CIMAT

Análisis computacional mediante simulaciones

Introducción a Ciencia de Datos.

Autores: Brain de Jesús Salazar García [†], Rodrigo Gonzaga Sierra[†].
Profesor: Dr. Marco Antonio Aquino López [†]

Centro de Investigación en Matemáticas A. C.[†]

RESUMEN:
En este reporte se describe la implementación en Python de un análisis computacional mediante simulaciones comparando el desempeño de diversos clasificadores frente al clasificador óptimo de Bayes. El sistema genera datos sintéticos a partir de distribuciones normales multivariadas y evalúa métodos como LDA, QDA, Naive Bayes, Fisher y k-NN en diferentes escenarios controlados.

Índice

1. Introducción

En este trabajo, se realiza el estudio del desempeño de distintos modelos de clasificación como, Naive Bayes gaussiano, LDA, QDA, criterio de Fisher y K-NN y se comparan los riesgos de clasificación de cada metodo contra el riesgo óptimo de Bayes. Estás comparaciones se realizan bajo 4 distintos entornos controlados los cuales van desde casos en que las poblaciones tienen distintos comportamientos por clase, hasta el caso en que los comportamientos son bastante similares entre ambas clases. Así mismo, se analizan los distintos escenarios en los que se pueden encontrar las poblaciones, es decir, el desbalance de poblaciones y casos en que las diferencias entre poblaciones no son grandes.

Para poder realizar esa tarea, se creó un programa en el cual se generan dos muestras que simulan el comportamiento de cada población,

$$X|Y=0 \sim N(\mu_0, \Sigma_0) \quad \text{y} \quad X|Y=1 \sim N(\mu_1, \Sigma_1)$$

y se establecen las distribuciones a priori de Y dadas por $\pi_0 = \mathbb{P}[Y=0]$ y $\pi_1 = \mathbb{P}[Y=1]$. Para poder abordar los distintos escenarios mencionados anteriormente, se asignan distintos valores a los parámetros de las distribuciones de cada población, $\Sigma_0 = \Sigma_1, \Sigma_0 \neq \Sigma_1$, con distintos grados de correlación y vectores de medias cercanos y lejanos. En cada escenario se varían los tamaños muestrales por clase ($n \in \{50, 100, 200, 500\}$) y el parámetro k para el número de vecinos, ($k \in \{1, 3, 5, 11, 21\}$). Por último, los riesgos de clasificación se obtienen mediante validación cruzada y el riesgo de Bayes mediante integración Monte Carlo.

2. Análisis.

De acuerdo con [Aquino López, 2024a], se sabe que el clasificador de Bayes es el método con la menor probabilidad de error que se pueda tener. En este sentido, se espera que los demás métodos se encuentren con errores por encima del de Bayes, así, resulta relevante analizar cual de estos métodos se aproxima más al error de Bayes.

Se inicia analizando el caso en que las matrices de varianza y covarianza son iguales. Observe que en este caso se satisface con las hipótesis del modelo LDA, por lo que es razonable observar que LDA logré acercarse al riesgo óptimo de Bayes, en la figura (3) se puede observar que la brecha es de 0.004, mientras que en la figura (1) se puede observar que el riesgo se aleja bastante para $n = 200$ y se aproxima conforme n crece. Por otro lado, se puede observar que el comportamiento del modelo Naive Bayes, dependerá más de la forma de la matriz de varianzas y covarianzas, ya que en el caso en que esta matriz tiene variables independientes o con poca correlación, el modelo Naive Bayes proporciona buenos resultados, esto se puede apreciar mejor en la figura (1), así como se puede ver en la figura (3), que para este caso la brecha es de 0.002. Por otro lado, cuando las matrices de varianzas y covarianzas pertenecen a variables muy correlacionadas, el modelo Naive Bayes comienza a tener problemas. Note también que el modelo QDA tiene síntomas de sobre ajuste cuando la muestra es pequeña, en la figura (1) se puede ver que para n pequeño el riesgo se aproxima bastante al óptimo, sin embargo no se mantiene

conforme crece n. Mientras que el clasificador de Fisher muestra en la figura (1) un desempeño regular y bastante estable, se puede observar en la figura (3) que a brecha para este método es de 0.004. Por último, el método de k vecinos más cercanos no muestra tanta estabilidad además de que parece ser el método con peor desempeño ya que se aproxima poco al óptimo de Bayes.

