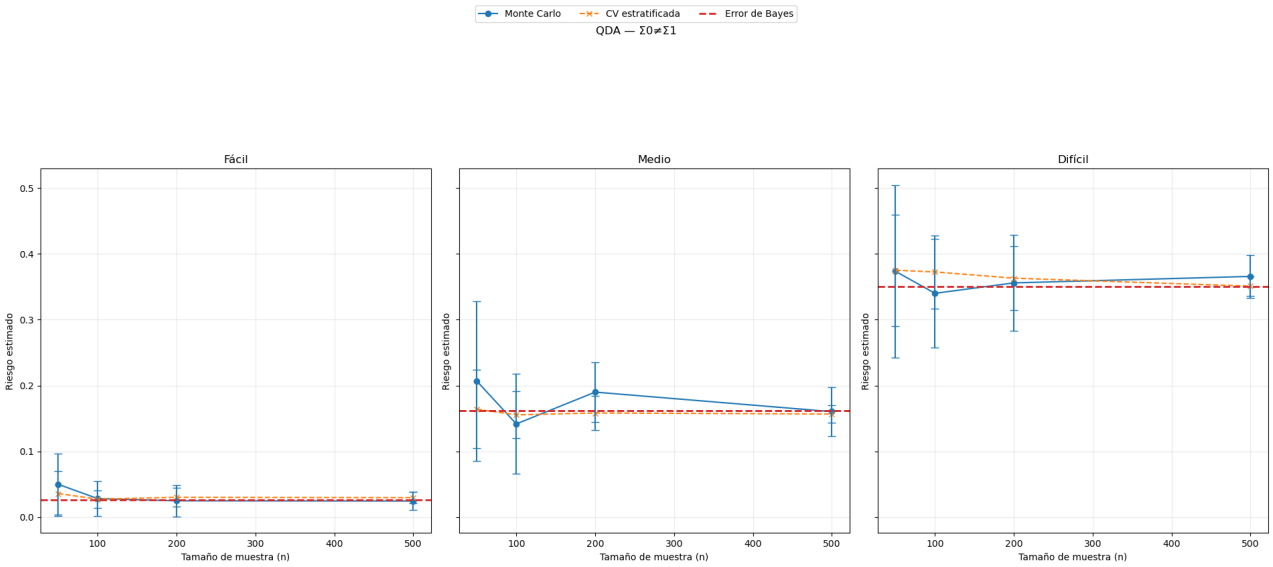


Figura 32: QDA: riesgo verdadero vs. validación para diferentes dificultades, covarianzas distintas y clases balanceadas.

Este es el escenario bien especificado para QDA: la frontera óptima de Bayes es cuadrática. Las curvas observadas son coherentes con esa predicción:

- **Fácil.** L_{MC} y L_{CV} quedan muy próximos a R^* desde tamaños muestrales pequeños; la brecha es mínima y la varianza rápidamente decrece. La ligera elevación de L_{CV} respecto de L_{MC} es el sesgo positivo típico de la CV.
- **Medio.** Con n pequeño se ve una sobreestimación inicial (parámetros $\hat{\mu}_k, \hat{\Sigma}_k$ poco precisos), seguida de un descenso sostenido hacia R^*. La CV permanece pegada a R^*
- **Difícil.** El solapamiento entre clases eleva R^* y amplifica la varianza inicial; en n bajos aparecen oscilaciones (incluso algún punto por debajo de R^*). A medida que crece n, tanto L_{MC} como L_{CV} convergen y se estabilizan alrededor de R^*.

