Análisis computacional del Riesgo de clasificadores mediante simulaciones

INTRODUCCIÓN A CIENCIA DE DATOS

01 de Octubre de 2025

Jessica Rubí Lara Rosales jessica.lara@cimat.mx
Iván García Mestiza ivan.garcia@cimat.mx
Luis Erick Palomino Galván luis.palomino@cimat.mx

Introducción

En este trabajo estudiaremos el riesgo de clasificación de varios métodos (Naive Bayes, LDA, QDA y k-NN) frente al riesgo del clasificador óptimo de Bayes. Para ello, se diseñarán y ejecutarán simulaciones con datos sintéticos, evaluando el desempeño de cada método frente al riesgo óptimo y frente a estimadores obtenidos mediante técnicas de validación. Este enfoque permitirá analizar cómo los distintos clasificadores se aproximan al desempeño teórico ideal y cómo varía su eficiencia en escenarios prácticos de entrenamiento y prueba. Mostraremos de manera resumida cómo se calculan los discriminantes para los modelos trabajados y luego haremos una presentación de resultados de manera gráfica. Con lo anterior podremos observar que al incrementar el tamaño de muestra para datos sintéticos, el riesgo de clasificación de los distintos modelos tiende al riesgo óptimo de Bayes, aunque unos se acercan más lentamente que otros. En el caso particular del k-NN, resaltaremos la importancia de un "buen balance" entre k y n para obtener mejores clasificadores.

Exploración visual

Consideremos un escenario sintético con dos clases $Y \in \{0, 1\}$ con probabilidades *a priori* π_0 y $\pi_1 = 1 - \pi_0$. Las observaciones condicionales se generan como:

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

donde $\mu_0, \mu_1 \in \mathbb{R}^p$ y $\Sigma_0, \Sigma_1 \in \mathbb{R}^p \times \mathbb{R}^p$ representan las medias y matrices de covarianza de cada clase. Una pregunta interesante es ¿qué pasa si las covarianzas son iguales? Intuitivamente, la frontera óptima es una línea recta. Como LDA está diseñado específicamente para encontrar la mejor frontera lineal posible, la solución coincide con la solución óptima de Bayes. De hecho, si las covarianzas son iguales ($\Sigma_0 = \Sigma_1$), entonces el clasificador LDA es el mismo que el de Bayes, mientras que si las covarianzas son distintas, la frontera óptima es una curva, y en este caso el clasificador QDA es el mismo que el de Bayes. Estos resultados fueron vistos en clase y se encuentran en [1] y [2]

Notemos que la distancia entre los centros de las clases y el tamaño y forma de los datos nos pueden crear casos fáciles y difíciles de clasificar. Idealmente, nubes de puntos que están muy lejos una de otra y que minimicen la superposición, es un escenario fácil. Esto ocurre cuando las medias μ_k son muy distantes entre sí y cuando se usan matrices de covarianza con varianzas pequeñas, lo que hace que las nubes de puntos sean densas. Por otro lado, si las medias están muy cerca una de la otra, o las medias están separadas pero las varianzas son muy grandes, o se usan covarianzas distintas y orientadas de forma conflictiva, entonces estamos frente a un caso difícil.

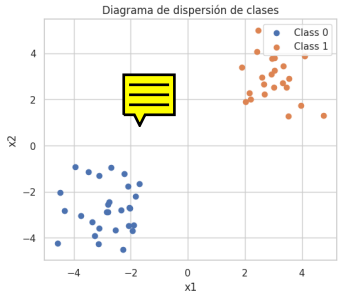

Figura 1: Diagrama de dispersión, caso fácil para $n = 50$.

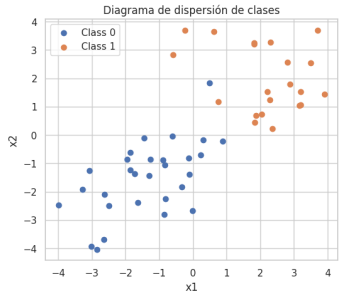

Figura 2: Diagrama de dispersión, caso medio para $n = 50$.

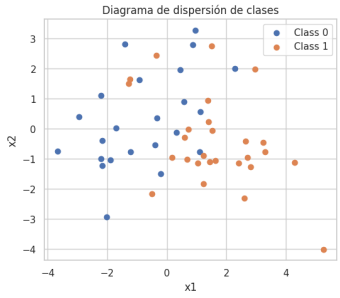

Figura 3: Diagrama de dispersión, caso difícil para $n = 50$.

Para el diagrama de la Figura 1, usamos medias muy separadas en el espacio y covarianzas iguales y pequeñas, por lo que cualquier clasificador debería obtener un rendimiento casi perfecto, ya que la frontera de decisión es muy evidente. Para el diagrama de la Figura 2 usamos medias más cercanas pero con covarianzas un poco más grandes y con algo de correlación, por lo que en este caso esperaríamos unos pocos errores en los clasificadores, pero no demasiados. Por último, para el diagrama de la Figura 3, usamos medias muy cercanas y covarianzas grandes y con orientación opuesta, lo que causa que la zona de mayor densidad de una clase se proyecte en la otra.

Veamos cómo analizar un ejemplo de un caso medio. Para la simulación, se generaron 50 observaciones, asignando cada muestra a la clase 0 con probabilidad $\pi_0 = 0.5$ o a la clase 1 con probabilidad $\pi_1 = 0.5$. Cada observación fue generada de manera independiente siguiendo las distribuciones mencionadas, inicialmente con los parámetros

$$\mu_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 2 & -1 \\ -1 & 4 \end{bmatrix}. \tag{1}$$

Un ejemplo de datos generados con estas distribuciones se muestra en la Figura 4.

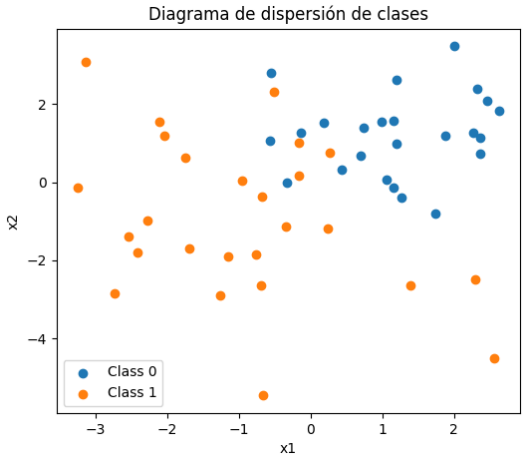

Figura 4: Diagrama de dispersión de clases para $n = 50$ con covarianzas distintas.

Este conjunto de datos lo dividimos en un 70 % para entrenamiento y un 30 % para prueba mediante el uso de la función *stratify*. Luego, se implementaron cuatro tipos de clasificadores: **Naive Bayes Gaussiano**, que parte del supuesto de independencia condicional entre las características; Análisis Discriminante Lineal (**LDA**), que asume covarianzas iguales entre clases y genera una frontera de decisión lineal; Análisis Discriminante Cuadrático (**QDA**),

que permite covarianzas distintas y en consecuencia fronteras de decisión cuadráticas; y **k-NN**, que es un método no paramétrico basado en la noción de distancia de puntos en el espacio.