Por otro lado, para el caso en que las matrices de varianzas y covarianzas son distintas, es razonable observar que el modelo LDA presenta un mal desempeño, esto debido a que se está violando el supuesto de matrices de varianzas y covarianzas iguales, en la figura (1) se puede observar que si bien el riesgo parece bajar conforme crece n, no parece suficiente para que el riesgo se aproxime al riesgo optimo, del mismo modo, en la figura (3) se puede observar que la brecha es de 0.009. Por otro lado, el modelo QDA presenta una ventaja notable ya que se satisface con las hipótesis adecuadas para poder implementar este modelo, observe que en eser caso, el riesgo de QDA se acerca bastante al riesgo óptimo de Bayes, según la figura (1), cuando n no es tan grande. El desempeño del método Naive Bayes va a depender mayormente del tipo de covarianza que tenga en las matrices, ya que si se ajustan las matrices de tal manera que aproxime ala independencia de las covariables, el método mejora considerablemente su desempeño, sin embargo de manera general, esto no es así. Para este caso, se puede observar que el riesgo se mantiene bastante estable conforme crece n.

De manera general, se puede observar que en los casos de desbalance, los métodos favorecen mayormente a la población mayoritaria. La mayoría de los métodos tienen comportamientos similares a excepción del modelo de clasificación de Fisher, el cual tuvo un comportamiento bastante alejado del resto de los modelos. Una forma de tratar de mitigar este problema, es mediante el ajuste de las distribuciones a priori, sin embargo esto no necesariamente resuelve el problema de la sensibilidad de la clase minoritaria, sino que motiva que una mayor parte de la población pueda ser predicada como de la clase minoritaria. Por último, observe que para los casos en que se tiene un mal condicionamiento y mucha correlación, los modelos LDA y QDA muestran problemas, lo que puede ser un síntoma a la hora de calcular las matrices inversas. Mientras que modelos como Naive Bayes y k vecinos se mostraron un poco más estables; sin embargo, los riesgos siguen siendo altos.

3. Presentación de resultados.

En esta sección, se presentan los resultados obtenidos en forma de gráficos, con los cuales se pretende ilustrar de una mejor manera del comportamiento que se tuvo cuando se varían ciertos parámetros.

A continuación, se presentan los valores del riesgo de clasificación para los distintos modelos conforme va cambiando el tamaño de la muestra.

