Comparación de Técnicas de Modelos de Clasificación Bajo Distintos Supuestos

Rodolfo Jesús Ramirez Lucario
Omar García Ramos
Eric Ernesto Moreles Abonce

1 Introducción

El propósito de este trabajo es el de estudiar, en un entorno controlado donde la distribución verdadera es conocida, el comportamiento de los clasificadores vistos en el curso frente al clasificador óptimo de Bayes. Para esto, hemos programado el algoritmo del clasificador óptimo de Bayes y mediante simulación montecarlo hemos estimado el error teórico de este clasificador y lo hemos comparado con los errores de otros modelos cuyos errores se han estimado bajo la metodología de Validación Cruzada (CV). Los otros modelos a considerar son: Naive Bayes, Quadratic Discriminant, KNN1 (K-Nearest Neighbors with 1 neighbor), KNN3, KNN5, KNN11 y KNN21.
Evaluamos la sensibilidad de los modelos ante situaciones típicas como desbalance de la muestra, varianzas diferentes (o iguales) en los grupos y correlaciones fuertes o débiles.
Se supondrá que la v.a. condicionada tiene una distirbución normal. Formalmente, consideremos el par de v.a. X y Y tales que X una v.a. que, condicionada a Y (una v.a. discreta binaria), tiene las siguientes distribuciones:

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

además, la distribución a priori de Y está dada por $\pi_0 = \mathbb{P}[Y=0]$ y $\pi_1 = \mathbb{P}[Y=1]$.

2 Análisis de Sensibilidad

Por lo visto en clase, el Clasificador de Bayes es el mejor clasificador en el sentido de que es el que menor probabilidad de error tiene i.e.

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

dónde g es el clasificador. En otras palabras, el clasificador de Bayes establece un límite teórico de desempeño. Ningún clasificador puede hacerlo mejor en términos de probabilidad de error. Los demás clasificadores que utilicemos en la simulación deberán, en promedio, presentar un error mayor o, en el mejor de los casos, aproximarse al riesgo de Bayes.

A continuación consideraremos los siguientes casos de hiperparámetros para Σ_0, Σ_1, μ_0 y μ_1:

1) $\Sigma_0 = \Sigma_1 = 2I$ y $\mu_0 = [-1, 1]$, $\mu_1 = [1, -1]$,

2) $\Sigma_0 = I, \Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$ y $\mu_0 = [-2, 2], \mu_1 = [0, 0]$

3) $\Sigma_0 = \Sigma_1 = \begin{bmatrix} 4 & 5{,}7 \\ 9 & 4 \end{bmatrix}$ y $\mu_0 = [-2, 2], \mu_1 = [0, 0]$

4) $\Sigma_0 = \begin{bmatrix} 4 & 5{,}7 \\ 5{,}7 & 9 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & -3{,}6 \\ -3{,}6 & 16 \end{bmatrix}$ y $\mu_0 = [-2, 2], \mu_1 = [0, 0]$

2.1 Misma Varianza e Independencia

Analicemos el caso **1)** dónde estamos suponiendo que las covarianzas de las densidades de las poblaciones son iguales y además las covarianzas son diferentes. Para este caso, los datos satisfacen los supuestos del modelos Discriminante Lineal (lda) por lo que es de esperarse que este tenga un error cercano al error del Clasificador Óptimo de Bayes. Una simulación de este caso se muestra en la figura 1 para clases balanceadas (50-50) y en la 2 para clases des-balanceadas (50-350). Estimamos el error $L(g)$ para cada clasificador g y los resultados se presentan en la Tabla 1 y en la Figura 3 este último gráfico muestra cómo evoluciona el error de los diferentes clasificadores (eje y) cuándo se deja fija la población 0 en 50 individuos y hacemos crecer la población 1 (el eje x).

Cuadro 1: Error de Clasificación $L(g)$ y Desviación Estándar para Diferentes Modelos en el caso 1.

n1	**Naive Bayes**		**LDA**		**QDA**		**k-NN (k=1)**		**k-NN (k=3)**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,130	0,110	0,120	0,117	0,130	0,110	0,240	0,150	0,180	0,133
100	0,120	0,065	0,133	0,060	0,113	0,060	0,153	0,079	0,127	0,047
150	0,135	0,071	0,140	0,073	0,145	0,072	0,240	0,070	0,165	0,067
200	0,100	0,065	0,104	0,067	0,104	0,067	0,160	0,069	0,136	0,090
250	0,090	0,070	0,087	0,072	0,093	0,074	0,157	0,058	0,107	0,053
300	0,091	0,036	0,094	0,031	0,094	0,031	0,151	0,042	0,134	0,051
350	0,070	0,027	0,077	0,024	0,072	0,017	0,117	0,034	0,087	0,026

n1	**k-NN (k=5)**		**k-NN (k=11)**		**k-NN (k=21)**		**Bayes**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,160	0,111	0,150	0,136	0,130	0,119	0,150	0,036
100	0,120	0,040	0,133	0,060	0,113	0,052	0,144	0,029
150	0,135	0,032	0,130	0,046	0,130	0,051	0,125	0,024
200	0,108	0,074	0,112	0,056	0,104	0,041	0,115	0,019
250	0,100	0,052	0,110	0,067	0,103	0,057	0,098	0,015
300	0,103	0,043	0,117	0,035	0,106	0,029	0,089	0,012
350	0,085	0,017	0,075	0,025	0,072	0,017	0,083	0,013

Dado que las densidades condicionales son normales y la covarianza es 0 en las matrices de covarianzas, estamos asumiendo independencia entre las covariables por lo que, como era de esperarse, el modelo Naive Bayes tiene un buen desempeño frente a los demás y está bastante cerca del error del clasificador óptimo de Bayes. Adicionalmente, al asumir varianzas iguales, el LDA también tiene un muy buen desempeño. Otro fenómeno observable en la Figura 3 y que se seguirá repitiendo en las demás gráficas es que, conforme aumenta el tamaño de la población 1, el error disminuye; este efecto lo podemos atribuir al sesgo del error a clasificar correctamente la población más grande.