Para la comparación de los clasificadores, graficamos las fronteras de decisión de cada clasificador sobre el conjunto de prueba para interpretar cómo el modelo separa las clases. Dichas gráficas corresponden a las Figuras 5 a 8.

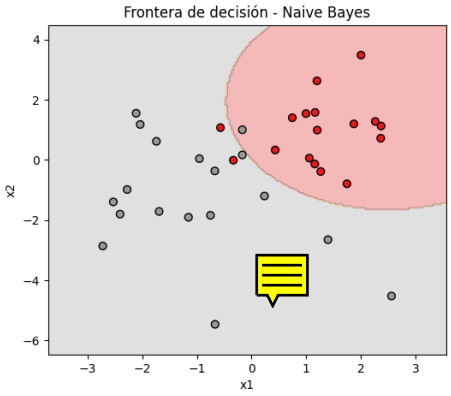

Figura 5: Frontera de decisión de Naive Bayes para $n = 50$.

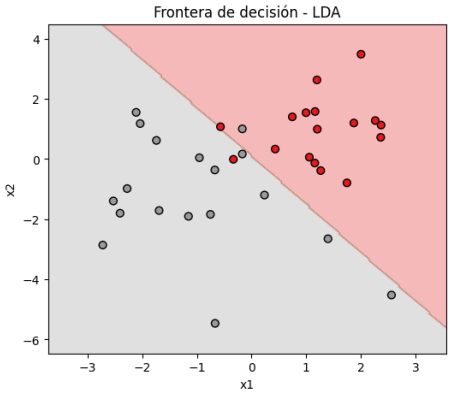

Figura 6: Frontera de decisión de LDA para $n = 50$.

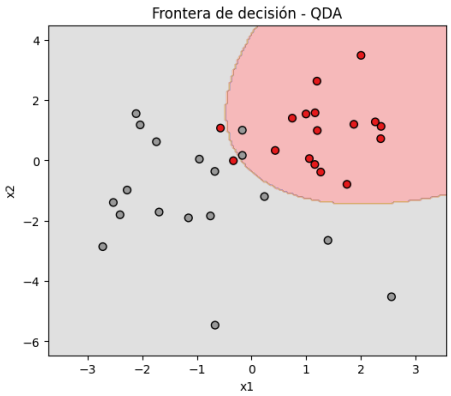

Figura 7: Frontera de decisión de QDA para $n = 50$.

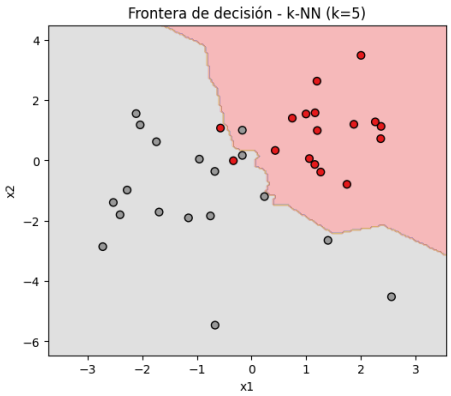

Figura 8: Frontera de decisión de k-NN para $n = 50$.

A simple vista podríamos concluir que todos los modelos hacen un buen trabajo, pero es conveniente usar métricas de evaluación para hacer conclusiones más precisas. En la Figura 9 se muestran las matrices de confusión respectivas para los modelos.

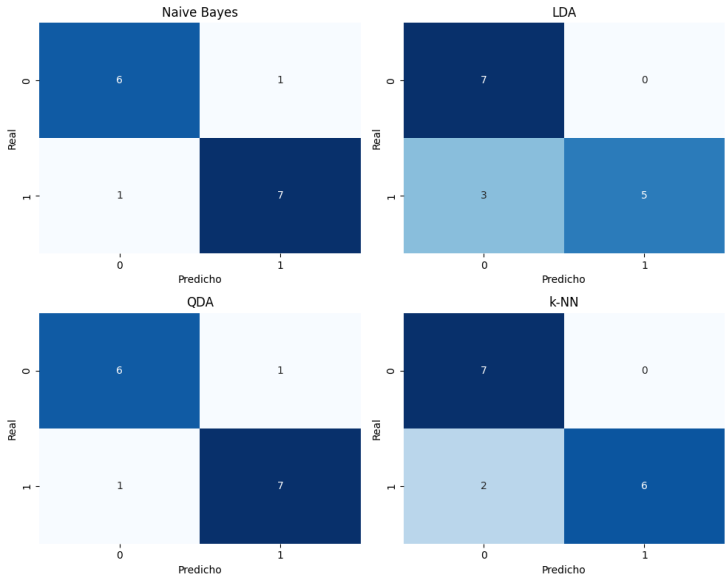

Figura 9: Matrices de confusión de los clasificadores Naive Bayes, LDA, QDA y k-NN para una muestra de 50 observaciones de covarianzas distintas.

Como el tamaño de muestra es chico, y el conjunto de prueba consiste de únicamente 15 datos, saber cuál es el mejor clasificador en este caso es un poco difuso, por lo que posteriormente haremos estos análisis para tamaños de muestra más grandes. Sin embargo, podemos usar como línea base para comparar el clasificador Óptimo de Bayes, puesto que en este caso conocemos las distribuciones exactas de donde provienen los datos. Los resultados se presentan en las Figuras 10 y 11. Como se había mencionado previamente, en el caso de $\Sigma_0 \neq \Sigma_1$, el clasificador QDA coincide con el de Bayes, lo cual se aprecia en dichas figuras.

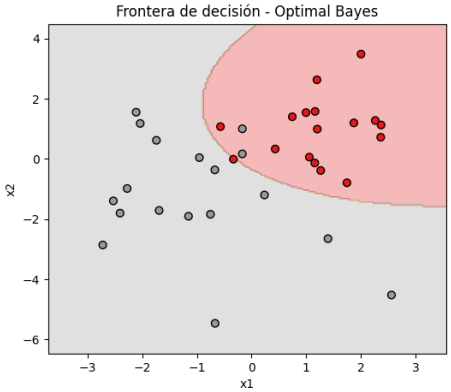

Figura 10: Frontera de decisión de *Optimal Bayes*.

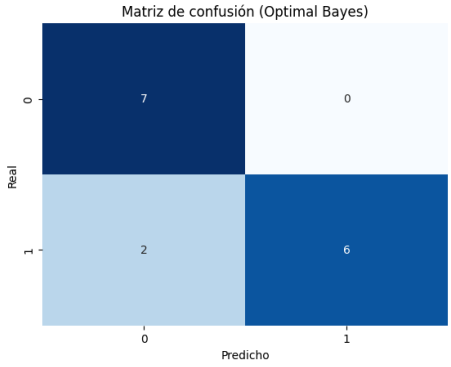

Figura 11: Matrices de confusión del clasificador *Optimal Bayes* para una muestra de 50 observaciones.

Como podemos observar, el modelo k-NN con $k = 5$ tiene la misma matriz de confusión que el clasificador óptimo. Sin embargo, el resto de modelos tienen matrices de predicción también muy similares, y en este caso en particular no podemos descartar a alguno simplemente con estos criterios. Por último, considerando el modelo Weighted k-NN, se tienen los resultados de las Figuras 12 y 13, que muestran que este también es un estimador bueno, pero las regiones en las que se dividen los datos no se ven tan parecidas a las del estimador óptimo y tienen formas más irregulares, por lo que podría ser mejor considerar los modelos anteriores.

