Proyecto 2: Clasificación supervisada (simulaciones)
Introducción a la Ciencia de Datos

Integrantes:	Avendaño Caballero, Joksan; Rodríguez Villagrán, Juan Pablo
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas
Profesor:	Dr. Marco Antonio Aquino López

Resumen

Mediante datos simulados provenientes de distribuciones normales bivariadas, se comparan los clasificadores Naïve Bayes, LDA, QDA y k-NN con el comparador óptimo de Bayes. Para esta comparación, se someten los clasificadores a seis escenarios: clases balanceadas con $\mu_0 \neq \mu_1$ y $\Sigma_0 = \Sigma_1$, clases balanceadas con $\mu_0 \neq \mu_1$ y $\Sigma_0 \neq \Sigma_1$, clases desbalanceadas con $\mu_0 \neq \mu_1$ y $\Sigma_0 = \Sigma_1$, clases balanceadas con $\mu_0 \approx \mu_1$ y $\Sigma_0 \approx \Sigma_1$, clases balanceadas con $\mu_0 \approx \mu_1$ y $\Sigma_0 = \Sigma_1$ y clases desbalanceadas con $\mu_0 \approx \mu_1$ y $\Sigma_0 = \Sigma_1$.

1. Introducción

El problema de clasificación en estadística consiste en proponer criterios estadísticos para separar los datos en niveles o categorías bien definidas. En el caso de la clasificación con categorías binarias, el problema se puede interpretar como uno de decisión en el que una de las categorías sea que sí se cumple una propiedad y la otra que no se cumpla.

En este trabajo se presentan ejemplos de uso de diversos clasificadores ante el problema de clasificación binaria. Formalmente, dada una variable aleatoria binaria $Y \in \{0, 1\}$, con proporciones $\pi_y = \mathbb{P}[Y = y]$ y un vector de predictores $X \in \mathbb{R}^p$, para una regla $g : \mathbb{R}^p \to \{0, 1\}$ el riesgo con pérdida 0-1 es

$$L(g) = \mathbb{P}[g(X) \neq Y].$$

Aquí es donde aparece el **clasificador de Bayes**

$$g^*(x) = \mathbb{1}_{\{\eta(x) \geq \frac{1}{2}\}}, \qquad \text{con} \quad \eta(x) = \mathbb{P}[Y = 1 | X = x],$$

el cual minimiza $L(g)$. Sin embargo, en la práctica, las funciones $\eta(x)$, $p(x|y)$ y π_y son desconocidas y se reemplazan por estimadores, dando lugar a clasificadores que modelan, ya sea a $p(x|y)$ y π_y o a $\eta(x)$ y $\mathbb{P}[Y|X]$.

En este trabajo se busca estudiar, mediante simulaciones controladas, cómo se comportan distintos clasificadores como Naïve Bayes, LDA, QDA y k-NN frente al óptimo g^* bajo escenarios Gaussianos donde se puede calcular explícitamente $L(g)$. Para esta comparación, se contemplan cuatro casos principales:

1. Covarianzas iguales
2. Covarianzas distintas
3. Desbalanceo en π_y
4. Correlaciones o mal condicionamiento en Σ_y

2. Descripción de los clasificadores

Como se mencionó en la introducción, el clasificador óptimo es el de Bayes, para el cual se pueden hacer cálculos explícitos bajo hipótesis de normalidad. En esta sección se hace una revisión breve de los clasificadores más usuales. Una discusión más completa de éstos se puede revisar en [HTF09] y [DGL96].

2.1. Clasificador de Bayes

Este clasificador es el óptimo y surge de maximizar la función de verosimilitud posterior o, equivalentemente, minimizar la función de riesgo L. Para distribuciones a priori π_y y densidades $f_y(x) = p(x|y)$, el clasificador de Bayes está dado por

$$g^*(x) = \arg \max_{y \in \{0,1\}} \ln \pi_y + \ln f_y(x).$$

Este clasificador es imposible de obtener en la práctica ya que requiere conocimiento sobre la distribución de los datos. En el caso Gaussiano se puede calcular fácilmente.

2.2. Análisis de Discriminante Lineal (LDA)

Para obtener este primer clasificador derivado del clasificador de Bayes, se trabaja bajo la hipótesis de que $(X|Y=y) \sim N(\mu_y, \Sigma)$, con covarianza común entre X y Y. Así, se clasifica por el mayor discriminante lineal

$$\delta_y(x) = x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln \pi_y,$$

el cual induce la frontera lineal $\{\delta_1(x) = \delta_0(x)\}$. Este clasificador es óptimo cuando $\Sigma_0 = \Sigma_1$, por lo que es sensible a diferencias en covarianza o indicios de no linealidad. Es estable para valores moderados de la dimensión ambiente p.