4.2.4. Covarianzas distintas, $\pi_0 = 0.8$ (desbalance).

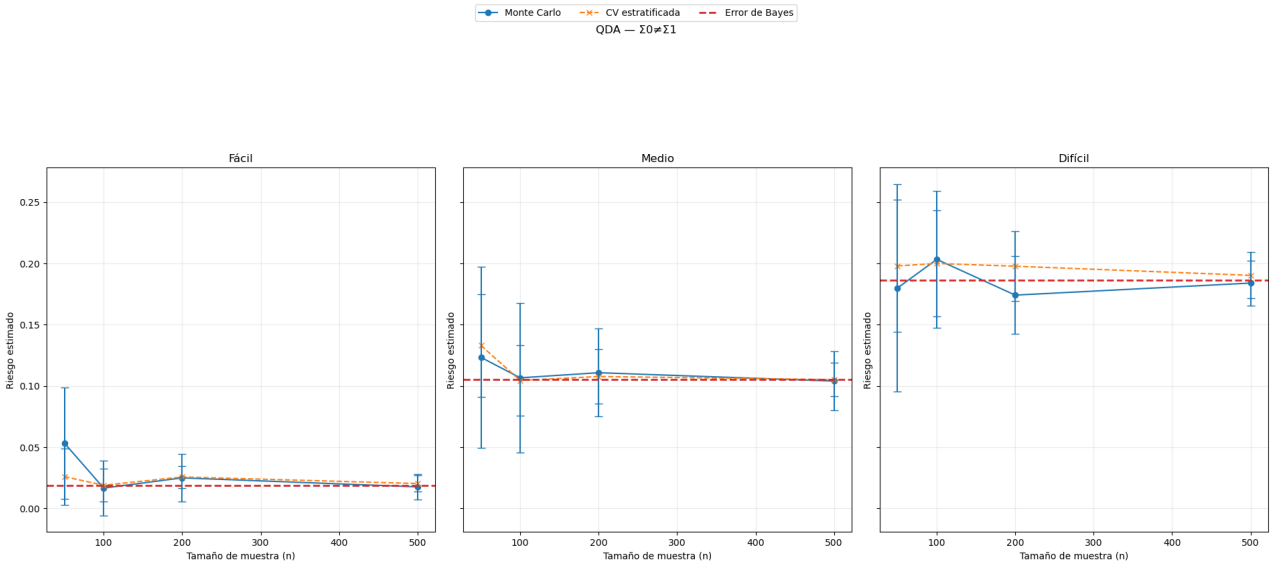


Figura 33: QDA: riesgo verdadero vs. validación para diferentes dificultades, covarianzas distintas y desbalance $\pi_0 = 0.8$.

- **Fácil.** Tanto L_{MC} como L_{CV} caen rápidamente y se ubican muy cerca de R^* a partir de tamaños muestrales moderados. La CV estratificada queda levemente por encima de MC (sesgo positivo esperado) para $n \approx 100$, pero con variabilidad menor.
- **Medio.** Con n pequeño aparece una sobreestimación inicial (debida a la imprecisión de $\widehat{\mu}_k$ y, sobre todo, de las $\widehat{\Sigma}_k$), seguida de un descenso sostenido hacia R^*. La CV reproduce el mismo comportamiento y converge prácticamente al mismo nivel que MC al aumentar n.
- **Difícil.** La mayor superposición entre clases eleva R^* y amplifica la varianza para n pequeños; se observa una oscilación de MC en n bajos que se estabiliza al crecer la muestra. En todo el rango, L_{CV} permanece ligeramente por encima de MC, pero ambas estimaciones se acercan gradualmente a R^*.

4.3. Naive Bayes

4.3.1. Covarianzas iguales, $\pi_0 = 0.5$ (balance).

- **Fácil.** L_{MC} y L_{CV} son prácticamente indistinguibles de R^* desde $n \approx 100$; la brecha es mínima y la variabilidad pequeña. Esto es consistente con que la suposición de NB coincide con el generador (independencia y varianzas iguales).
- **Medio.** Para n pequeños aparece una ligera sobreestimación para MC atribuible a error de estimación en medias/varianzas por clase; dicha brecha se reduce conforme aumenta el tamaño muestral y ambas curvas se alinean con R^*.
- **Difícil.** La mayor superposición entre clases eleva el nivel de riesgo y la varianza inicial. Observamos también que la estimación del error por validación cruzada converge más rápido a R^* que MC.