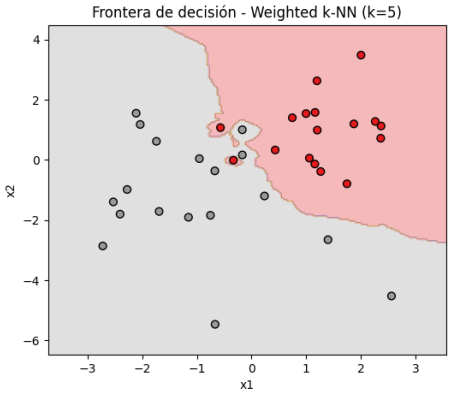

Figura 12: Frontera de decisión de Weighted k-NN.

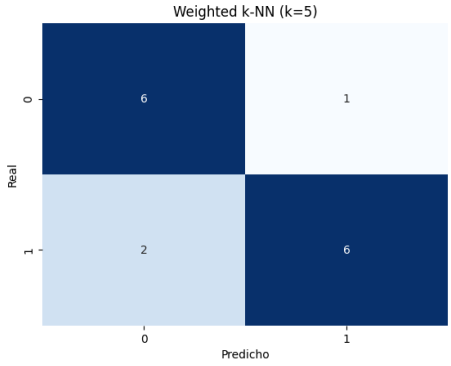

Figura 13: Matrices de confusión del clasificador Weighted k-NN para una muestra de 50 observaciones.

Antes de buscar formalizar de alguna manera los análisis anteriores, mostraremos un ejemplo de datos con covarianzas iguales, y tomemos la covarianza común igual a Σ_0 de (1), manteniendo las mismas medias que en dicha ecuación. Un ejemplo de datos generados con estas distribuciones se muestra en la Figura 14.

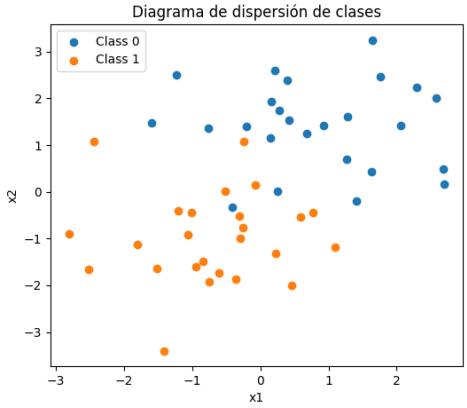

Figura 14: Diagrama de dispersión de clases para $n = 50$ con covarianzas iguales.

Para la comparación de los clasificadores, graficamos las fronteras de decisión de cada clasificador sobre el conjunto de prueba para interpretar cómo el modelo separa las clases. Dichas gráficas corresponden a las Figuras 15 a 18.

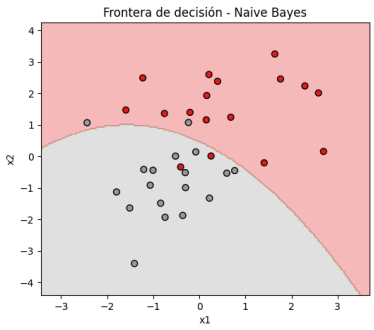

Figura 15: Frontera de decisión de Naive Bayes para $n = 50$.

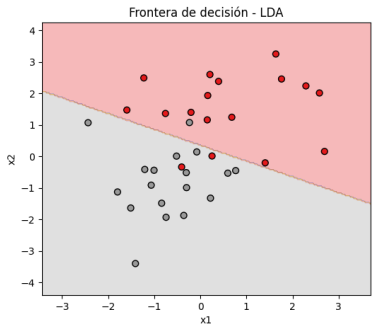

Figura 16: Frontera de decisión de LDA para $n = 50$.

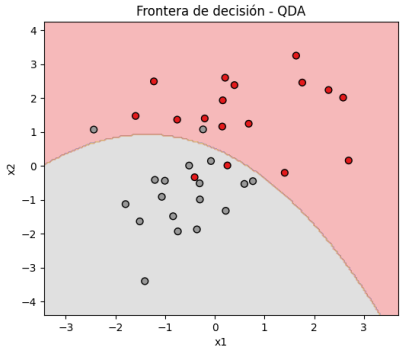

Figura 17: Frontera de decisión de QDA para $n = 50$.

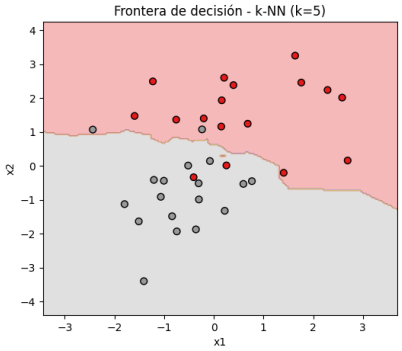

Figura 18: Frontera de decisión de k-NN para $n = 50$.

A simple vista podríamos concluir que todos los modelos hacen un buen trabajo, pero es conveniente usar métricas de evaluación para hacer conclusiones más precisas. En la Figura 19 se muestran las matrices de confusión respectivas para los modelos.

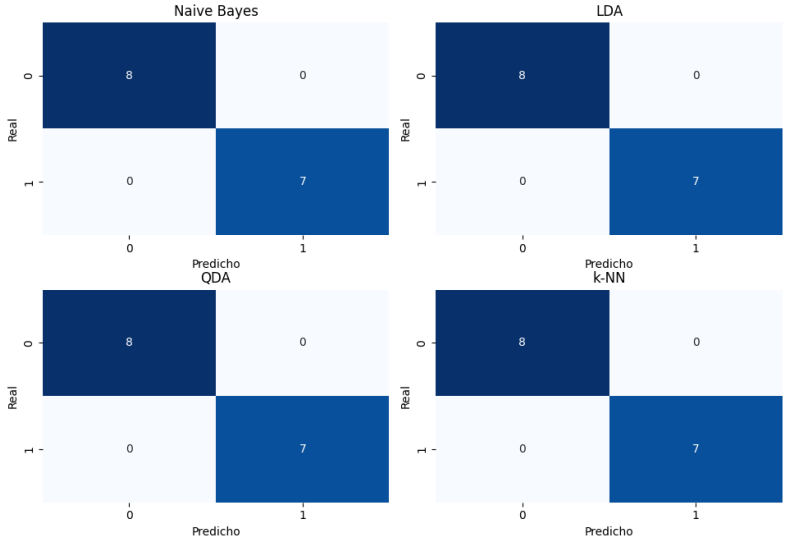

Figura 19: Matrices de confusión de los clasificadores Naive Bayes, LDA, QDA y k-NN para una muestra de 50 observaciones de covarianzas iguales.

Nuevamente podemos usar como línea base para comparar el clasificador Óptimo de Bayes, puesto que en este caso conocemos las distribuciones exactas de donde provienen los datos. Los resultados se presentan en las Figuras 20 y 21. Como se había mencionado previamente, en el caso de $\Sigma_0 = \Sigma_1$, el clasificador LDA coincide con el de Bayes, lo cual se aprecia en dichas figuras.