2.3. Clasificador de Fisher

Lo que deriva en el criterio de Fisher es buscar la proyección w^Tx que maximice la función

$$J(w) = \frac{w^TS_Bw}{w^TS_Ww}, \quad \text{donde,} \quad S_B = (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \quad \text{y} \quad S_W = \Sigma_0 + \Sigma_1.$$

Es binario por ortogonalidad ya que $w \propto S_W^{-1}(\mu_1 - \mu_0)$. Cuando $\Sigma_0 = \Sigma_1$, el criterio de Fisher coincide con el de LDA. Comúnmente es utilizado como herramienta de reducción de dimensión.

2.4. Análisis de Discriminante Cuadrático (QDA)

En contraste con el clasificador de LDA, en este caso la hipótesis es $(X|Y=y) \sim N(\mu_y, \Sigma_y)$, con covarianzas específicas por clase. Recibe su nombre porque ahora la frontera a maximizar

$$\delta_y(x) = -\frac{1}{2}\ln\det\Sigma_y - \frac{1}{2}(x-\mu_y)^T\Sigma_y^{-1}(x-\mu_y) + \ln\pi_y,$$

es una forma cuadrática. Este clasificador es óptimo precisamente cuando $\Sigma_0 \neq \Sigma_1$. Naturalmente las fronteras que define este clasificador son intersecciones de superficies cuádricas asociadas a Σ_y^{-1}. Como se vio en el trabajo con datos, este clasificador puede ser inestable cuando existe alta correlación entre variables por lo que se suele regularizar por un término constante λI.

2.5. Naïve Bayes

Este clasificador introduce hipótesis en la relación entre la densidad condicional de x dado y con un supuesto de independencia condicional, es decir, considerando

$$p(x|y) = \prod_{j=1}^{p} p(x_j|y).$$

En el caso Gaussiano, la hipótesis es $(x_j|y) \sim N(\mu_{jy}, \sigma_{jy}^2)$. Este clasificador suele ser robusto ante la dimensión. El principal problema radica en la hipótesis de independencia, la cual rara vez se cumple; del mismo modo, no hay criterios definidos para estimar σ_{jy}^2, por lo que puede ser ruidoso y complicado de calibrar.

2.6. k vecinos más cercanos (k-NN)

Este clasificador decide la clase de un punto x de acuerdo a un proceso de votación de los k vecinos más cercanos a x respecto a la métrica euclideana. Uno de los resultados básicos de este clasificador es que su eficiencia está acotada superiormente por el doble del riesgo de Bayes. La principal ventaja de este clasificador es que depende de un método no paramétrico, pero al ser dependiente de la métrica es muy sensible a la maldición de la dimensionalidad.

3. Simulaciones

En cada escenario se utilizan tamaños de muestra de $n = 50, 100, 200, 500$ y valores de $k = 1, 3, 5, 7$. Para replicabilidad, se utilizó la semilla en `Python` 1108, además de que en el código anexo se cuenta con una función para calcular estimadores bootstrap del riesgo en cada escenario con los distintos tamaños de muestra n y de vecinos k.

3.1. Escenario 1: clases balanceadas y matrices de covarianza iguales

Para este escenario, se simularon muestras con $\pi_y = (1/2, 1/2)$ de un vector tal que

$$(X|Y=0) \sim N((0,0)^T, \Sigma) \quad \text{y} \quad (X|Y=1) \sim N((3,3)^T, \Sigma) \quad \text{donde} \quad \Sigma = \begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}.$$

Para mostrar de manera sintética el rendimiento de los distintos clasificadores respecto al estimador óptimo, se presenta en la figuras 1 la comparación del riesgo de Bayes para cada uno de los clasificadores respecto al tamaño de muestra y en la figura 2 la comparación del riesgo de k-NN para distintos valores de k.

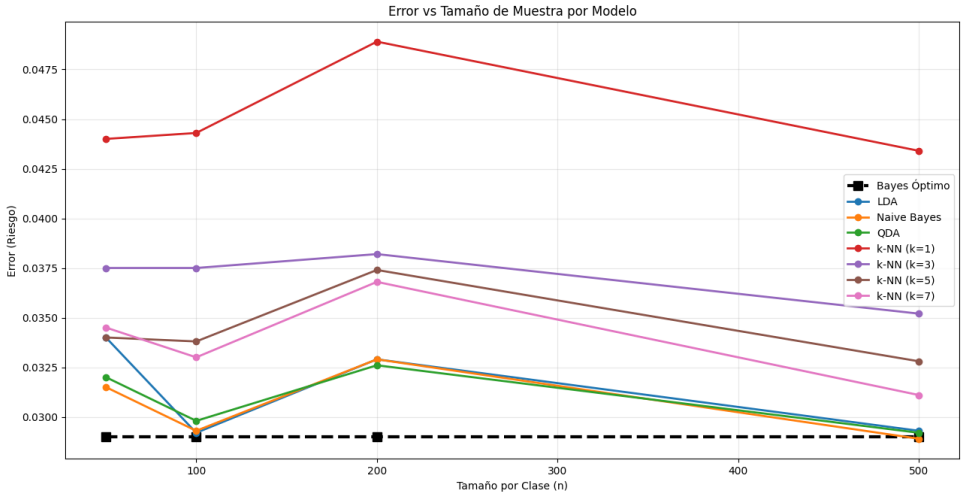

Figura 1: Comparación del riesgo de Bayes de los distintos clasificadores para el caso balanceado con matrices de covarianzas iguales para distintos tamaños de muestra.