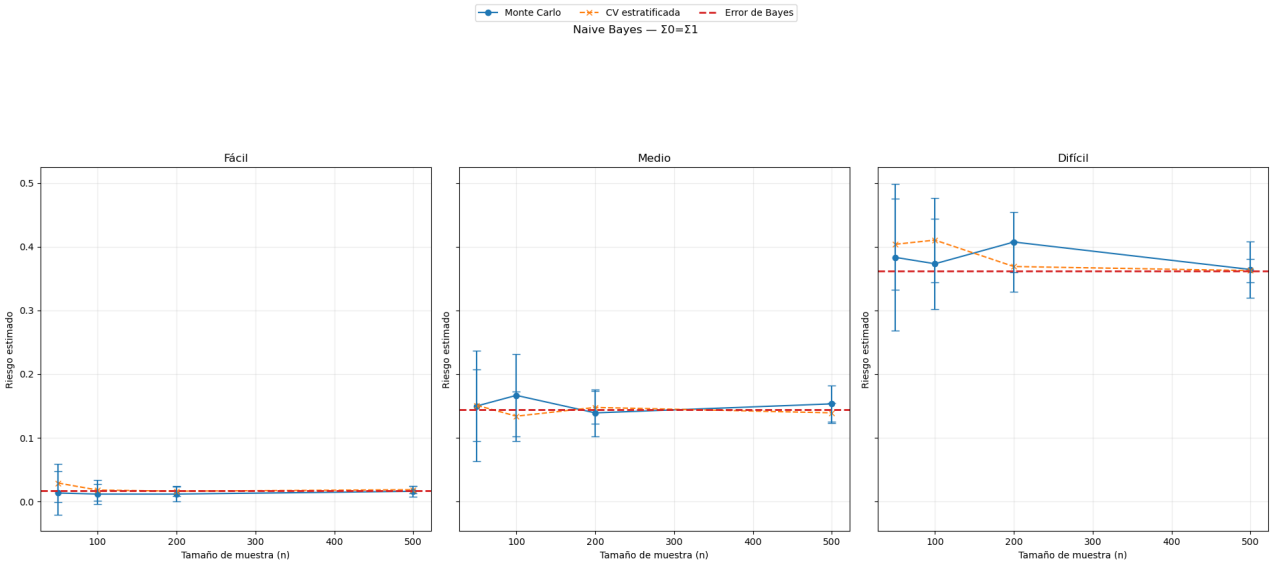


Figura 34: Naive Bayes: riesgo verdadero vs. validación para diferentes dificultades, covarianzas iguales y balance.

4.3.2. Covarianzas iguales, $\pi_0 = 0.8$ (desbalance).

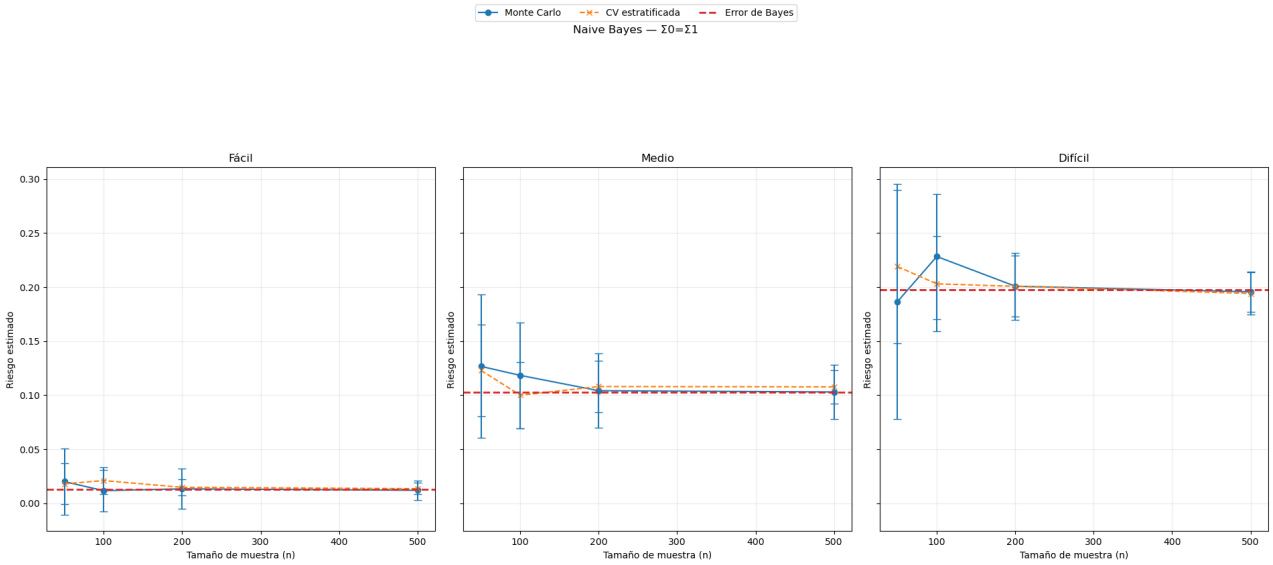


Figura 35: Naive Bayes: riesgo verdadero vs. validación para diferentes dificultades, covarianzas iguales y desbalance $\pi_0 = 0.8$.

- **Fácil.** L_{MC} y L_{CV} quedan prácticamente sobre R^* para n grandes; la brecha es mínima y la variabilidad pequeña. La CV estratificada aparece levemente por encima de MC (sesgo positivo esperado), con menor dispersión.
- **Medio.** Para n pequeños se observa una ligera sobreestimación (errores de estimación en medias y varianzas por clase); a medida que n crece, ambas curvas descienden y se alinean con R^*.