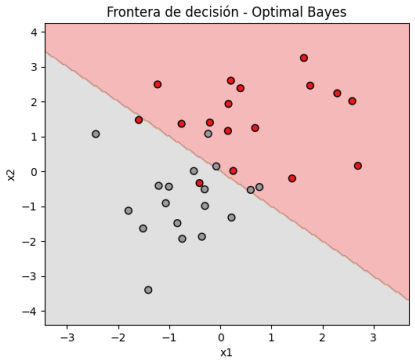

Figura 20: Frontera de decisión de *Optimal Bayes*.

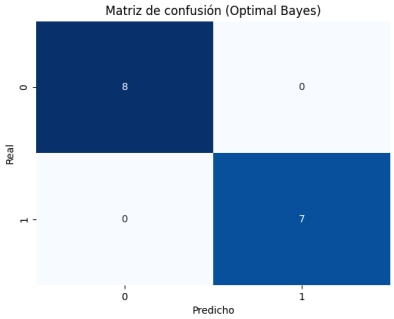

Figura 21: Matrices de confusión del clasificador *Optimal Bayes* para una muestra de 50 observaciones.

Como podemos observar, todos los modelos clasifican muy bien y lo hacen con la misma precisión que el Óptimo de Bayes, y resulta sorprendente que en el conjunto de prueba todos lograron los mismos resultados. Sin embargo, esto puede deberse a la poca cantidad de datos que hay en dicho conjunto, así que, de nuevo, este análisis no es suficiente para concluir si alguno de ellos es el mejor. Por último, considerando el modelo Weighted k-NN, se tienen los resultados de las Figuras 22 y 23, que muestran que este también es un estimador bueno, y a diferencia del caso de covarianzas distintas, las regiones no se ven demasiado irregulares.

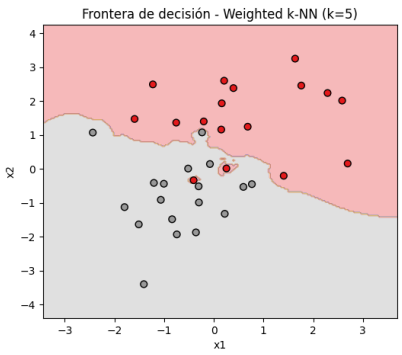

Figura 22: Frontera de decisión de Weighted k-NN.

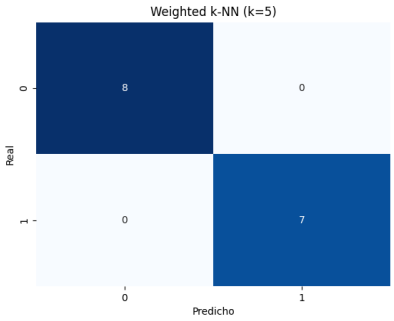

Figura 23: Matrices de confusión del clasificador Weighted k-NN para una muestra de 50 observaciones.

En los dos ejemplos anteriores vimos que en general casi todos los clasificadores son buenos cuando hay aproximadamente la misma proporción de datos de las distintas clases. Sin embargo, podríamos preguntarnos cómo cambiarían los resultados si se cambiara la distribución *a priori*. Para hacer este análisis, simularemos datos de tal manera que $\pi_0 = 0.2$ y $\pi_1 = 0.8$, conservando el resto de parámetros de (1).

Un ejemplo de datos generados con estas distribuciones se muestra en la Figura 24.

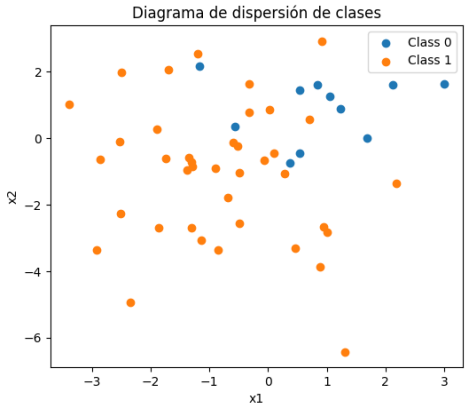

Figura 24: Diagrama de dispersión de clases para $n = 50$ con datos desproporcionados.

Para la comparación de los clasificadores, graficamos las fronteras de decisión de cada clasificador sobre el conjunto de prueba para interpretar cómo el modelo separa las clases. Dichas gráficas corresponden a las Figuras 25 a 28.

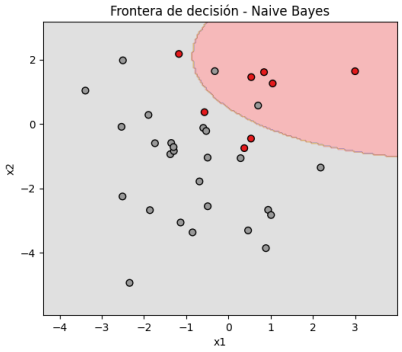

Figura 25: Frontera de decisión de Naive Bayes para $n = 50$.

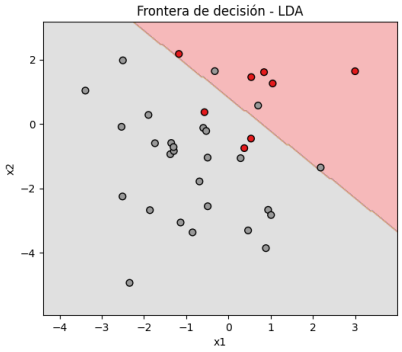

Figura 26: Frontera de decisión de LDA para $n = 50$.

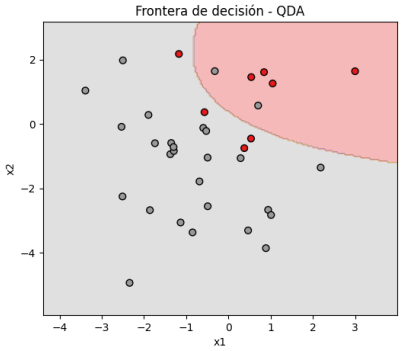

Figura 27: Frontera de decisión de QDA para $n = 50$.

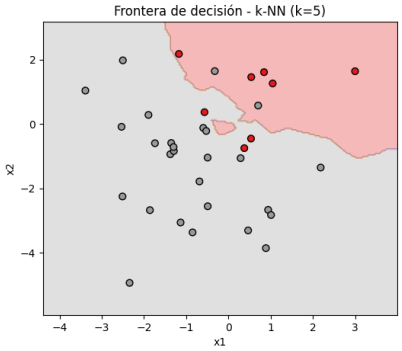

Figura 28: Frontera de decisión de k-NN para $n = 50$.

En este caso no parece muy buena la clasificación que hacen los estimadores, y para evaluarlo en la Figura 29 se muestran las matrices de confusión respectivas para los modelos.