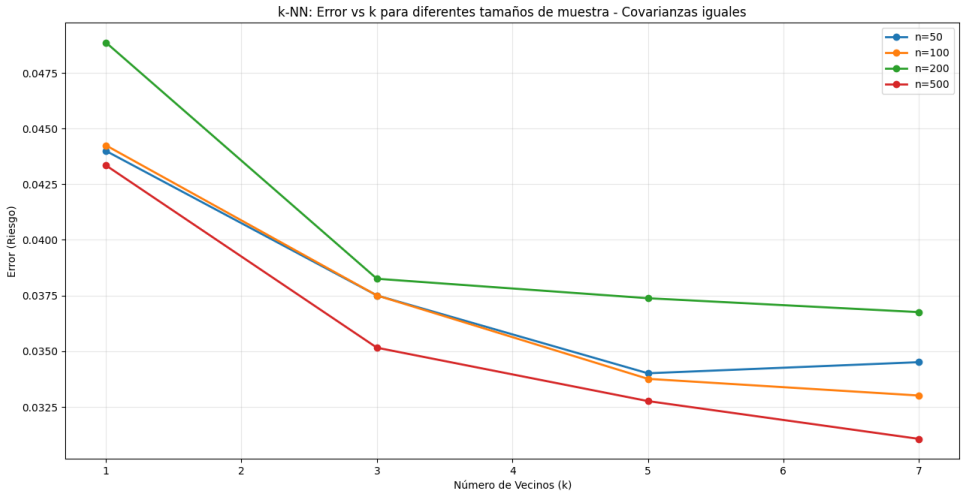

Figura 2: Comparación del riesgo de Bayes de los distintos clasificadores de k-NN balanceado con matrices de covarianzas iguales para distintos valores de k.

Notemos que los clasificadores que se mantienen siempre con un riesgo de Bayes bajo son LDA, QDA y Naïve Bayes. No es de sorprender que LDA y Naïve Bayes mantengan el riesgo bajo ya que se ven favorecidos por normales con la misma varianza y correlaciones pequeñas.

3.2. Escenario 2: clases balanceadas y matrices de covarianza distintas

Para este escenario, se simularon muestras con $\pi_y = (1/2, 1/2)$ de un vector tal que

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

$$\text{donde} \quad \Sigma_0 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \quad \text{y} \quad \Sigma_0 = \begin{pmatrix} 2 & -1/2 \\ -1/2 & 1 \end{pmatrix}.$$

Para mostrar de manera sintética el rendimiento de los distintos clasificadores respecto al estimador óptimo, se presenta en la figuras **??** la comparación del riesgo de Bayes para cada uno de los clasificadores respecto al tamaño de muestra y en la figura **??** la comparación del riesgo de k-NN para distintos valores de k.