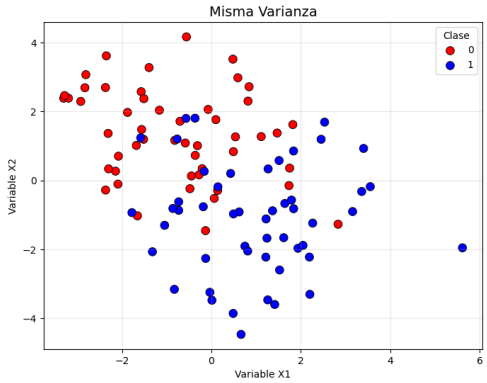

Figura 1: Poblaciones balanceadas con la misma varianza, en un caso con baja separación entre las clases y covariables independientes.

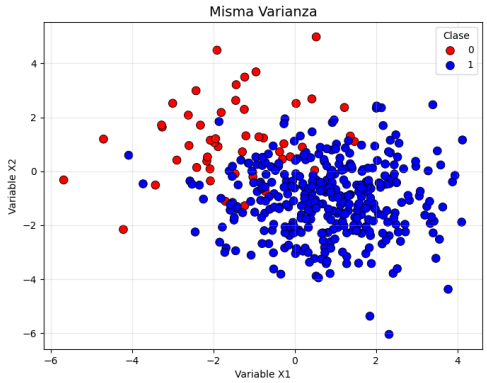

Figura 2: Poblaciones des-balanceadas con la misma varianza, en un caso con baja separación entre las clases y covariables independientes.

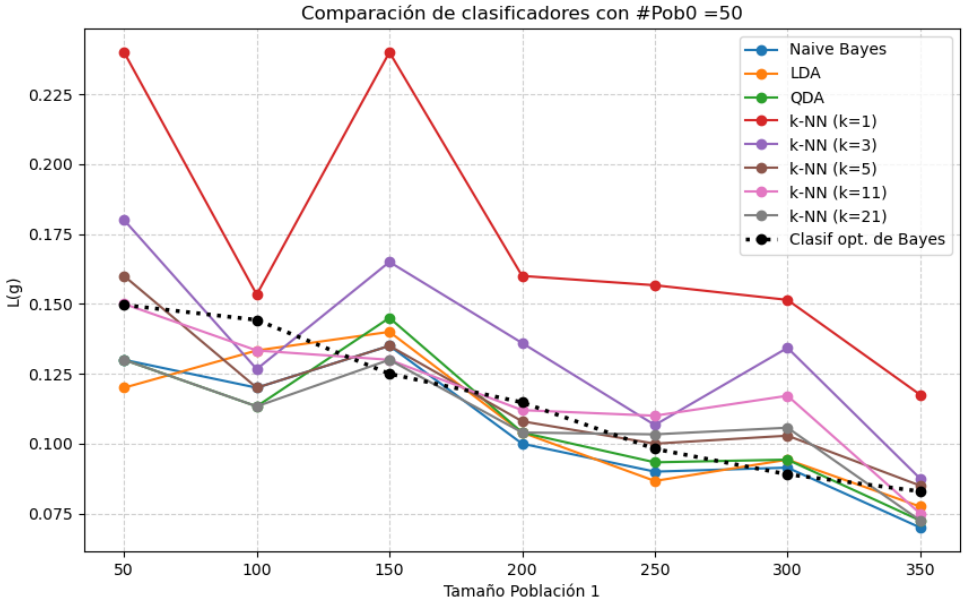

Figura 3: Este gráfico muestra cómo evoluciona el error de los diferentes clasificadores (eje y) cuándo se deja fija la población 0 en 50 individuos y hacemos crecer la población 1 (el eje x). Este es el caso de poblaciones con misma varianza, en un caso con baja separación entre las clases y covariables independientes Caso **1)**.

El gráfico 4 nos indica la diferencia entre la media obtenida de error en los modelos respecto al error del clasificador óptimo de Bayes (el primero menos el segundo), esto mientras variamos el tamaño de la población 1 y dejamos fija la población 0 en 50. Podemos notar que hay muchos colores azules, lo que indica que en diversas ocasiones el error de los demás clasificadores fue mejor que el error real, pero esto lo podemos atribuir al azar pues en otras simulaciones el heatmap es en su mayoría rojo. De los mejores algoritmos en este caso fue el modelo Naive Bayes y un ejemplo de su clasificación para una simulación de este caso lo podemos ver en la Figura 5.