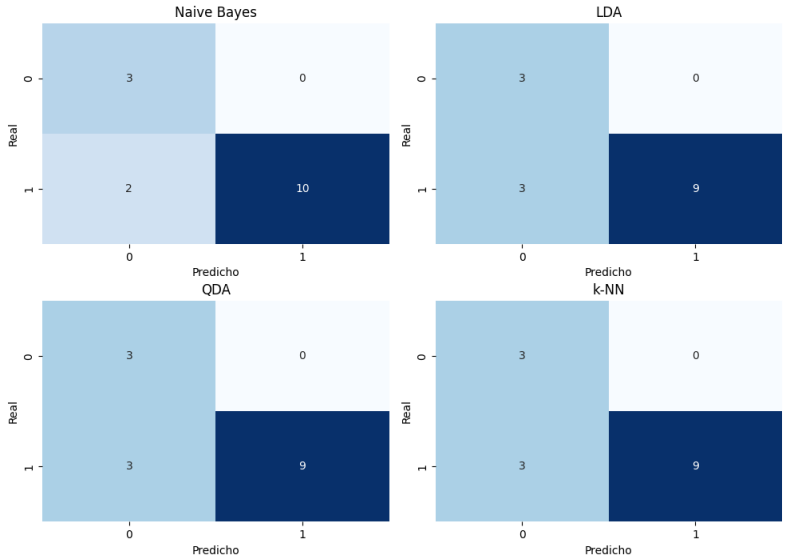

Figura 29: Matrices de confusión de los clasificadores Naive Bayes, LDA, QDA y k-NN para una muestra de 50 con datos desproporcionados.

Como en los análisis previos, compararemos dichos clasificadores con el Óptimo de Bayes. Los resultados se presentan en las Figuras 30 y 31.

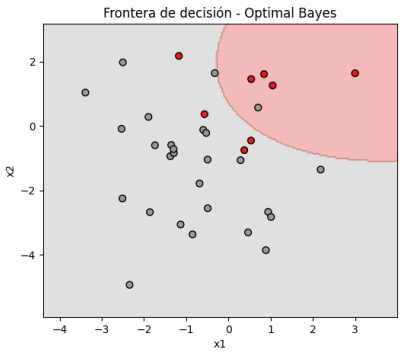

Figura 30: Frontera de decisión de *Optimal Bayes*.

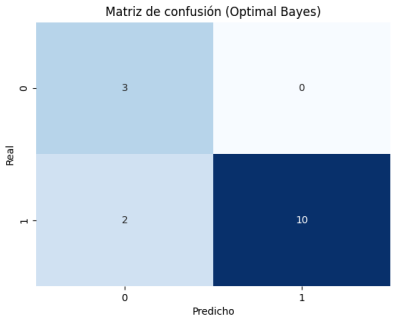

Figura 31: Matrices de confusión del clasificador *Optimal Bayes* para una muestra de 50 observaciones.

Como podemos observar, los modelos clasifican de manera parecida al Óptimo de Bayes, aunque los resultados no son tan buenos, pues en este caso a los modelos le cuesta más aprender sobre los datos de los que hay poca proporción. Por último, considerando el modelo Weighted k-NN, se tienen los resultados de las Figuras 32 y 33, que muestran de nuevo regiones bastante irregulares y con huecos dentro de las distintas áreas.

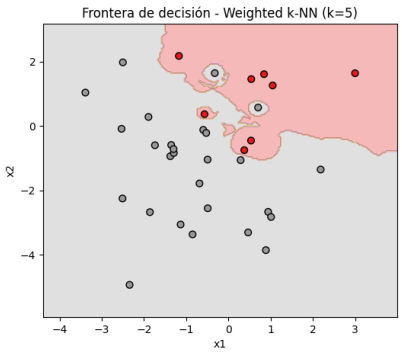

Figura 32: Frontera de decisión de Weighted k-NN.

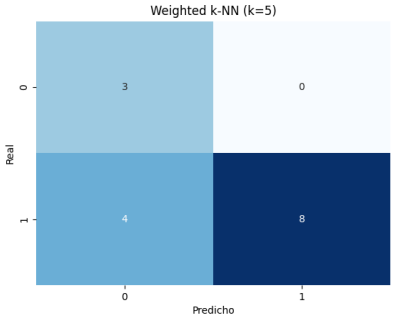

Figura 33: Matrices de confusión del clasificador Weighted k-NN para una muestra de 50 observaciones.

Discriminantes de clasificación en diversos modelos

Para poder comparar qué tan buenos son los clasificadores en términos teóricos y prácticos, evaluaremos los riesgos de clasificación mediante diversas simulaciones. Para ello presentaremos brevemente cómo se calculan dichos riesgos. Recordemos que el modelo en el que estamos trabajando consiste en una distribución *a priori* (π_0, π_1), y con

$$X \mid Y = g \sim N(\mu_g, \Sigma_g), \qquad g \in \{0, 1\},$$

Luego, la regla de decisión basada en verosimilitudes (o bien, regla de Bayes) compara los discriminantes

$$\delta_g(x) = \ln(\pi_g) + \ln f_g(x),$$

y asigna x a la clase que tenga mayor $\delta_g(x)$, en donde f_g representa la densidad correspondiente de la normal considerada.

Por otro lado, como en este caso las distribuciones de donde provienen los datos son conocidas, se puede calcular el riesgo óptimo usando la regla de Bayes. Para ello,

Para una regla $h : \mathbb{R}^d \to \{0, 1\}$ el riesgo de clasificación es

$$R(h) = \mathbb{P}(h(X) \neq Y) = \pi_0 \int \mathbf{1}\{h(x) = 1\} f_0(x)\, dx + \pi_1 \int \mathbf{1}\{h(x) = 0\} f_1(x)\, dx.$$

La regla de Bayes h^* minimiza $R(h)$, y consiste en comparar las distribuciones posteriores:

$$h^*(x) = \begin{cases} 1 & \text{si } \eta(x) := \mathbb{P}(Y = 1 \mid X = x) = \dfrac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} > \frac{1}{2}, \\ 0 & \text{en otro caso.} \end{cases}$$

El riesgo óptimo de Bayes se expresa como

$$R^* = R(h^*) = \mathbb{E}\big[\min\{\eta(X), 1 - \eta(X)\}\big].$$

Por otro lado, en el caso Naive Bayes se asume independencia condicional, por lo que Σ_g es una matriz diagonal, y en este caso se tiene que, si

$$x_j \mid Y = g \sim N(\mu_{gj}, \sigma^2_{gj}), \qquad j = 1, \ldots, d,$$

entonces la densidad conjunta de la muestra se factoriza como $f_g(x) = \prod_{j=1}^{d} f_{g,j}(x_j)$, en donde

$$f_{g,j}(x_j) = \frac{1}{\sqrt{2\pi\sigma_{gj}^2}} \exp\left(-\frac{(x_j - \mu_{gj})^2}{2\sigma_{gj}^2}\right).$$

En este caso el discriminante logarítmico está dado por

$$L_g(x) = \ln(\pi_g) + \sum_{j=1}^{d} \ln f_{g,j}(x_j) = \ln(\pi_g) - \frac{1}{2}\sum_{j=1}^{d}\left[\ln(2\pi\sigma_{gj}^2) + \frac{(x_j - \mu_{gj})^2}{\sigma_{gj}^2}\right].$$

Por otra parte, para calcular el riesgo del clasificador LDA (Linear Discriminant Analysis), si $\Sigma_0 = \Sigma_1 = \Sigma$ se tiene que

$$L_g(x) = x^T\Sigma^{-1}\mu_g - \tfrac{1}{2}\mu_g^T\Sigma^{-1}\mu_g + \ln(\pi_g).$$

Definiendo $w = \Sigma^{-1}(\mu_1 - \mu_0)$, la clasificación lineal puede escribirse como

$$\text{decidir } 1 \iff w^T x > c, \qquad c = \tfrac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) - \ln\Big(\frac{\pi_1}{\pi_0}\Big).$$

En el caso del clasificador QDA (Quadratic Discriminant Analysis), si $\Sigma_0 \neq \Sigma_1$ entonces

$$L_g(x) = -\tfrac{1}{2}\ln\det(\Sigma_g) - \tfrac{1}{2}(x - \mu_g)^T\Sigma_g^{-1}(x - \mu_g) + \ln(\pi_g).$$

En este caso la frontera de decisión es cuadrática, pero no existe una expresión analítica cerrada para calcular el riesgo, así que en la práctica hay que usar métodos numéricos o realizar diversas repeticiones y estimar el riesgo por validación cruzada o bootstrap.