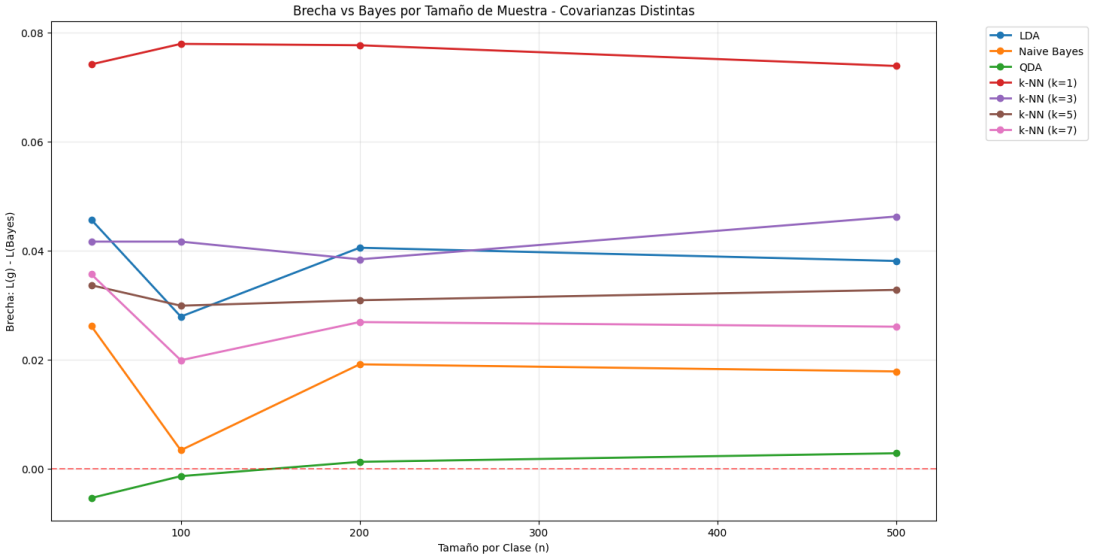

Figura 3: Comparación del riesgo de Bayes de los distintos clasificadores para el caso balanceado con matrices de covarianzas distintas para distintos tamaños de muestra. Se muestra la brecha entre el riesgo de Bayes y el de los clasificadores

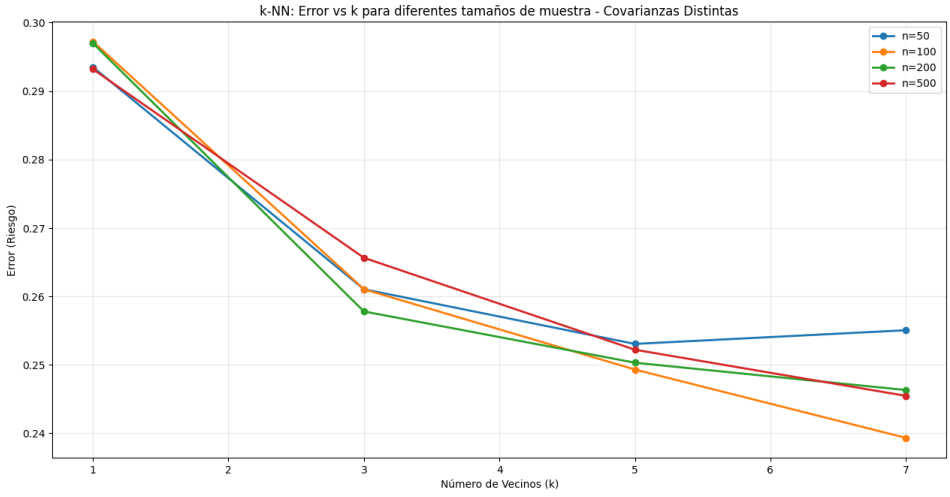

Figura 4: Comparación del riesgo de Bayes de los distintos clasificadores de k-NN balanceado con matrices de covarianzas distintas para distintos valores de k.

Notemos que en este caso el clasificador QDA casi alcanza la optimalidad. En la exposición inicial se mencionó que éste es el caso del cual surge QDA y este ejemplo lo ilustra.

3.3. Escenario 3: clases desbalanceadas y matrices de covarianza iguales

Para este escenario, se simularon muestras con $\pi_y = (4/5, 1/5)$ de un vector tal que

$$(X|Y=0) \sim N((0,0)^T, \Sigma) \quad \text{y} \quad (X|Y=1) \sim N((3/2,3/2)^T, \Sigma), \quad \text{donde} \quad \Sigma_0 = \begin{pmatrix} 1 & 3/10 \\ 3/10 & 1 \end{pmatrix}.$$

Para mostrar de manera sintética el rendimiento de los distintos clasificadores respecto al estimador óptimo, se presenta en la figuras 5 la comparación del riesgo de Bayes para cada uno de los clasificadores respecto al tamaño de muestra y en la figura 6 la comparación del riesgo de k-NN para distintos valores de k.