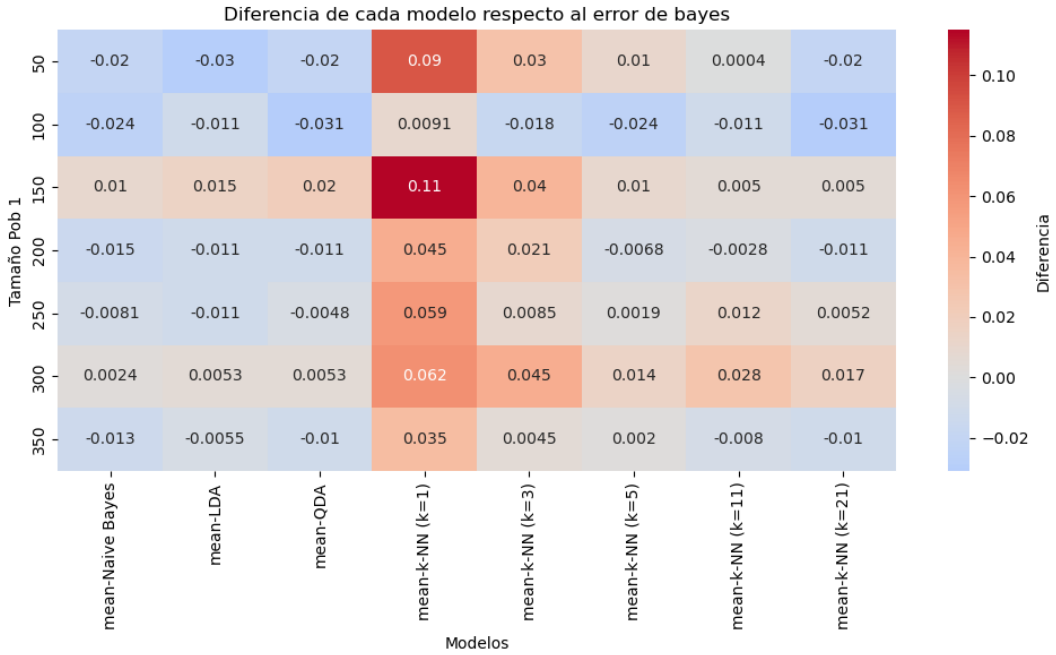

Figura 4: Este grafico nos indica la diferencia entre la media obtenida de error en los modelos respecto al error del clasificador óptimo de Bayes (el primero menos el segundo), esto mientras variamos el tamaño de la población 1 y dejamos fija la población 0 en 50. Este es el caso de poblaciones con la misma varianza, en un caso con baja separación entre las clases y covariables independientes.

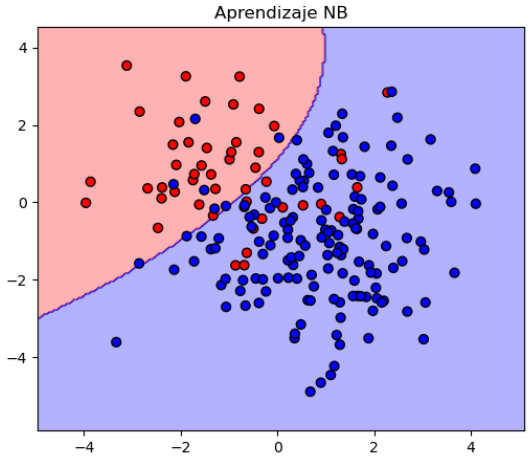

Figura 5: Este gráfico nos muestra región de decisión que genera el algoritmo de NB bajo una simulación de datos del tipo **1)**.

2.2 Diferente Varianza e Independencia

Consideremos ahora el Caso **2)** que es cuándo tenemos diferentes matrices de covarianza e independencia entre las covariables. Una simulación de este caso se muestra en la Figura 6. Podemos ver en la evolución del errror de la Figura 8 cómo el modelo LDA tiene un mal desempeño, esto dado que se ha violado el supuesto

de igualdad de varianzas. Los datos de la estimación junto con las desviaciones estándar las podemos ver en la Tabla 2. Por otro lado, por la misma razón que en el análisis anterior el modelo Naive Bayes tiene un buen desempeño así como el QDA, pues las varianzas son diferentes. La región aprendida por el modelo QDA cuándo la población está equilibrada (50-50) la podemos ver en la Figura 7. El heatmap de las diferencias las vemos en la Figura 9.

Cuadro 2: Error de Clasificación $L(g)$ y Desviación Estándar para Diferentes Modelos en el caso 2.

n1	**Naive Bayes**		**LDA**		**QDA**		**k-NN (k=1)**		**k-NN (k=3)**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,170	0,119	0,180	0,087	0,180	0,117	0,170	0,078	0,170	0,090
100	0,147	0,111	0,187	0,078	0,140	0,101	0,207	0,092	0,187	0,126
150	0,115	0,067	0,155	0,091	0,125	0,081	0,195	0,065	0,175	0,084
200	0,164	0,066	0,184	0,045	0,172	0,076	0,208	0,069	0,168	0,050
250	0,157	0,050	0,180	0,045	0,150	0,056	0,200	0,052	0,183	0,034
300	0,111	0,041	0,129	0,059	0,109	0,040	0,160	0,073	0,151	0,075
350	0,117	0,054	0,135	0,025	0,122	0,047	0,120	0,046	0,132	0,028

n1	**k-NN (k=5)**		**k-NN (k=11)**		**k-NN (k=21)**		**Bayes**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,200	0,118	0,190	0,114	0,200	0,100	0,134	0,035
100	0,160	0,100	0,147	0,115	0,133	0,107	0,145	0,024
150	0,135	0,067	0,155	0,093	0,150	0,084	0,135	0,027
200	0,184	0,054	0,156	0,068	0,176	0,065	0,125	0,019
250	0,223	0,042	0,180	0,048	0,163	0,053	0,123	0,018
300	0,137	0,067	0,143	0,056	0,123	0,041	0,113	0,015
350	0,125	0,039	0,132	0,037	0,135	0,050	0,103	0,014

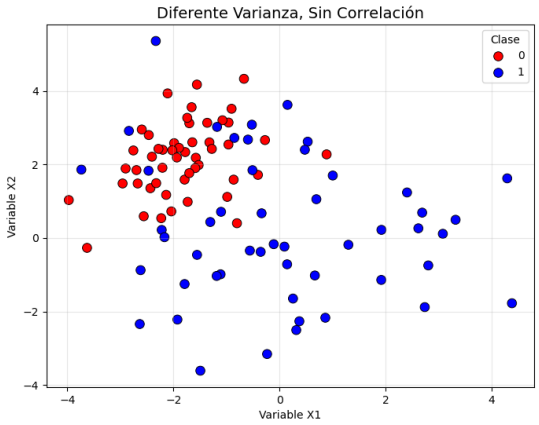

Figura 6: Poblaciones balanceadas con diferente varianza, en un caso con baja separación entre las clases y covariables independientes Caso **2**).