Por otro lado, el criterio de Fisher busca una proyección en una dimensión $w \in \mathbb{R}^d$ que maximice la separabilidad entre dos clases:

$$J(w) = \frac{(w^T(\mu_1 - \mu_0))^2}{w^T S_w w}, \qquad S_w = \Sigma_0 + \Sigma_1,$$

en donde S_w es la matriz de dispersión intraclase. La dirección óptima de proyección es

$$w^* \propto S_w^{-1}(\mu_1 - \mu_0).$$

Tras proyectar $Z = w^{*T}X$ en esta dirección, se elige un umbral t (por ejemplo minimizando el error empírico) y el riesgo se evalúa como en LDA.

Medición computacional del riesgo de clasificación

Con el objetivo de formalizar los análisis hechos en la sección de Exploración visual y poder comparar los modelos no solamente con las matrices de confusión o las gráficas, sino usando las mediciones de riesgo de clasificación de la sección anterior, hicimos una evaluación del desempeño de dichos estimadores bajo distintos escenarios: consideramos un barrido de parámetros de

$$n \in \{50, 100, 200, 500\}, \quad k \in \{1, 3, 5, 11, 21\},$$

y con cada combinación de ellos creamos 20 repeticiones con el objetivo de estimar el riesgo de clasificación de cada método y compararlo con el riesgo óptimo. Cada uno de estos riesgos lo medimos de las siguientes maneras:

el verdadero, por validación cruzada, y por las reglas de Bootstrap .632 y .632+, obteniendo como resultados una tabla como en el Cuadro 1.

Modelo	n	Replicación	k	Riesgo Verdadero	Riesgo CV	Riesgo Bootstrap 0.632	Riesgo Bootstrap 0.632+
Optimal Bayes	50	0	NaN	0.096	0.10	0.104	0.232
Naive Bayes	50	0	NaN	0.104	0.12	0.124	0.256
LDA	50	0	NaN	0.116	0.12	0.139	0.267
QDA	50	0	NaN	0.098	0.10	0.119	0.255
Weighted k-NN	50	0	5	0.140	0.04	0.126	0.264
Fisher	50	0	NaN	0.118	0.12	0.144	0.267
k-NN	50	0	1	0.178	0.10	0.145	0.270

Cuadro 1: Comparación de riesgos para distintos modelos y parámetros.

Sin embargo, un estimador puntual no siempre da toda la información de los parámetros, por lo que también es conveniente calcular otras estadísticas descriptivas como la media y desviación estándar. La implementación se encuentra en el código adjunto, y podemos observar que, en el modelo kNN, para valores chicos de n, en realidad no importa mucho el valor de k escogido, y en ocasiones tienen un mejor desempeño los valores más chicos bajo las métricas anteriores. Esto es razonable, puesto que mientras menos datos haya, valores de k mayores implican revisar una proporción más grande de la muestra, lo cual hace que las clasificaciones sean más globales y menos locales, y esto no es conveniente para muestras desproporcionadas.

Por otra parte, para valores más grandes de n se hace evidente la mejora al tomar mayores valores de k, aunque no siempre los más grandes son los mejores. Esto se debe a que, por lo visto en clases, para acercarse al riesgo óptimo, el valor de k debería ser tal que $k/n \to 0$ cuando $n \to \infty$, de manera que debe haber un equilibrio entre k y n.

Para facilitar la interpretación de los resultados, en la Figura 34 se muestra una gráfica de barras de errores sobre la evolución de $L(g)$ vs n para los distintos modelos considerados. Para cada punto se grafica su valor estimado y se coloca una barra vertical centrada en dicho punto que representa el error de dicha estimación (que medimos con la desviación estándar).

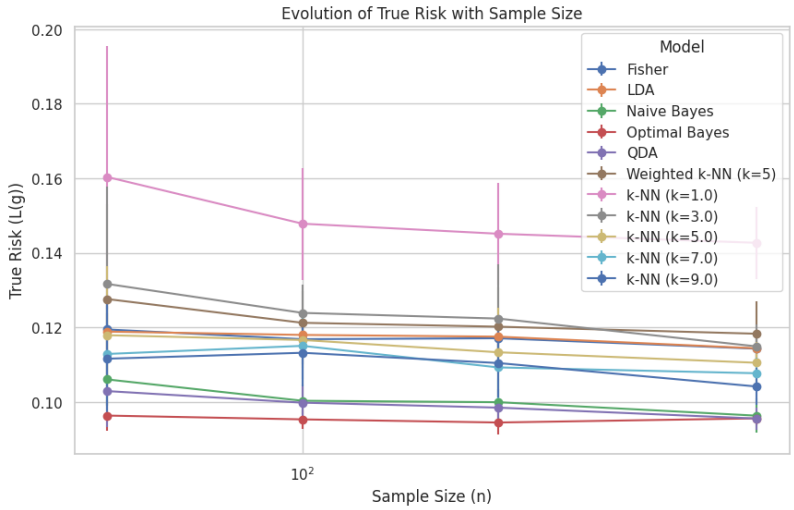

Figura 34: Evolución $L(g)$ vs n para distintos modelos.

Como podemos notar, los modelos más alejados del óptimo de Bayes son los k-NN con valores de k chicos, y lo modelos QDA y Naive Bayes se comportan muy bien en el caso presentado, en donde se tomaron distintas matrices de covarianza. Podemos notar además que en general la mayoría de los modelos presenta una tendencia hacia abajo, es decir, tienden a parecerse al Clasificador Óptimo de Bayes, aunque algunos lo hacen con mayor lentitud que otros. Esto es razonable, puesto que mientras más información, los clasificadores pueden aprender más de los datos. Nótese que esto solo es válido cuando se tiene más información de manera estructurada, y en la práctica no siempre es el caso, puesto que muchas veces agregar información representa introducir ruido que no se desea.

Ahora bien, para enfocarnos en el modelo k-NN, en la Figura 35 se muestra una gráfica de barras de errores sobre la evolución de $L(g)$ vs n para distintos valores de k en k-NN. Como podemos observar, manteniendo k fijo, el aumentar el tamaño de muestra ayuda a disminuir el error hasta cierto punto, pero después se nota una tendencia hacia arriba en todos los casos. En efecto, esto es de esperarse, puesto que valores muy chicos de k con respecto a n hacen que el kNN se vuelva muy local, y valores mayores hacen que obtenga información de una mayor proporción de la muestra, lo cual inclina el modelo hacia la población con mayor proporción.