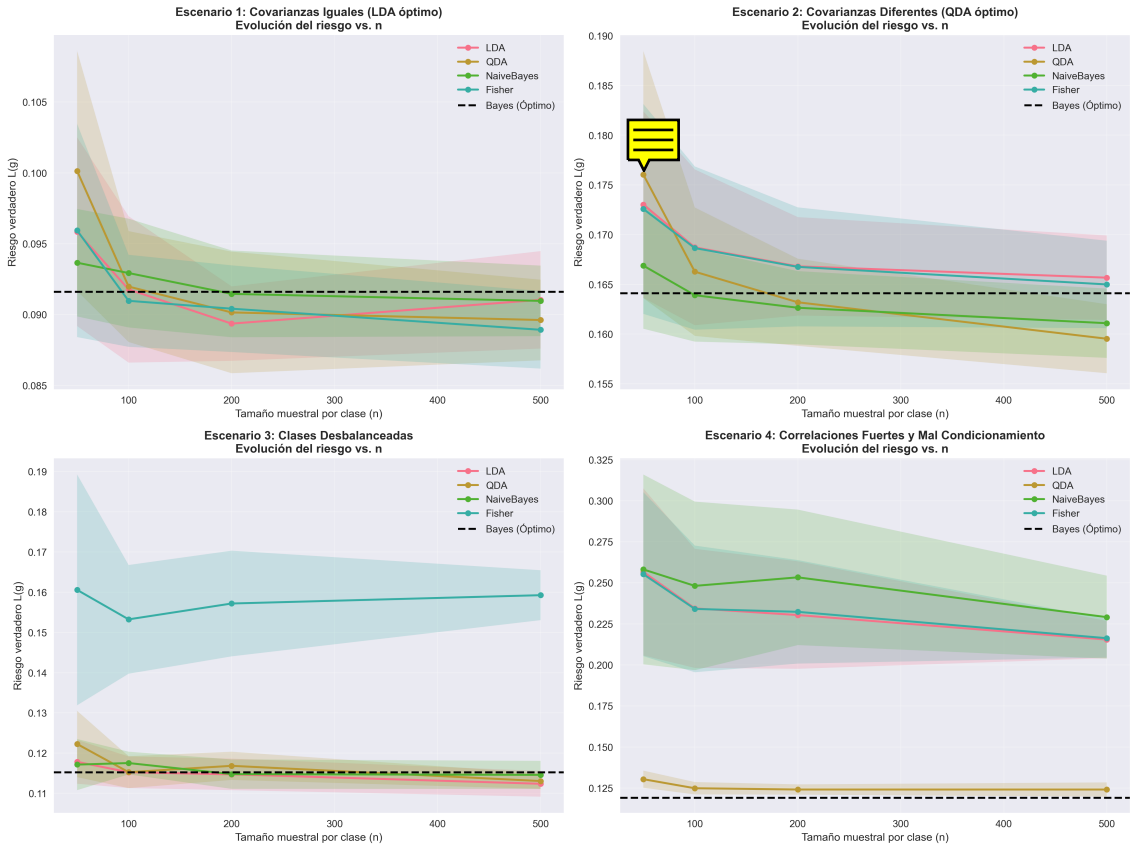

Figura 1: Evolución del riesgo vs Tamaño muestral.

Con esta figura es más claro observar que el modelo QDA muestra una gran variabilidad y un aparente sobre ajuste cuando el tamaño de la muestra e pequeño. Además, observe que el modelo Naive Bayes se muestra estable conforme crece el tamaño de la muestra. El modelo LDA también muestra un buen desempeño conforme crece el tamaño de la muestra, acercándose suficientemente al óptimo conforme crece la muestra.

En la figura (2) se puede observar el comportamiento del riesgo para el método de k vecinos más cercanos conforme se varía el número de vecinos a considerar, k. Observe que el comportamiento en cada escenario va a depender del valor de k, en los casos bien planteados, los comportamientos parecen regulares, sin embargo para el caso de desbalance, se comienzan a ver problemas para valores pequeños de k. De manera general, se puede observar que en valores pequeños de k, la varianza es grande, cuando k está en valores intermedios se comporta de manera aceptable y cuando k crece el riesgo crece. No fue posible alcanzar el óptimo de Bayes en este caso.