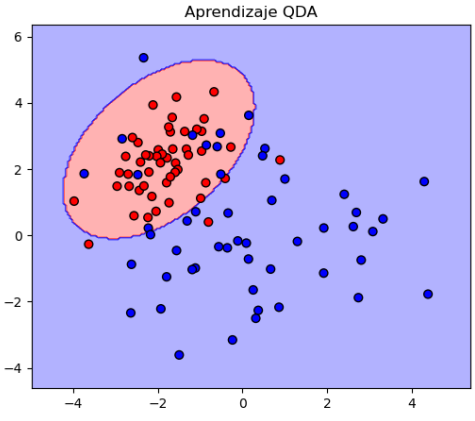

Figura 7: Región de decisión aprendida por el modelo QDA para una población balanceada con diferente varianza, en un caso con baja separación entre las clases y covariables independientes Caso **2**).

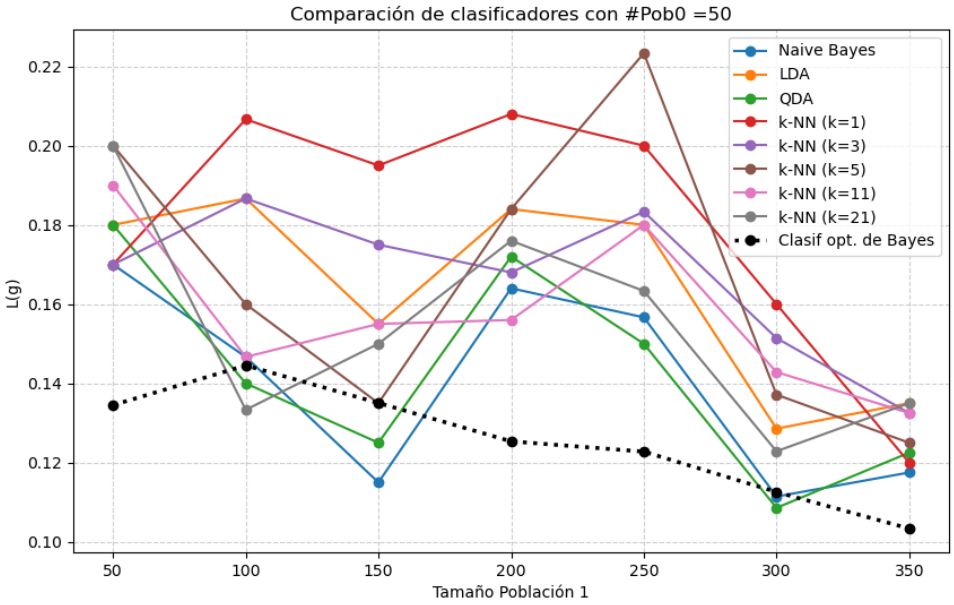

Figura 8: Este gráfico muestra cómo evoluciona el error de los diferentes clasificadores (eje y) cuándo se deja fija la población 0 en 50 individuos y hacemos crecer la población 1 (el eje x). Este es el caso de poblaciones con diferente varianza, en un caso con baja separación entre las clases y covariables independientes Caso **2)**.

Podemos notar en la figura 9 que esta es en su mayoría de color rojo, lo que nos indica que en la mayoría de los casos el clasificador óptimo de Bayes fue el mejor modelo (como era de esperarse).

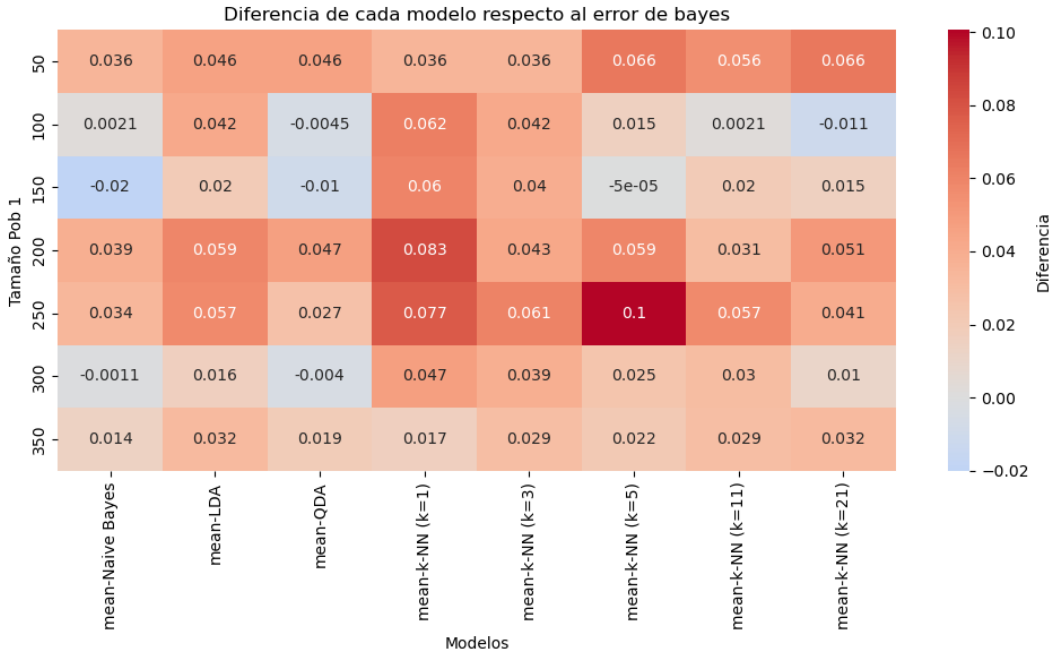

Figura 9: Este grafico nos indica la diferencia entre la media obtenida de error en los modelos respecto al error del clasificador óptimo de Bayes (el primero menos el segundo), esto mientras variamos el tamaño de la población 1 y dejamos fija la población 0 en 50. Este es el caso de poblaciones con diferente varianza, en un caso con baja separación entre las clases y covariables independientes Caso **2)**.