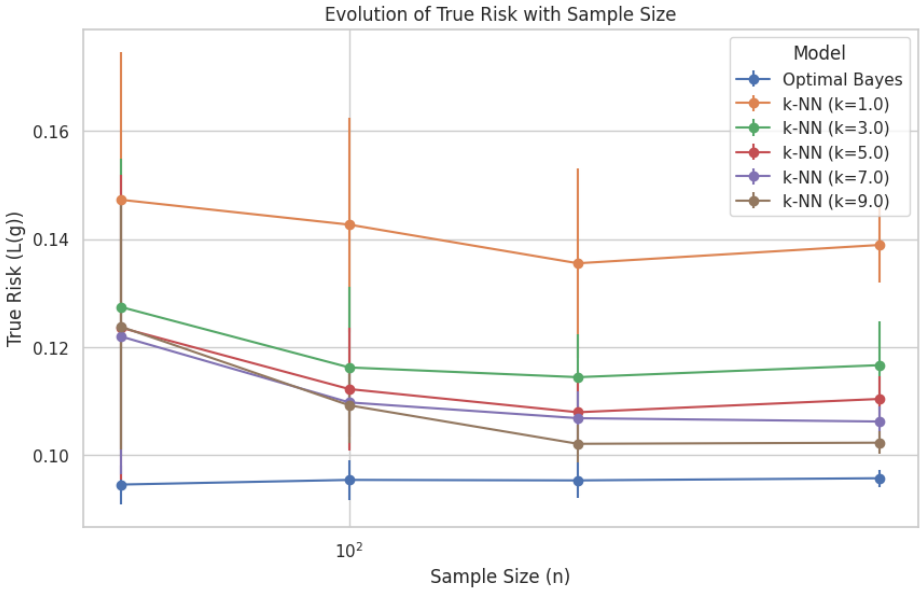

Figura 35: Evolución $L(g)$ vs n para distintos valores de k en kNN.

Por otro lado, la Figura 36 presenta la evolución de $L(g)$ vs k para distintos valores de n en k-NN, y nuevamente se puede apreciar el fenómeno anterior, en donde debe haber cierto equilibrio entre k y n para que el método k-NN clasifique de manera adecuada.

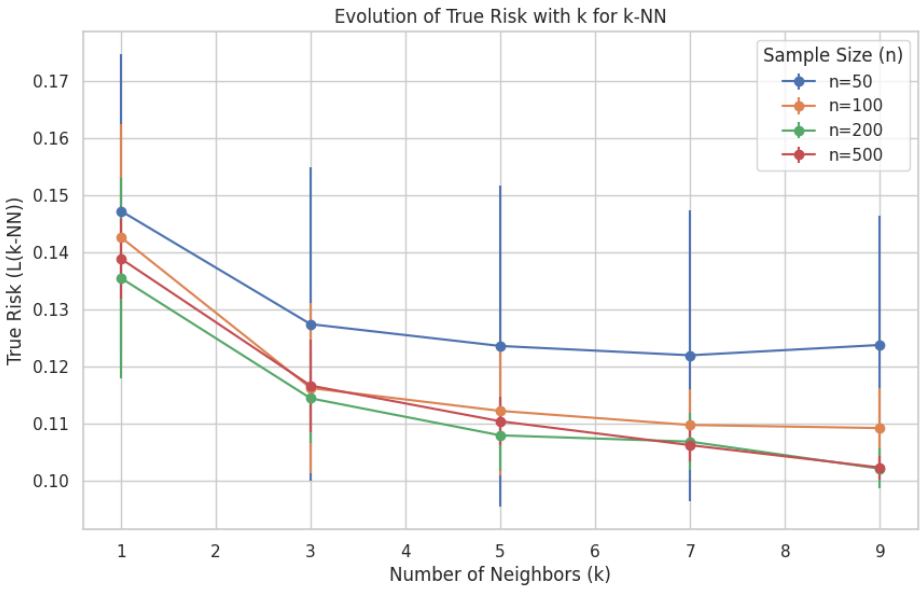

Figura 36: Evolución $L(g)$ vs k para distintos valores de k en kNN.

Ahora bien, para ayudar a la comparación de los errores y ver qué tan cerca están del clasificador óptimo los distintos modelos, en la Figura 37 se presentan las gráficas de brechas $L(\text{Bayes}) - L(g)$ vs n para los distintos modelos trabajados. En esta visualización es más clara una tendencia hacia arriba, que al estar graficando diferencias, indica que el valor de $L(g)$ tiende al de Bayes conforme se aumenta el tamaño de muestra.

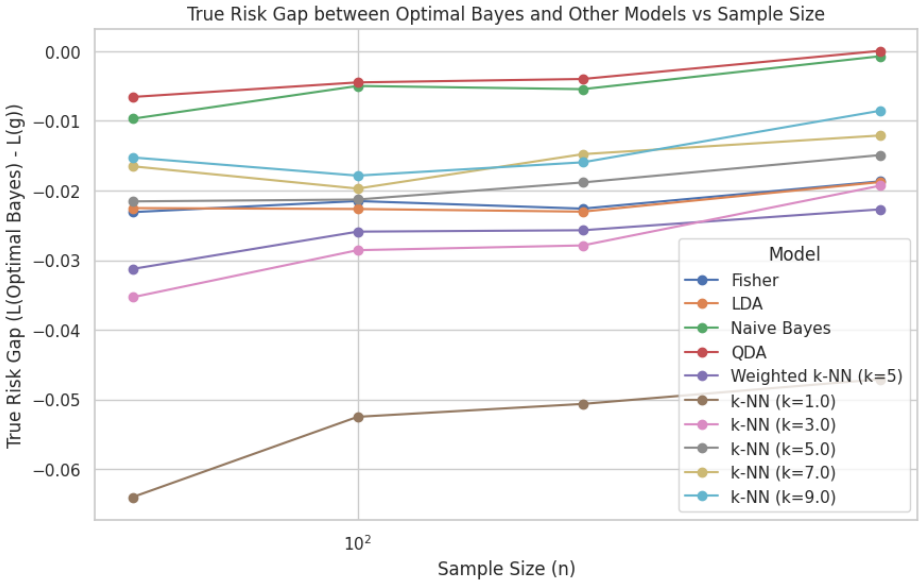

Figura 37: Brechas $L(\text{Bayes}) - L(g)$ vs n para los distintos modelos.

Los cálculos anteriores para las estimaciones de los riesgos de clasificación fueron realizados al hacer distintas replicaciones de muestras (20 por cada combinación de parámetros) y calcular los estimadores correspondientes, su media y desviación estándar. Sin embargo, esto hace que surja la necesidad de comparar el valor real de dichos errores contra el que ya estimamos. Con esto en mente, podemos calcular el riesgo por el método de Monte Carlo,

el cual garantiza que los resultados obtenidos tienden a parecerse con los verdaderos. En la Figura 38 se grafica el riesgo por el método Monte Carlo, que llamamos Riesgo Verdadero, contra el riesgo por validación cruzada que calculamos anteriormente para los distintos modelos. La línea diagonal punteada es la recta identidad, así que un punto arriba de ella significa que el modelo está sobreestimando el error, mientras que un punto por debajo significa que lo está subestimando. Sin embargo, como podemos notar, la mayoría caen muy cercanos a la recta identidad.