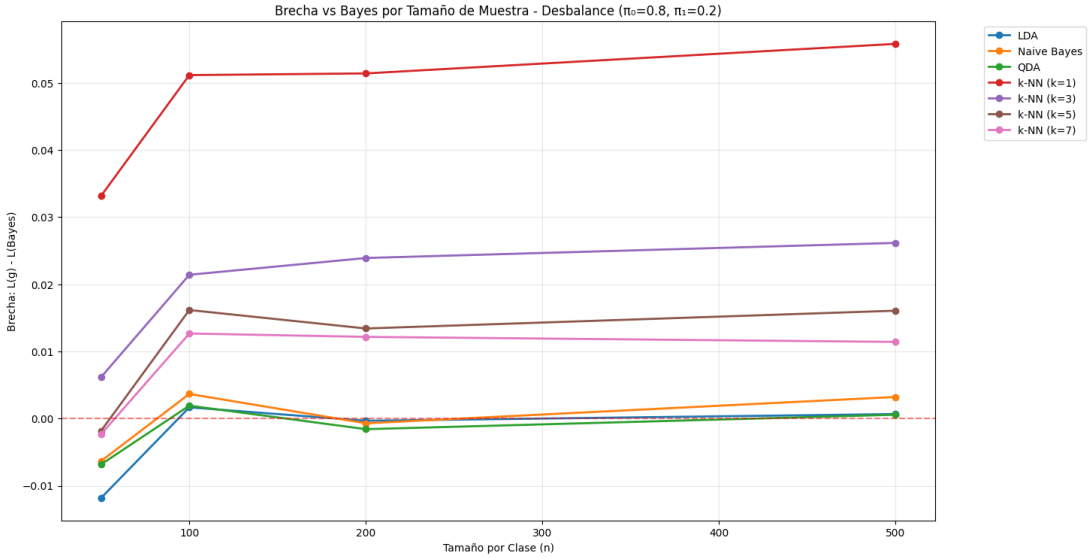

Figura 5: Comparación del riesgo de Bayes de los distintos clasificadores para el caso desbalanceado con matrices de covarianzas iguales para distintos tamaños de muestra. Se muestra la brecha entre el riesgo de Bayes y el de los clasificadores

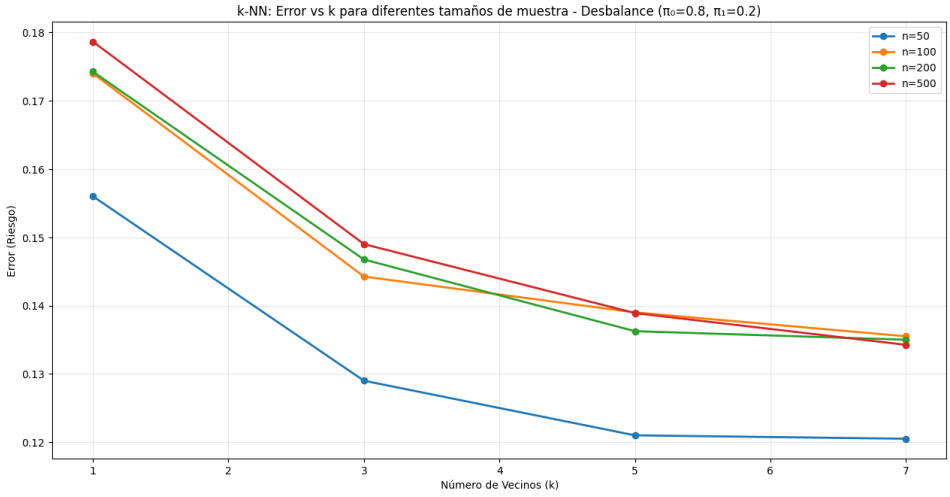

Figura 6: Comparación del riesgo de Bayes de los distintos clasificadores de k-NN desbalanceado con matrices de covarianzas iguales para distintos valores de k.

Notemos que en este, como en el primero, los clasificadores de LDA, QDA y Naïve Bayes son los que menos riesgo presentan. Es interesante notar que en el caso de los k-NN, el caso con una muestra más pequeña presenta menos riesgo que los otros. También se aprecia el comportamiento aproximadamente acotado del riesgo de los clasificadores k-NN.

3.4. Escenario 4: clases balanceadas, medias parecidas y matrices de covarianza iguales casi singular

Para este escenario, se simularon muestras con $\pi_y = (1/2, 1/2)$ de un vector tal que

$$(X|Y=0) \sim N((0,0)^T, \Sigma) \quad \text{y} \quad (X|Y=1) \sim N((1/2,1/2)^T, \Sigma), \quad \text{donde} \quad \Sigma_0 = \begin{pmatrix} 1 & 95/100 \\ 95/100 & 1 \end{pmatrix}.$$

Para mostrar de manera sintética el rendimiento de los distintos clasificadores respecto al estimador óptimo, se presenta en la figuras 7 la comparación del riesgo de Bayes para cada uno de los clasificadores respecto al tamaño de muestra y en la figura 8 la comparación del riesgo de k-NN para distintos valores de k.