- **Difícil.** La mayor superposición entre clases eleva R^* y la varianza inicial: MC exhibe barras más amplias y valores por encima del óptimo cuando n es bajo. A partir de $n \approx 200$ los errores convergen y se estabilizan muy cerca de R^*.

4.3.3. Covarianzas distintas, $\pi_0 = 0.5$ (balance).

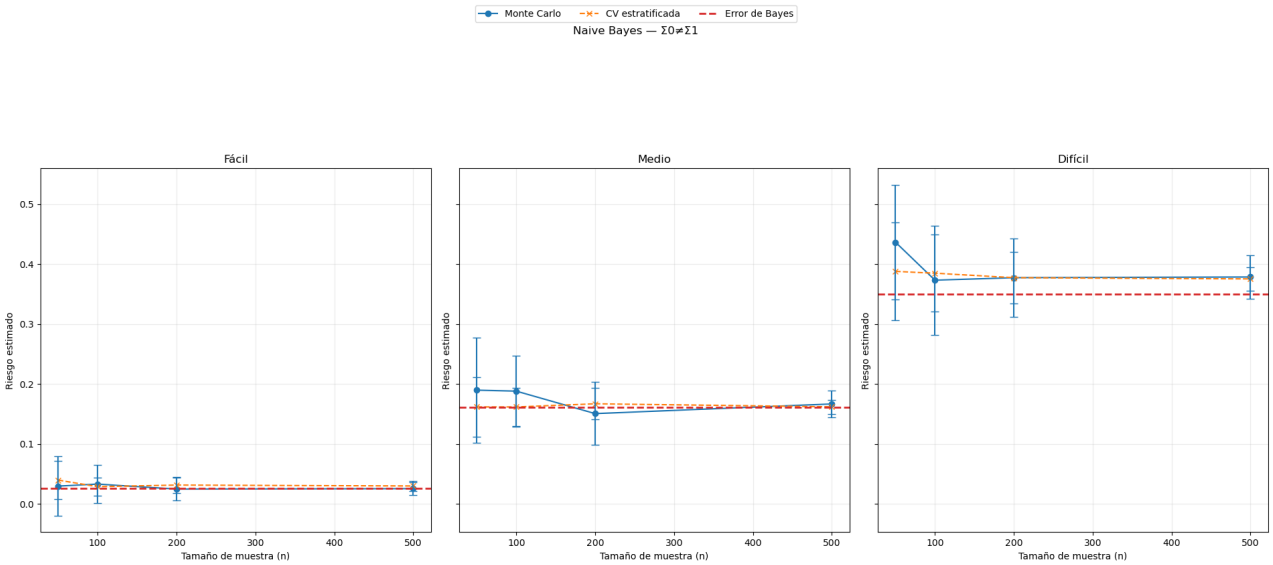


Figura 36: Naive Bayes: riesgo verdadero vs. validación para diferentes dificultades, covarianzas distintas y balance.

- **Fácil.** L_{MC} y L_{CV} descienden rápidamente y quedan muy cerca de R^* para $n \gtrsim 100$. La independencia condicional es poco costosa cuando las clases están bien separadas, y la variabilidad es pequeña.
- **Medio.** Con n pequeño se observan oscilaciones (algún punto de MC puede caer ligeramente por debajo de R^* por variabilidad), y luego ambas curvas se estabilizan alrededor del nivel de R^* conforme mejoran las estimaciones de medias y varianzas por clase.
- **Difícil.** La mayor superposición entre clases y la correlación no modelada por NB generan una brecha estructural: tanto L_{MC} como L_{CV} se estabilizan por encima de R^* incluso con n grande (brecha positiva persistente) pero ambas curvas muestran el mismo límite.