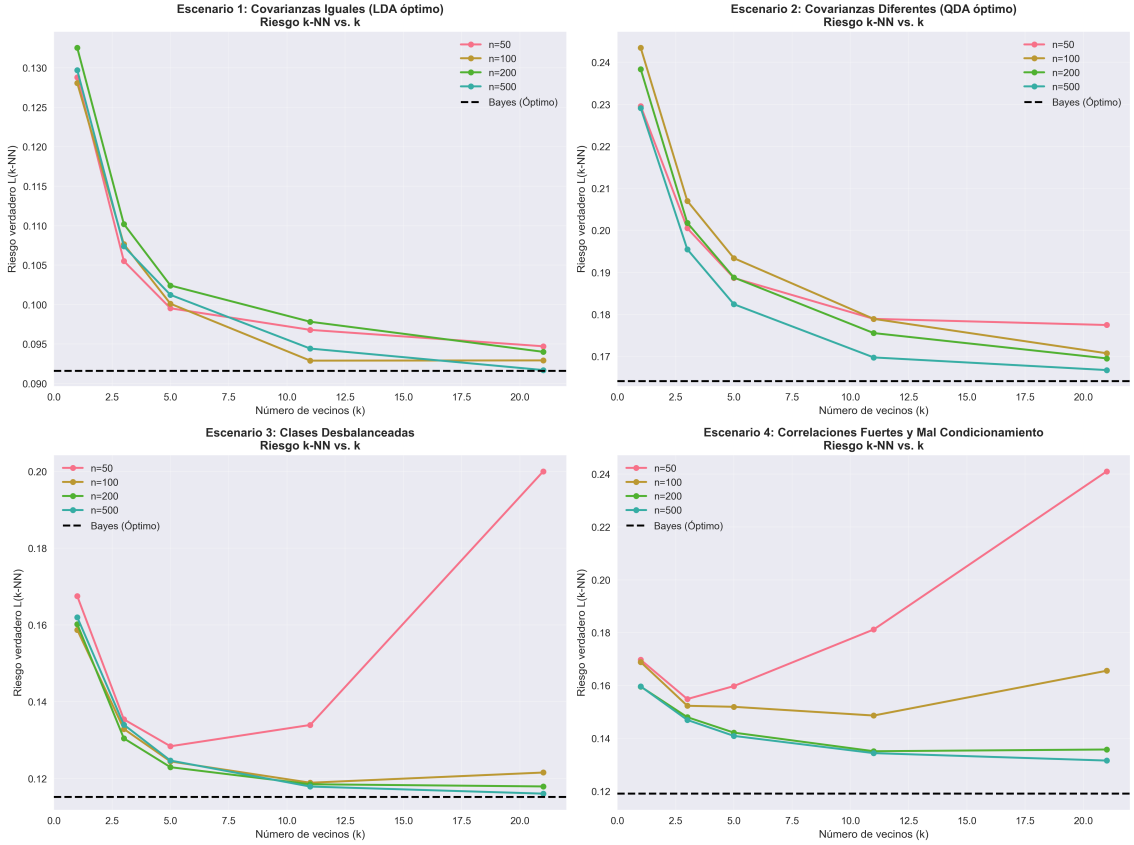

Figura 2: Comportamiento de riesgo para K-NN con distintos valores de K y distintos tamaños de población.

En la figura (3) se puede observar el comportamiento de los sesgos en cada escenario. En los casos de varianzas iguales y distintas, se puede notar que los mayores sesgos se obtienen cuando la población es pequeña y mejora de distintas maneras para cada método, conforme crece la población. En esta tabla se puede ver más claramente que el modelo LDA tiene un buen desempeño par el caso de varianzas iguales con una población moderada, mientras que para el caso en que las varianzas son distintas, el modelo Naive Bayes y QDA muestran un mejor desempeño con poblaciones moderadas. Observe que en el caso de clases desbalanceadas, el modelo que muestra más problemas es el de clasificación de Fisher y para el caso de correlaciones fuertes y mal condicionamiento, la mayoría de los modelos tienen muy mal desempeño menos QDA, que tiene un desempeño regular en este caso.