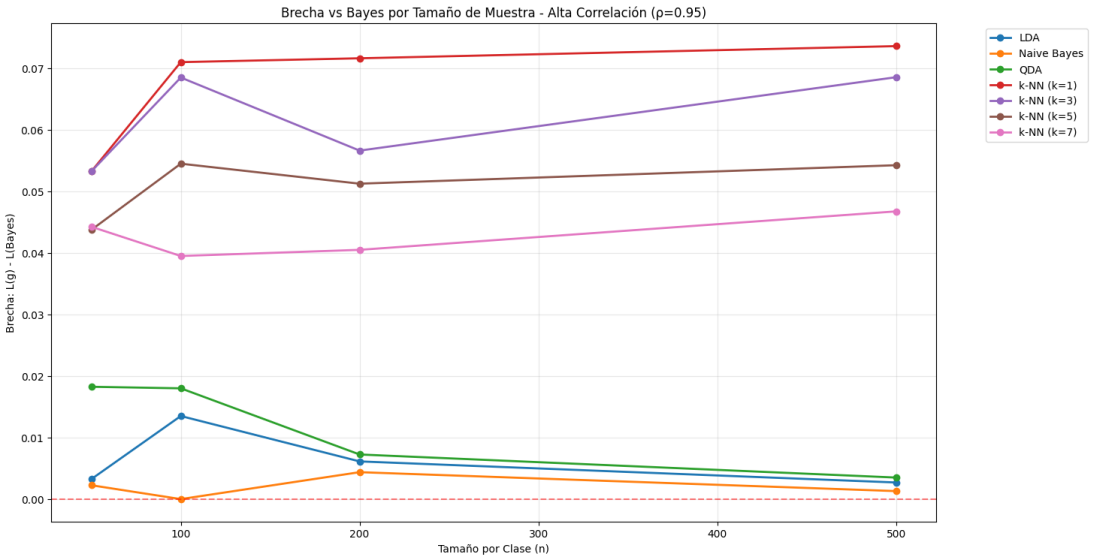

Figura 7: Comparación del riesgo de Bayes de los distintos clasificadores para el caso de clases balanceadas, medias parecidas y matrices de covarianza iguales casi singular para distintos tamaños de muestra. Se muestra la brecha entre el riesgo de Bayes y el de los clasificadores

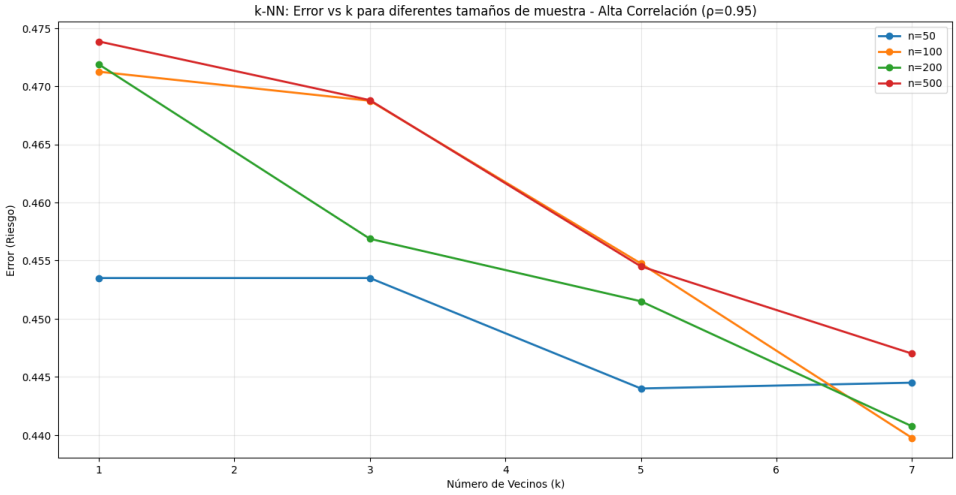

Figura 8: Comparación del riesgo de Bayes de los distintos clasificadores de k-NN para clases balanceadas, medias parecidas y matrices de covarianza iguales casi singular para distintos valores de k.

Notemos que en este, como en el primero, los clasificadores de LDA, QDA y Naïve Bayes son los que menos riesgo presentan. También llama la atención que los clasificadores de k-NN no tienen un riesgo que se acerque a 0.

3.5. Escenario 5: clases balanceadas, medias parecidas y matrices de covarianza iguales

Para este escenario, se simularon muestras con $\pi_y = (1/2, 1/2)$ de un vector tal que

$$(X|Y=0) \sim N((0,0)^T, \Sigma) \quad \text{y} \quad (X|Y=1) \sim N((2/5, 2/5)^T, \Sigma), \quad \text{donde} \quad \Sigma = \begin{pmatrix} 1 & 3/10 \\ 3/10 & 1 \end{pmatrix}.$$

Para mostrar de manera sintética el rendimiento de los distintos clasificadores respecto al estimador óptimo, se presenta en la figuras 9 la comparación del riesgo de Bayes para cada uno de los clasificadores respecto al tamaño de muestra y en la figura 10 la comparación del riesgo de k-NN para distintos valores de k.