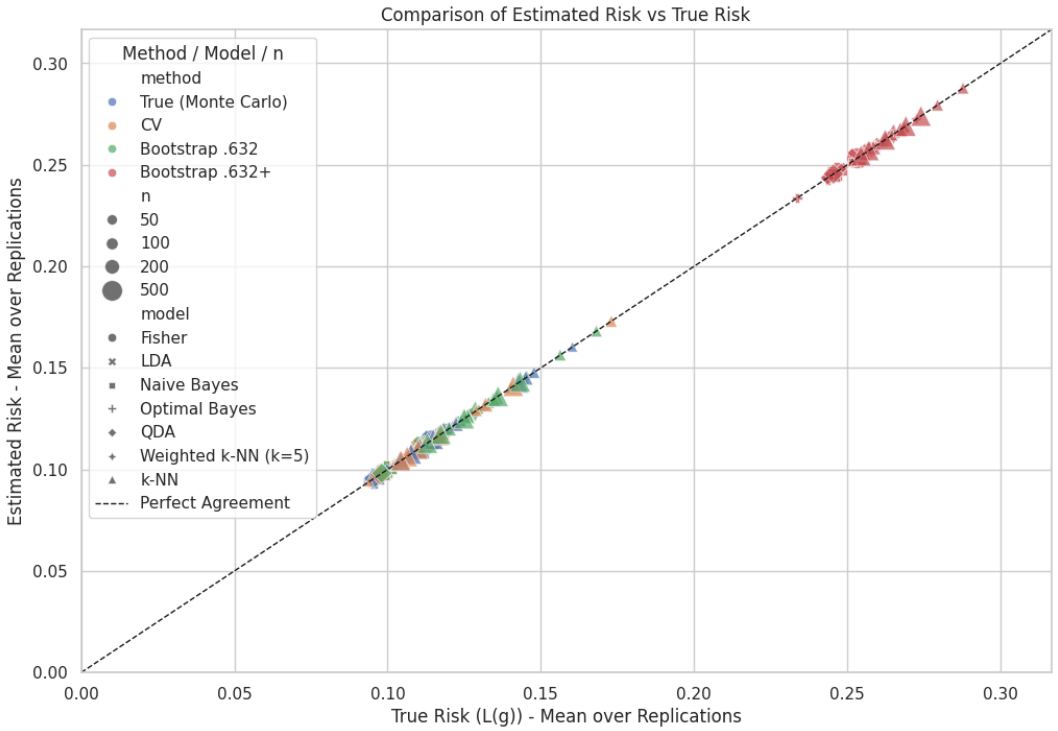

Figura 38: Riesgo estimado vs Verdadero de los distintos modelos.

Los tamaños de las figuras corresponden a los distintos valores de n, siendo las marcas más chicas las de muestras de $n = 50$, y las marcas más grandes las de muestras de $n = 500$. Además, cada modelo tiene una figura diferente, y se muestra un color distinto para cada tipo de cálculo de error, ya sea por validación cruzada, Bootstrap .632, Bootstrap .623+ o Monte Carlo. Podemos apreciar que para tamaños de muestra más grandes, la estimación mejora, y que las estimaciones por CV y Bootstrap .632 tienden a estar cerca de la diagonal, por lo que son "más confiables".

Por otro lado, la estimaciones por Bootstrap .623+ por lo general sobreestiman el error cuando el riesgo verdadero es alto, particularmente para tamaños de muestra pequeños, pero para valores más grandes de n, se acercan más al verdadero valor.

Conclusiones

En este trabajo analizamos el riesgo de distintos clasificadores en un entorno controlado, en donde conocemos la distribución verdadera de los datos y por ende podemos compararlos con el clasificador óptimo de Bayes. Como pudimos observar, cuando hay una desproporción de los datos o estos se encuentran muy cercanos, en general es difícil encontrar buenos clasificadores, y en ocasiones ni siquiera el óptimo de Bayes hace un trabajo que nos deje razonablemente satisfechos. Sin embargo, dicho estimador es una línea de partida muy útil, y en el caso en el que se conocen las distribuciones verdaderas de los datos nos da una cota sobre el mejor posible desempeño que puedan tener los clasificadores.

Desafortunadamente, en la vida real no siempre se conocen las distribuciones verdaderas de los datos, por lo que es conveniente conocer el comportamiento de los distintos tipos de clasificadores y estudiar cuáles son las mejores técnicas en las que se pueden medir los errores y riesgos de clasificación.

Como pudimos observar, el desempeño de k-NN depende mucho de un buen balance entre k y n, pues no siempre un valor grande de k es la mejor opción. De hecho, valores muy chicos hacen que la clasificación sea muy local y no necesariamente represente el comportamiento total de la muestra, mientras que valores muy grandes no son los indicados si es que tenemos muestras desproporcionadas, puesto que el estimador tendería a clasificar a la mayoría de los puntos con el valor que tiene más representación, sin considerar tanto el entorno local de ellos. Como alternativa se encuentra el Weighted k-NN, que permite añadir más peso a los puntos más cercanos y en general tiene un mejor desempeño, pero de nuevo, es importante un buen balance entre k y n.

Por otro lado, los estimadores LDA, QDA y el Criterio de Fisher dependen mucho de la estructura de los datos, y son más restrictivos puesto que sus fronteras de decisión son hiperplanos, curvas cuadráticas o rectas cuando se hace la proyección sobre algún eje. Sin embargo, si por alguna razón reconocemos la estructura de los datos y vemos que las covarianzas son muy parecidas, el LDA podría ser una buena opción, o si son distintas, el QDA podría servir en el caso en el que los datos sigan una distribución normal aproximada, lo cual se puede comprobar con métodos de inferencia estadística.

Es importante evaluar el riesgo de los distintos clasificadores y conocer su comportamiento general, porque en la práctica no se tiene acceso a la distribución verdadera de los datos, así que no hay una línea de base con la cual comparar y decidir si un clasificador es bueno o no. Como pudimos observar, el estimador de riesgo por Bootstrap $.632+$ en general sobreestima el error (lo cual también se conoce como que es un estimador más pesimista), pero esto puede ser útil en situaciones en donde se desea evitar lo más posible el riesgo. Por otro lado, los estimadores por Validación Cruzada y Bootstrap $.632$ son la mejor opción cuando no es tan crítico evitar el mayor riesgo posible, y como pudimos observar en las gráficas, siempre tienden a parecerse mucho al riesgo verdadero sin importar el tamaño de muestra, aunque ciertamente, a tamaños mayores, los estimadores son más confiables.

En conclusión, antes de escoger modelos de clasificación, es importante conocer la estructura o naturaleza de los datos y analizar qué tan crítico es evitar errores de clasificación. Además, es útil conocer el contexto de los datos en el problema en el que se trabaje, para que de esta manera se pueda tomar una decisión informada y fundamentada sobre los clasificadores a utilizar.

Referencias

[1] Aquino López, M. A. (2025). *Notas de Ciencia de Datos*. GitHub. Recuperado de `https://github.com/maquinolopez/Ciencia_De_Datos.git`

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