2.3 Misma Varianza y Alta Correlación

Ahora analicemos el caso en el que existe una covarianza fuerte entre las covariables ; para esto, consideramos primero el caso de densidades con la misma matriz de covarianzas i.e. el **Caso 3)**. Una simulación de este caso se muestra en la Figura 10 dónde las clases parecieran formar líneas paralelas. Podemos ver en la figura 11 el gráfico del error para los algoritmos KNN para $k's$ diferentes, con $k \in \{1, 3, 5, 11, 21\}$ y aumentando el desbalanceo de la muestra. Observamos que para valores de k muy grandes, el algoritmo se vuelve menos preciso.

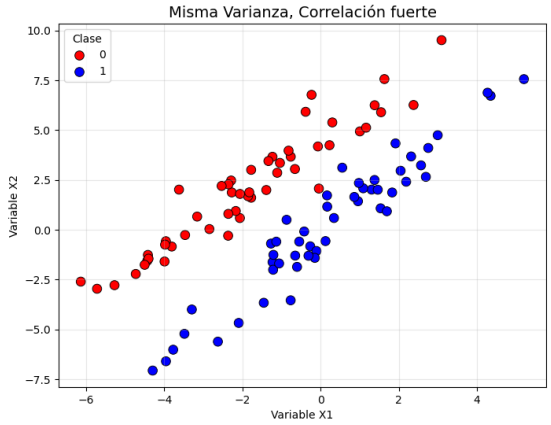

Figura 10: Poblaciones balanceadas con la misma varianza, en un caso con baja separación entre las clases y una alta correlación entre las covariables Caso **3)**.

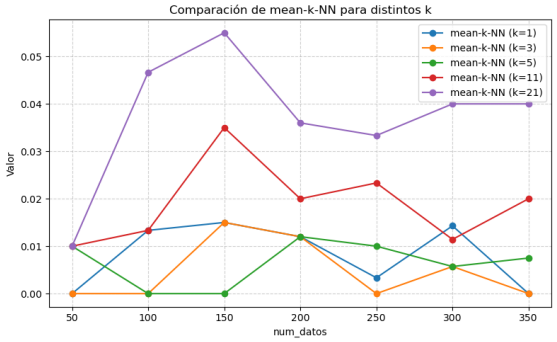

Figura 11: Gráfica de Error de los modelos KNN moviendo el número de vecinos con 1, 3, 5, 11 y 21. Esto para el modelo Caso **3)**.

Algo interesante a notar en la figura de la evolución de $L(g)$ bajo todos los modelos en la Figura 12, es que al ya no haber independencia entre las covariables, el algoritmo de Naive Bayes tiene un mal desempeño mientras que LDA sigue teniendo un buen desempeño por la igualdad entre las matrices de Covarianzas.

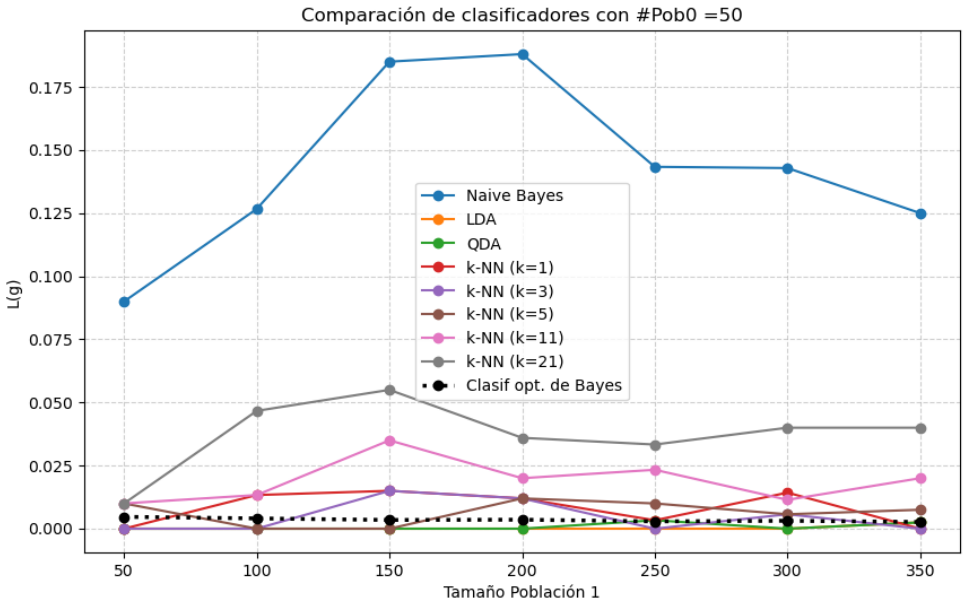

Figura 12: Este gráfico muestra cómo evoluciona el error de los diferentes clasificadores (eje y) cuándo se deja fija la población 0 en 50 individuos y hacemos crecer la población 1 (el eje x). Este es el caso de poblaciones con misma varianza, en un caso con baja separación entre las clases y con una alta correlación entre covariables Caso **3)**.

Las estimaciones de $L(g)$ junto con sus desviaciones estandar para este caso las podemos ver en la Tabla 3. Así como el ejemplo de un ajuste a una muestra de estos datos en la Figura 13.