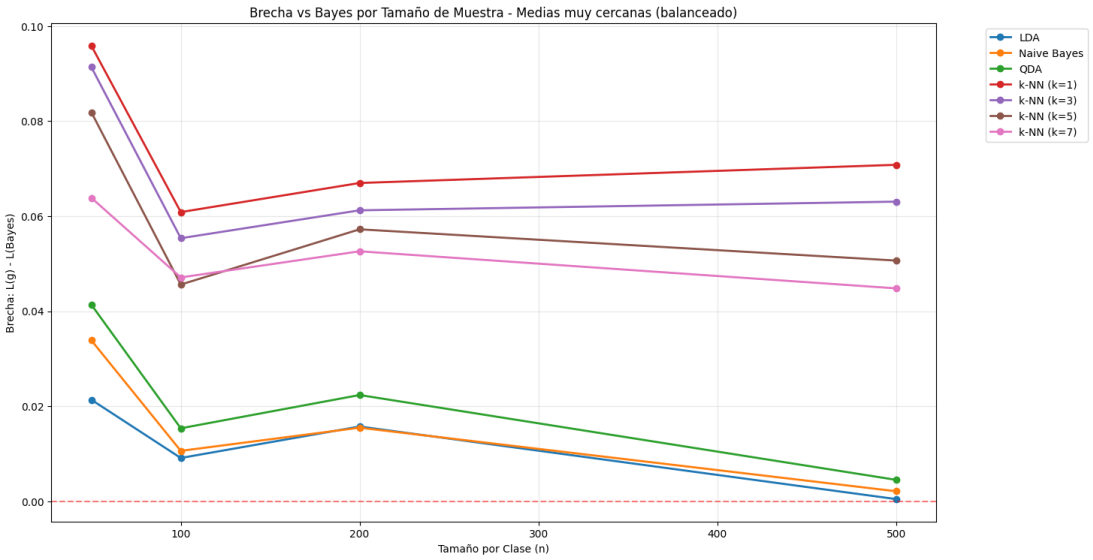

Figura 9: Comparación del riesgo de Bayes de los distintos clasificadores para el caso de clases balanceadas, medias parecidas y matrices de covarianza iguales para distintos tamaños de muestra. Se muestra la brecha entre el riesgo de Bayes y el de los clasificadores

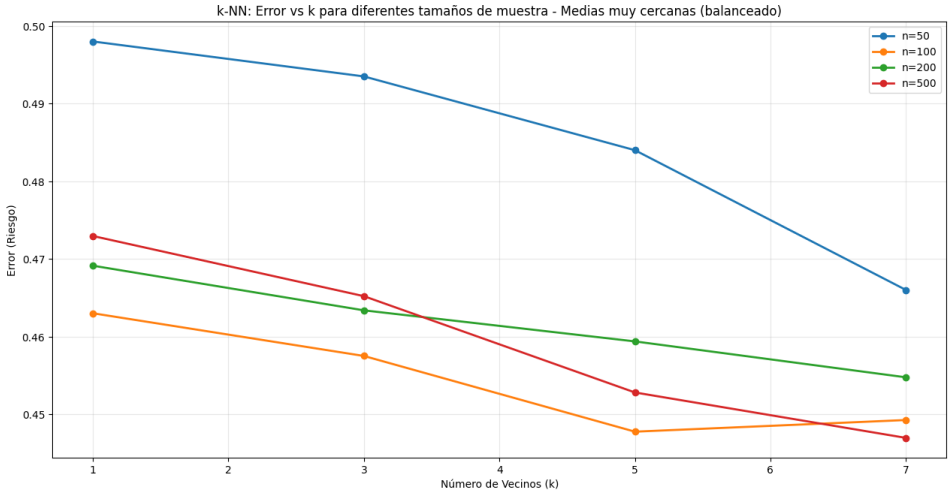

Figura 10: Comparación del riesgo de Bayes de los distintos clasificadores de k-NN para clases balanceadas, medias parecidas y matrices de covarianza iguales para distintos valores de k.

En este caso, notemos que como intuitivamente una muestra es una traslación pequeña de la otra, los clasificadores LDA, QDA y Naïve Bayes necesitan tamaños de muestra más grandes para poder identificar las distintas clases.

3.6. Escenario 6: clases desbalanceadas, medias parecidas y matrices de covarianza iguales

Para este escenario, se simularon muestras con $\pi_y = (4/5, 1/5)$ de un vector tal que

$$(X|Y=0) \sim N((0,0)^T, \Sigma) \quad \text{y} \quad (X|Y=1) \sim N((1/2,1/2)^T, \Sigma), \quad \text{donde} \quad \Sigma = \begin{pmatrix} 1 & 3/10 \\ 3/10 & 1 \end{pmatrix}.$$

Para mostrar de manera sintética el rendimiento de los distintos clasificadores respecto al estimador óptimo, se presenta en la figuras 11 la comparación del riesgo de Bayes para cada uno de los clasificadores respecto al tamaño de muestra y en la figura 12 la comparación del riesgo de k-NN para distintos valores de k.