4.3.4. Covarianzas distintas, $\pi_0 = 0.8$ (desbalance).

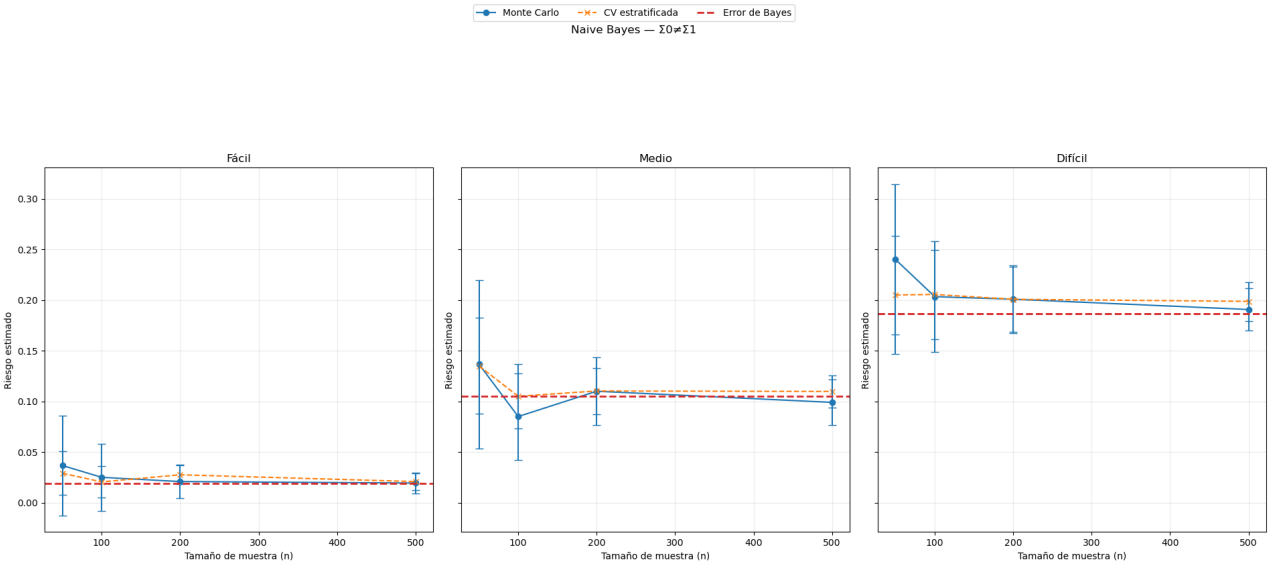


Figura 37: Naive Bayes: riesgo verdadero vs. validación para diferentes dificultades, covarianzas distintas y desbalance $\pi_0 = 0.8$.

- **Fácil.** L_{MC} y L_{CV} caen rápido y quedan muy cerca de R^* desde $n \approx 100$. La independencia condicional no penaliza casi nada cuando la separación geométrica es grande; la CV estratificada aparece levemente por encima de MC (sesgo positivo) y con menor varianza.
- **Medio.** Para n pequeños se observa una sobreestimación que se reduce al crecer n hasta alinearse con R^*.
- **Difícil.** La superposición entre clases y la correlación no modelada generan una brecha estructural: tanto L_{MC} como L_{CV} se estabilizan por encima de R^* incluso con n grande (aunque MC se acerca más). La CV mantiene un nivel apenas mayor que MC.

5. Conclusiones

En un entorno controlado con generadores gaussianos, el clasificador de Bayes (R^*) nos sirvió como línea base para entender qué tan cerca puede llegar cada método en función de la dificultad del problema, la correcta especificación del modelo y el tamaño muestral.

- **Efecto de la especificación del modelo.** Cuando $\Sigma_0 = \Sigma_1$, la frontera óptima es lineal y LDA tiende a R^* con rapidez; QDA está sobreparametrizado en este escenario y paga un costo de varianza para n pequeños sin obtener ganancia de sesgo. Cuando $\Sigma_0 \neq \Sigma_1$, la frontera óptima es cuadrática y QDA se vuelve el método mejor especificado. Naive Bayes funciona muy bien cuando la independencia condicional es una aproximación razonable o la separación geométrica es grande, pero exhibe una brecha estructural cuando hay correlaciones no modeladas.
- **Efecto del tamaño muestral y dificultad.** En los tres niveles (fácil, medio, difícil) observamos un patrón coherente con la intuición de sesgo–varianza: a mayor solapamiento de clases, mayor R^* y mayor varianza inicial de las estimaciones; con n creciente, los métodos bien especificados se estabilizan cercana o prácticamente en R^*.

- **k-NN.** El desempeño depende sensiblemente de k: valores pequeños reducen el sesgo pero aumentan la varianza; valores grandes hacen lo contrario. Con n suficiente existe una zona intermedia de k que aproxima bien a R^*, pero el método puede quedar por encima del óptimo cuando el problema es difícil o el k elegido no es adecuado.
- **Validación vs. riesgo verdadero.** Las curvas de validación cruzada estratificada (L_{CV}) replican sistemáticamente el perfil del riesgo Monte Carlo (L_{MC}) y del R^*, con un ligero sesgo esperado. En escenarios bien especificados y con n moderado, L_{CV} es un estimador práctico y confiable del desempeño fuera de muestra.

Referencias

[Friedman, 1989] Friedman, J. (1989). Regularized discriminant analysis.

[Fukunaga, 2013] Fukunaga, K. (2013). Introduction to statistical pattern recognition.