Cuadro 3: Error de Clasificación $L(g)$ y Desviación Estándar para Diferentes Modelos en el caso 3.

n1	**Naive Bayes**		**LDA**		**QDA**		**k-NN (k=1)**		**k-NN (k=3)**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,090	0,083	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
100	0,127	0,092	0,000	0,000	0,000	0,000	0,013	0,027	0,000	0,000
150	0,185	0,063	0,000	0,000	0,000	0,000	0,015	0,023	0,015	0,023
200	0,188	0,026	0,000	0,000	0,000	0,000	0,012	0,018	0,012	0,018
250	0,143	0,026	0,000	0,000	0,003	0,010	0,003	0,010	0,000	0,000
300	0,143	0,000	0,000	0,000	0,000	0,000	0,014	0,019	0,006	0,017
350	0,125	0,000	0,003	0,008	0,003	0,008	0,000	0,000	0,000	0,000

n1	**k-NN (k=5)**		**k-NN (k=11)**		**k-NN (k=21)**		**Bayes**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,010	0,030	0,010	0,030	0,010	0,030	0,005	0,006
100	0,000	0,000	0,013	0,027	0,047	0,052	0,004	0,006
150	0,000	0,000	0,035	0,050	0,055	0,052	0,003	0,003
200	0,012	0,018	0,020	0,020	0,036	0,033	0,004	0,004
250	0,010	0,015	0,023	0,015	0,033	0,021	0,003	0,003
300	0,006	0,011	0,011	0,014	0,040	0,029	0,003	0,003
350	0,008	0,011	0,020	0,022	0,040	0,023	0,003	0,003

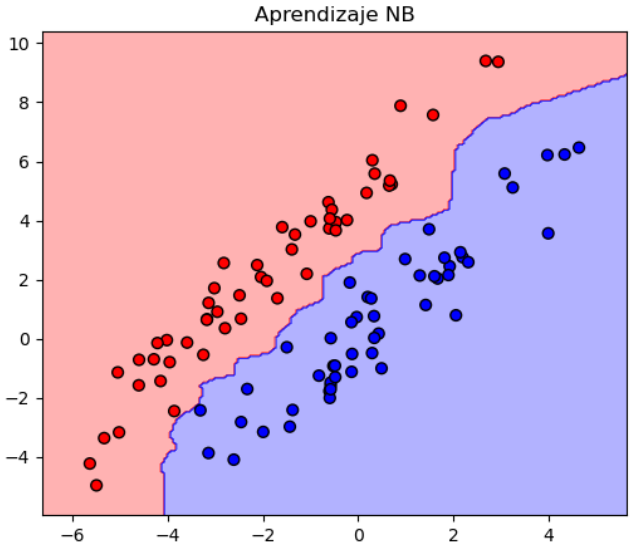

Figura 13: Región de decisión aprendida por el modelo NB para una población balanceada con la misma varianza, en un caso con baja separación entre las clases y covariables con alta dependencia Caso **3**).

EL heatmap para las diferencias de las medias respecto al error real de bayes se puede apreciar en el Heatmap de la Figura 14.

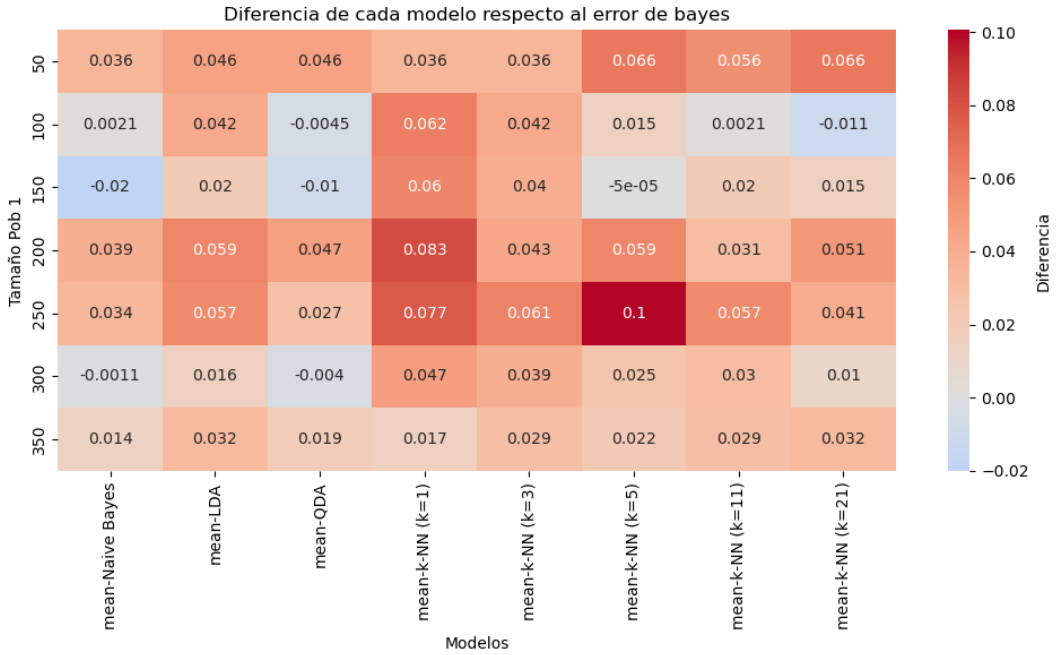

Figura 14: Este grafico nos indica la diferencia entre la media obtenida de error en los modelos respecto al error del clasificador óptimo de Bayes (el primero menos el segundo), esto mientras variamos el tamaño de la población 1 y dejamos fija la población 0 en 50. Este es el caso de poblaciones con diferente varianza, en un caso con baja separación entre las clases y covariables independientes Caso **3)**.

2.4 Diferente Varianza y Alta Correlación

Finalizamos analizando el error de los modelos en el caso en el que se tiene alta dependencia entre las covariables y las matrices de covarianza son diferentes i.e. Caso **4)**. Una simulación de este caso se muestra en la Figura 15 dónde las clases forman una X.