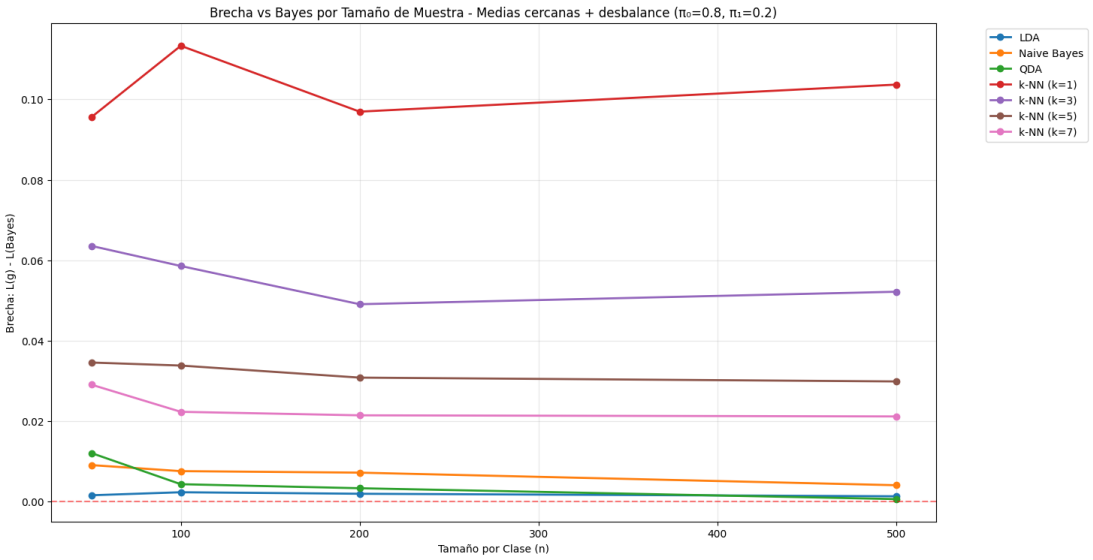

Figura 11: Comparación del riesgo de Bayes de los distintos clasificadores para el caso de clases desbalanceadas, medias parecidas y matrices de covarianza iguales para distintos tamaños de muestra. Se muestra la brecha entre el riesgo de Bayes y el de los clasificadores

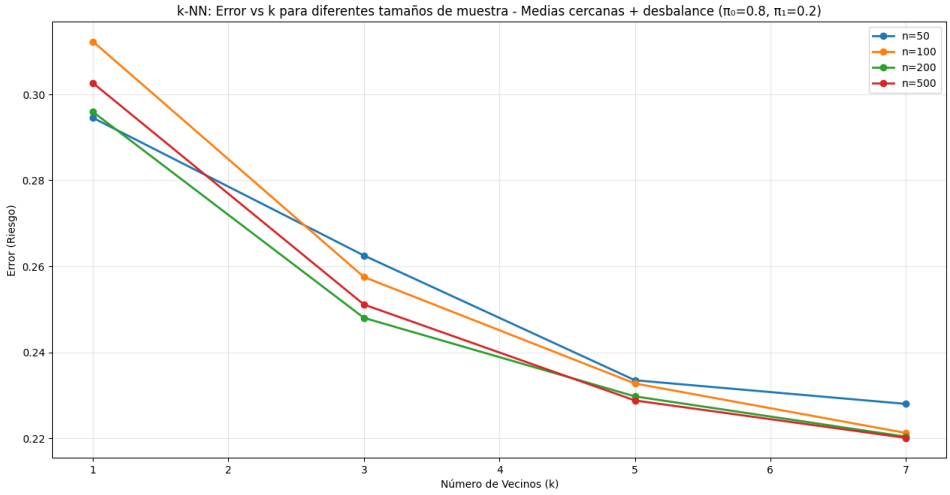

Figura 12: Comparación del riesgo de Bayes de los distintos clasificadores de k-NN para clases desbalanceadas, medias parecidas y matrices de covarianza iguales para distintos valores de k.

En este caso, notemos que como intuitivamente la muestra pequeña es una traslación pequeña de la otra, pero a pesar de ello, los clasificadores LDA, QDA y Naïve Bayes no tienen errores de Bayes de importancia. .

4. Conclusiones y discusión

La principal observación que se puede hacer de los seis escenarios presentados es que los clasificadores que se derivan de hipótesis de normalidad son aquellos con menor riesgo de Bayes. Esto es natural ya que los datos que se manejaron fueron generados a partir de una normal. Sobre los estimadores de k-NN, se aprecia también el hecho de que, conforme el tamaño de la muestra crece, la selección razonable de k también debería crecer para disminuir el riesgo de Bayes en la clasificación. Otra observación importante sobre el caso de k-NN es que cuando las medias son parecidas, el riesgo disminuye más lentamente. Esto podría deberse a que se llega a un efecto de homogeneidad/balance local de las clases.

Podría resultar interesante repetir este experimento de simulación con puntos generados por mecanismos distintos como procesos Poisson espaciales o distribuciones bivariadas más irregulares. Con estas estructuras de simulación podría determinarse la forma del clasificador de Bayes óptimo en cada contexto.

Referencias

[DGL96] Luc Devroye, László Györfi y Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. New York: Springer, 1996. DOI: 10.1007/978-1-4612-0711-5. URL: https://link.springer.com/book/10.1007/978-1-4612-0711-5.

[HTF09] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.