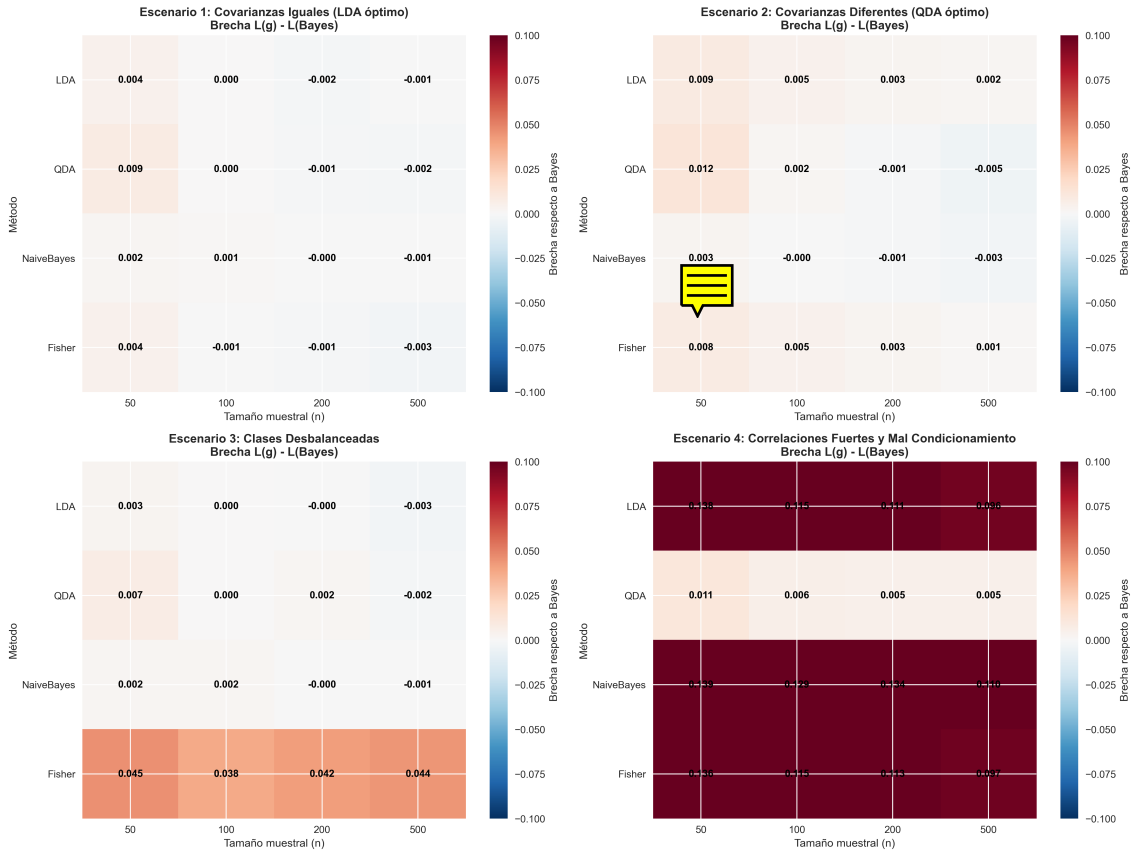

Figura 3: Tablas de sesgos de los riesgos respecto al óptimo de Bayes.

Por último, en la figura (**??**) se puede observar como se comporta el riesgo verdadero y el riesgo estimado mediante validación cruzada para distintos tamaños de muestra y para los distintos modelos. En este caso se puede observar que en varias ocasiones el riesgo estimado por validación cruzada se encuentra por encima de los valores reales del riesgo. Es natural observar que conforme crece n, los valores de los riesgos se aproximen.

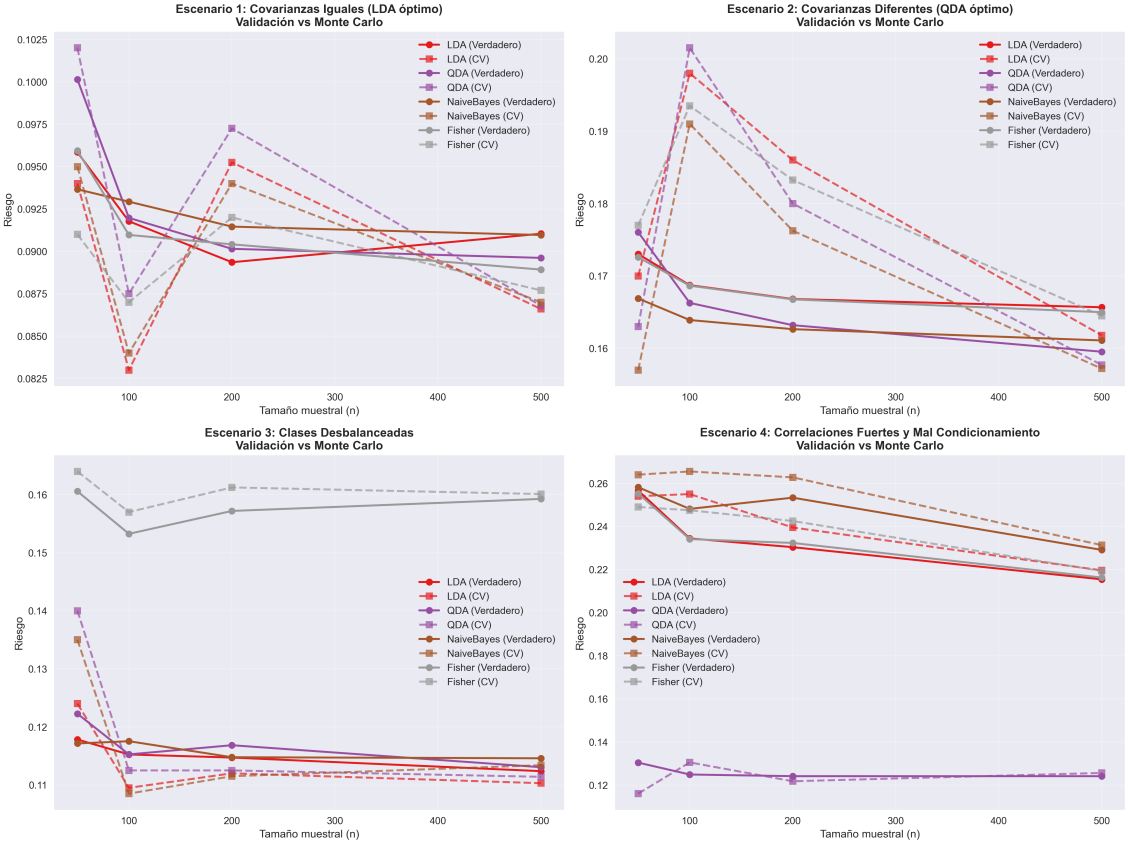

Figura 4: Validación cruzada vs Montecarlo.