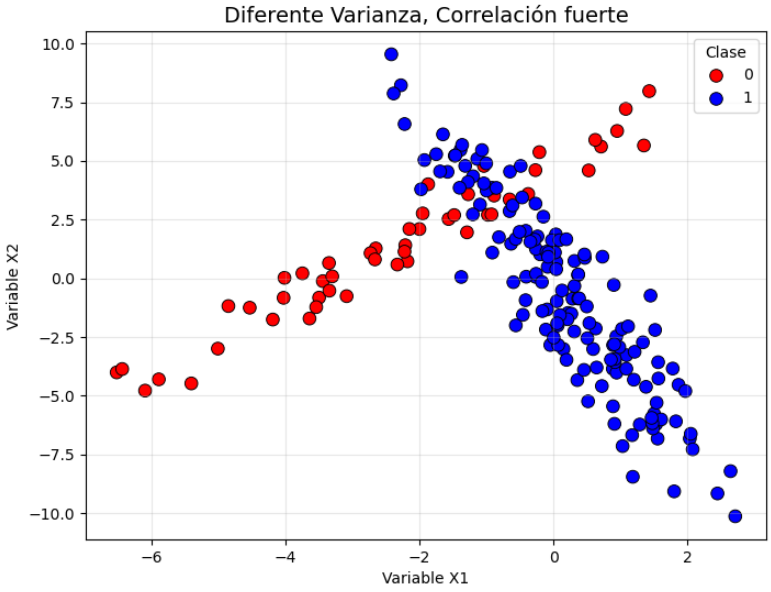

Figura 15: Poblaciones desbalanceadas (50-150) con diferente varianza, en un caso con baja separación entre las clases y una alta correlación entre las covariables Caso **4)**.

En la Figura 16 vemos la evolución del error del algoritmo KNN con $k \in [1, 3, 5, 11, 21]$ y observamos que en general el algoritmo que tiene un mejor desempeño sin importar que aumentamos la muestra es el que tiene $k = 5$.

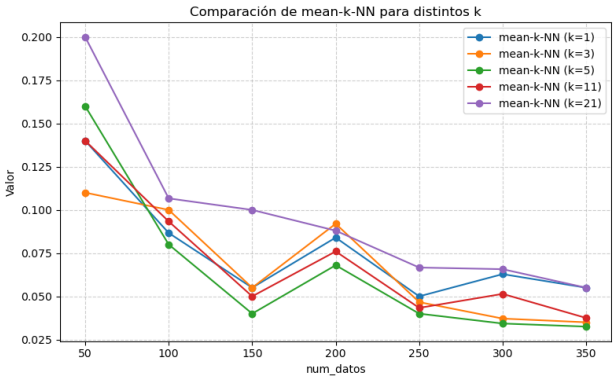

Figura 16: Gráfica de Error de los modelos KNN moviendo el número de vecinos con 1, 3, 5, 11 y 21. Esto para el modelo Caso **4)**.

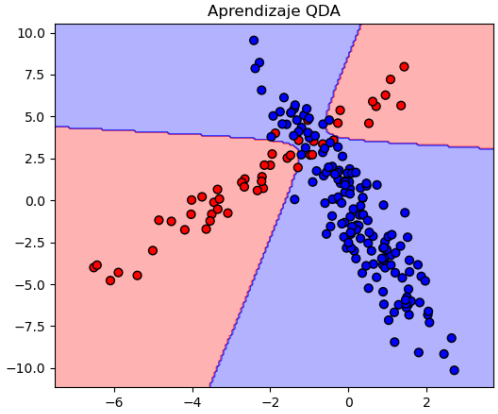

Figura 17: Región de decisión aprendida por el modelo QDA para una población balanceada con diferente varianza, en un caso con baja separación entre las clases y con una alta correlación entre covariables Caso **4)**.

En la Figura 18 vemos que el clasificador LDA ha dejado de dar buenos resultados a consecuencia de las diferencias en la matriz de covarianzas, mientras que QDA sigue dando buenos resultados (error cercano al clasificador óptimo de Bayes).

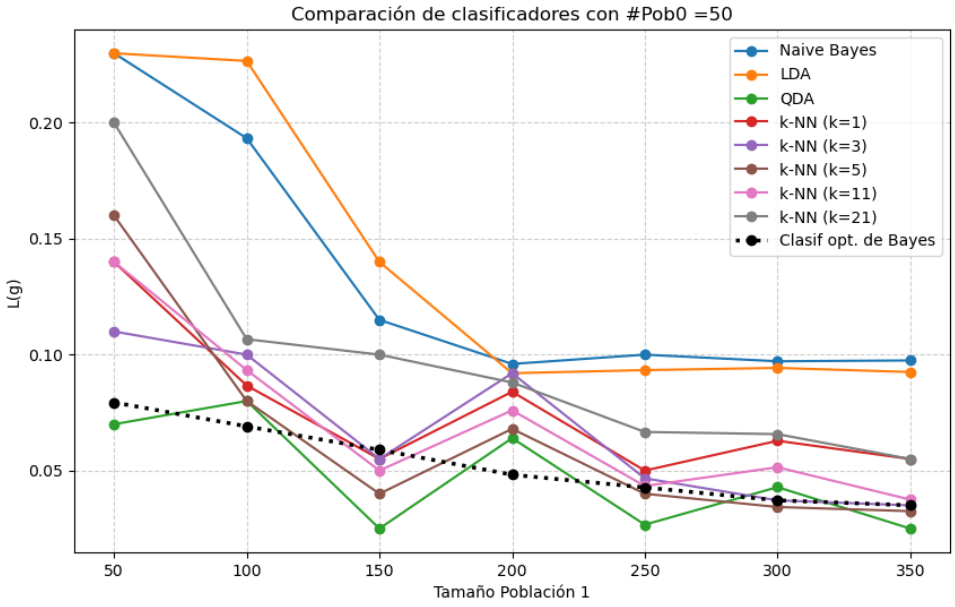

Figura 18: Este gráfico muestra cómo evoluciona el error de los diferentes clasificadores (eje y) cuándo se deja fija la población 0 en 50 individuos y hacemos crecer la población 1 (el eje x). Este es el caso de poblaciones con diferente varianza, en un caso con baja separación entre las clases y con una alta correlación entre covariables Caso **4)**.