3.1. Tablas de resumen de resultados

A continuación, se muestran resúmenes en formato de tabla de los resultados de la experimentación.

Cuadro 1: Resumen de Riesgos de Clasificación por Escenario (tamaño muestra = 500)

Método	Escenario 1	Escenario 2	Escenario 3	Escenario 4
Fisher	0.0889	0.1650	0.1593	0.2163
LDA	0.0910	0.1657	**0.1124**	0.2154
NaiveBayes	0.0910	0.1611	0.1146	0.2291
QDA	0.0896	**0.1595**	0.1131	**0.1241**
kNN_k1	0.1297	0.2292	0.1620	0.1597
kNN_k11	0.0944	0.1697	0.1179	0.1344
kNN_k21	0.0917	0.1667	0.1161	0.1316

- **Escenario 1 (Covarianzas Iguales)**: Métodos lineales (Fisher, LDA, QDA) muestran rendimiento similar, siendo Fisher ligeramente mejor. kNN con k alto se acerca al rendimiento óptimo.

- **Escenario 2 (Covarianzas Diferentes)**: QDA es el óptimo como se esperaba, seguido de cerca por NaiveBayes. Métodos lineales tienen mayor error.
- **Escenario 3 (Clases Desbalanceadas)**: LDA obtiene el mejor rendimiento, seguido por QDA y NaiveBayes. Fisher tiene pobre desempeño en este escenario.
- **Escenario 4 (Correlaciones Fuertes)**: QDA domina claramente, siendo el único método con error bajo (0.124). Los métodos lineales fallan significativamente debido al mal condicionamiento.

Método	**n=50**	**n=100**	**n=200**	**n=500**
Escenario 1: Covarianzas Iguales (LDA óptimo)				
LDA	0.0058	0.0018	-0.0006	0.0010
QDA	0.0101	0.0020	0.0001	-0.0004
Naive Bayes	0.0037	0.0029	0.0015	0.0010
Fisher	0.0059	0.0010	0.0004	-0.0011
k-NN (k=1)	0.0388	0.0381	0.0425	0.0397
k-NN (k=11)	0.0068	0.0029	0.0078	0.0044
Escenario 2: Covarianzas Diferentes (QDA óptimo)				
QDA	0.0160	0.0063	0.0032	-0.0005
LDA	0.0130	0.0087	0.0068	0.0057
Naive Bayes	0.0069	0.0039	0.0027	0.0011
Fisher	0.0126	0.0087	0.0068	0.0050
k-NN (k=1)	0.0696	0.0835	0.0784	0.0692
k-NN (k=11)	0.0189	0.0189	0.0155	0.0097
Escenario 3: Clases Desbalanceadas				
LDA	0.0078	0.0052	0.0047	0.0024
QDA	0.0123	0.0053	0.0068	0.0031
Naive Bayes	0.0071	0.0075	0.0048	0.0046
Fisher	0.0506	0.0433	0.0472	0.0493
k-NN (k=1)	0.0575	0.0487	0.0502	0.0520
k-NN (k=11)	0.0240	0.0089	0.0085	0.0079
Escenario 4: Correlaciones Fuertes				
QDA	0.0104	0.0049	0.0041	0.0041
k-NN (k=1)	0.0497	0.0489	0.0396	0.0397
k-NN (k=11)	0.0612	0.0286	0.0151	0.0144
LDA	0.1366	0.1144	0.1104	0.0954
Fisher	0.1353	0.1141	0.1123	0.0963
Naive Bayes	0.1382	0.1281	0.1333	0.1091

Cuadro 2: Brechas de riesgo respecto al clasificador óptimo de Bayes por escenario y tamaño muestral. Valores más cercanos a cero indican mejor performance.

1. **Escenario 1 (Covarianzas Iguales)**: Como se esperaba teóricamente, **LDA** muestra brechas muy pequeñas (cercanas a cero), (al igual que Naive Bayes), confirmando su optimalidad en este escenario.
2. **Escenario 2 (Covarianzas Diferentes)**: **QDA** es el método óptimo, mostrando las brechas más pequeñas que convergen a cero cuando n aumenta. Naive Bayes también performa notablemente bien.
3. **Escenario 3 (Clases Desbalanceadas)**: Los métodos **LDA, QDA y Naive Bayes** muestran brechas pequeñas y consistentes. Fisher y k-NN con k pequeño tienen dificultades con el desbalance.