Finalmente, podemos ver la región de predicción generada por el clasificador QDA para este último caso en la Figura 17. La Tabla 4 contiene las desviaciones estandar y medias de cada modelo. EL heatmap para las diferencias de las medias respecto al error real de bayes se puede apreciar en el Heatmap de la Figura 19.

Cuadro 4: Error de Clasificación $L(g)$ y Desviación Estándar para Diferentes Modelos en el Caso **4)**.

n1	**Naive Bayes**		**LDA**		**QDA**		**k-NN (k=1)**		**k-NN (k=3)**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,230	0,127	0,230	0,127	0,070	0,078	0,140	0,111	0,110	0,114
100	0,193	0,087	0,227	0,090	0,080	0,088	0,087	0,090	0,100	0,091
150	0,115	0,074	0,140	0,070	0,025	0,025	0,055	0,027	0,055	0,027
200	0,096	0,041	0,092	0,062	0,064	0,041	0,084	0,058	0,092	0,047
250	0,100	0,045	0,093	0,025	0,027	0,025	0,050	0,031	0,047	0,031
300	0,097	0,032	0,094	0,026	0,043	0,047	0,063	0,046	0,037	0,031
350	0,097	0,026	0,092	0,025	0,025	0,030	0,055	0,037	0,035	0,030

n1	**k-NN (k=5)**		**k-NN (k=11)**		**k-NN (k=21)**		**Bayes**	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
50	0,160	0,150	0,140	0,102	0,200	0,118	0,079	0,026
100	0,080	0,083	0,093	0,080	0,107	0,074	0,069	0,018
150	0,040	0,030	0,050	0,039	0,100	0,059	0,059	0,015
200	0,068	0,036	0,076	0,042	0,088	0,056	0,048	0,012
250	0,040	0,029	0,043	0,030	0,067	0,026	0,043	0,009
300	0,034	0,031	0,051	0,033	0,066	0,018	0,037	0,009
350	0,032	0,027	0,037	0,026	0,055	0,027	0,035	0,006

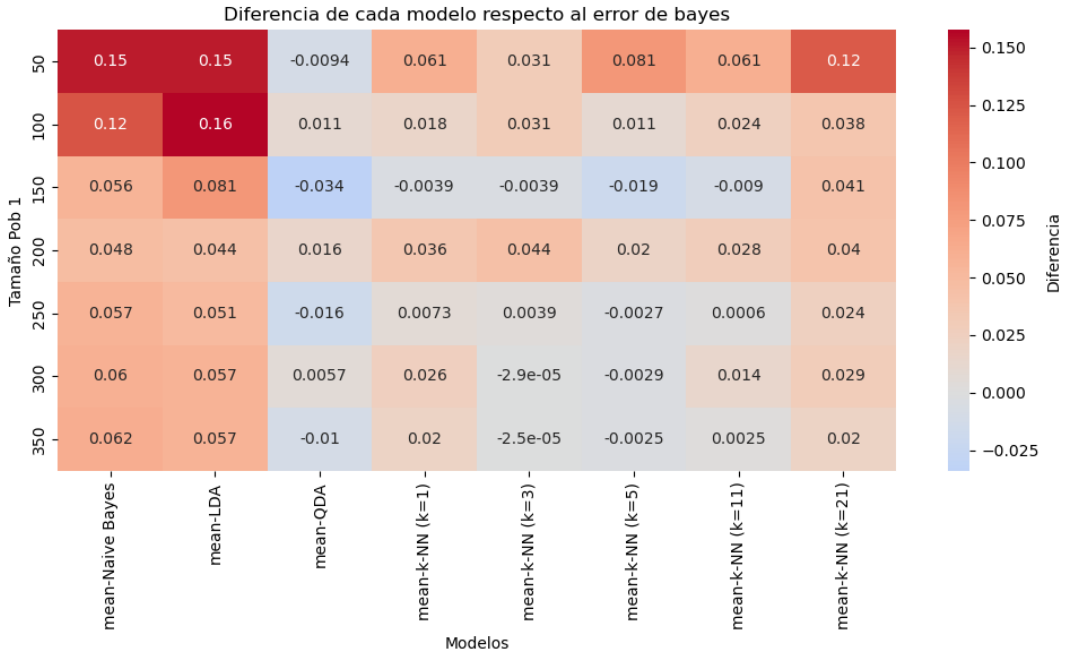

Figura 19: Este grafico nos indica la diferencia entre la media obtenida de error en los modelos respecto al error del clasificador óptimo de Bayes (el primero menos el segundo), esto mientras variamos el tamaño de la población 1 y dejamos fija la población 0 en 50. Este es el caso de poblaciones con diferente varianza, en un caso con baja separación entre las clases y covariables altamente dependientes Caso **4)**.

3 Conclusión

Recapitulando las conclusiones a las que llegamos en este proyecto, podemos decir que la estimación del error del clasificador óptimo de Bayes es un buen punto de referencia para saber qué tan bueno puede ser un modelo; lamentablemente, en la práctica no contamos con las densidades reales de los datos por lo que debemos recurrir a otros modelos como los presentados en este proyecto: NB, KNN, LDA Y QDA. Hemos visto que en caso de tener poblaciones en las que los datos presentan un patrón de dispersión muy diferente dependiendo de la clase, el modelo de LDA tiene resultados deficientes. Por otro lado, el modelo QDA tiene un muy buen desempeño.
Cómo lo esperábamos, la hipótesis de independencia entre las covariables es un supuesto muy fuerte, así cómo se mostró en las gráficas de los casos **3)** y **4)**, de usar el algoritmo NB para modelar y no comparar con otros, estaríamos obteniendo un pobre ajuste comparado con el que pudiéramos haber obtenido con otros modelos. Cuándo se usa el modelo de KNN es necesario comparar los errores bajo diferentes valores del hiperparámetro k esto pues dependiendo de la estructura de las clases, este valor puede ser muy diferente.