4. **Escenario 4 (Correlaciones Fuertes)**: **QDA** domina claramente con las brechas más pequeñas, seguido por k-NN con k moderado. Los métodos lineales (LDA, Fisher) y Naive Bayes struggle significativamente.

Cuadro 3: Resumen de Rendimiento de Métodos de Clasificación

	Escenario 1 (LDA)		Escenario 2 (QDA)	
	Riesgo Promedio	Gap Promedio	Riesgo Promedio	Gap Promedio
LDA	0.0924	0.0008	0.1712	0.0071
QDA	0.0956	0.0040	0.1721	0.0080
NaiveBayes	0.0929	0.0013	0.1656	0.0015
Fisher	0.0927	0.0011	0.1693	0.0052
kNN_k1	0.1328	0.0412	0.2213	0.0572
kNN_k3	0.1059	0.0143	0.1891	0.0250
kNN_k5	0.0989	0.0073	0.1809	0.0168
kNN_k11	0.0953	0.0037	0.1745	0.0104
kNN_k21	0.0938	0.0022	0.1716	0.0075

- **Escenario 1 (Covarianzas Iguales)**: Como era de esperar, **LDA** muestra el mejor rendimiento con un riesgo promedio de 0.0924, muy cercano al riesgo bayesiano teórico de 0.0916.
- **Escenario 2 (Covarianzas Diferentes)**: **QDA** y **NaiveBayes** obtienen los mejores resultados, con NaiveBayes mostrando el gap más pequeño (0.0015) respecto al riesgo bayesiano de 0.1641.

4. Conclusiones

La estimación del error del clasificador óptimo de Bayes es un buen punto de referencia óptima para poder . Los resultados muestran que:

- LDA es óptimo cuando las covarianzas son iguales, mientras que QDA y NaiveBayes funcionan mejor cuando las covarianzas difieren. kNN con k grande puede ser una alternativa robusta cuando no se conocen las suposiciones distribucionales.
- Al aumentar el tamaño de muestra, la mayoría de métodos mejoran o estabilizan su rendimiento. QDA muestra robustez en escenarios complejos, mientras LDA funciona bien en condiciones balanceadas.
- En general, las brechas disminuyen cuando n aumenta, mostrando convergencia hacia el óptimo de Bayes, excepto para algunos casos de k-NN que muestran inestabilidad.

- **k-NN** es muy dependiente de k. Valores extremos (k=1) generan brechas grandes, mientras que k moderado (k=11,21) puede ser competitivo en algunos escenarios. Su rendimiento mejora significativamente al aumentar el valor de k. **kNN_k21** se acerca al rendimiento de los métodos paramétricos óptimos en ambos escenarios.

- Los métodos paramétricos (LDA, QDA, NaiveBayes, Fisher) muestran mayor estabilidad que kNN, especialmente con valores pequeños de k.

En general, las simulaciones controladas permiten contrastar métodos en condiciones ideales, comprender sus limitaciones ante los supuestos y sus cambios, con esto poder tomar decisiones a la ahora de enfrentarse a un problema real.

4.1. Mini discusión

El código que se generó para este reporte se puede extender para estudiar dimensiones superiores ($p > 2$), además de incluir distribuciones no necesariamente Gaussianas, analizar métodos de regularización e incorporar técnicas de selección de modelos automáticamente.

Referencias

[Aquino López, 2024a] Aquino López, M. A. (2024a). Diapositivas del curso: Introducción a la ciencia de datos. Recuperado de: `https://github.com/maquinolopez/Ciencia_De_Datos/blob/main/Diapostivas/Presentation_8.pdf`. Accedido: 28 de septiembre del 2025.

[Aquino López, 2024b] Aquino López, M. A. (2024b). Diapositivas del curso: Introducción a la ciencia de datos. Recuperado de: `https://github.com/maquinolopez/Ciencia_De_Datos/blob/main/Diapostivas/Presentation_7.pdf`. Accedido: 28 de septiembre del 2025.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). The elements of statistical learning.

Contribuciones:
Desarrollo completo del código y del reporte: Rodrigo y Brain de Jesús.