Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 4

Victor Manuel Gómez Espinosa

4 de mayo de 2020

1. Problema 1

Para datos de clasificación binaria $\{(x_i, y_i)\}_{i=1}^{n}$, tenemos la siguiente función de costo

$$\sum_{i=1}^{n}(\theta(y_i)-(\beta' x_i-\beta_o))^2 \tag{1.1}$$

Y definimos n_+, n_- el número de observaciones con $y_i=1$ y $y_i=-1$, respectivamente c_+, c_- el centroide de las observaciones con $y_i=1$ y $y_i=-1$ y c el centroide de todos los datos.

Las matrices:

$$S_B=(c_+-c_-)(c_+-c_-)'=vv'$$

$$S_W=\sum_{i:y=1}^{n_+}(x_i-c_+)(x_i-c_+)'+\sum_{i:y=-1}^{n_-}(x_i-c_-)(x_i-c_-)'$$

También considerando las siguientes relaciones:

$$n=n_++n_-$$

$$\sum_{i=1}^{n}x_i=\sum_{i=1}^{n_+}x_i+\sum_{i=n_++1}^{n_-}x_i=n_+c_++n_-c_-$$

$$c = \frac{n_+ c_+ + n_- c_-}{n}$$

$$c_+ = {}^{1}\!/_{n_+} \sum_{i:y=1}^{n_+} x_i, c_- = {}^{1}\!/_{n_-} \sum_{i:y=-1}^{n_-} x_i$$

$$\theta(y_i = 1) = \frac{n}{n_+}, \theta(y_i = -1) = \frac{n}{n_-}$$

a) Verificando que:

$$S_W = \sum_{i:y=1}^{n_+} x_i x'_i + \sum_{i:y=-1}^{n_-} x_i x_i' - n_+ c_+ c_+' - n_- c_- c_-'$$

De:

$$S_W = \sum_{i:y=1}^{n_+} (x_i - c_+)(x_i - c_+)' + \sum_{i:y=-1}^{n_-} (x_i - c_-)(x_i - c_-)' = \sum_{i:y=1}^{n_+} (x_i x_i' - 2c_+ x_i' + c_+ c_+') + \sum_{i:y=-1}^{n_-} (x_i x_i' - 2c_- x_i' + c_- c_-')$$

$$= \sum_{i:y=1}^{n_+} x_i x_i' - 2c_+ \left(n_+ c_+'\right) + n_+ c_+ c_+' + \sum_{i:y=-1}^{n_-} x_i x_i' - 2c_- \left(n_- c_-'\right) + n_- c_- c_-'$$

$$S_W = \sum_{i:y=1}^{n_+} x_i x_i' + \sum_{i:y=-1}^{n_-} x_i x_i' - n_+ c_+ c_+' - n_- c_- c_-' \qquad (1.2)$$

O también: $\sum_{i=1}^{n} x_i x_i' = \sum_{i:y=1}^{n_+} x_i x_i' + \sum_{i:y=-1}^{n_-} x_i x_i' = S_W + n_+ c_+ c_+' + n_- c_- c_-'$

b) Ahora verificamos que el vector $S_B \beta$, es un múltiplo de $(c_+ - c_-) = v$:

$$S_B \beta = \lambda v$$

$$v v' \beta = \lambda v \rightarrow v' \beta = \lambda v^{-1} v$$

$$\lambda = v' \beta \qquad (1.3)$$

c) Ahora si deseamos encontrar el mínimo de la función de costo (1.1), es un problema de mínimos cuadrados:

$$X'X\beta_v = X'Y$$

Donde $X' = \begin{bmatrix} 1 & . & . & . & 1 \\ x_1 & \dots & x_{n_+} & \dots & x_{n_-} \end{bmatrix}$, $Y' = \begin{bmatrix} n/n_+ & \dots & n/n_+ & -n/n_- & \dots & -n/n_- \end{bmatrix}$ y el vector $\beta_v' = \begin{bmatrix} \beta_0 & \beta \end{bmatrix}$, es la solución.

Realizando el lado izquierdo:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i' \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i x_i' \end{bmatrix} = \begin{bmatrix} n & n_+c_+' + n_-c_-' \\ n_+c_+ + n_-c_- & S_W + n_+c_+c_+' + n_-c_-c_-' \end{bmatrix}$$

El lado izquierdo:

$$X'Y = \begin{bmatrix} n_+\left(n/n_+\right) - n_-\left(n/n_-\right) \\ \left(\sum_{i=1}^{n_+} x_i\right)\left(n/n_+\right) + \left(\sum_{i=n_++1}^{n} x_i\right)\left(-n/n_-\right) \end{bmatrix} = \begin{bmatrix} 0 \\ n\left(c_+ - c_-\right) \end{bmatrix}$$

Entonces tenemos:

$$\begin{bmatrix} n & n_+c_+' + n_-c_-' \\ n_+c_+ + n_-c_- & S_W + n_+c_+c_+' + n_-c_-c_-' \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ n\left(c_+ - c_-\right) \end{bmatrix}$$

El primer término:

$$n\beta_0 + \left(n_+c_+' + n_-c_-'\right)\beta = 0 \rightarrow \beta_0 = -\frac{\left(n_+c_+' + n_-c_-'\right)}{n}\beta = -c'\beta$$

$$\beta_0 = -\beta' c \tag{1.4}$$

El segundo:

$$\left(n_+c_+ + n_-c_-\right)\left(-\frac{\left(n_+c_+' + n_-c_-'\right)}{n}\beta\right) + \left(S_W + n_+c_+c_+' + n_-c_-c_-'\right)\beta = n\left(c_+ - c_-\right)$$

Desarrollando el lado izquierdo:

$$\left(S_W + n_+ c_+ c_+' + n\_c\_c\_' - \frac{\left(n_+^2 c_+ c_+' + 2n_+ n\_c\_c_+' + n_-^2 c\_c\_' \right)}{n} \right) \beta$$

Desarrollando lo del paréntesis:

$$= S_W - \frac{\left(n_+^2 c_+ c_+' \right)}{n} + n_+ c_+ c_+' - \frac{\left(2n_+ n\_c\_c_+' \right)}{n} - \frac{\left(n_-^2 c\_c\_' \right)}{n} + n\_c\_c\_'$$

$$= S_W + \left(-\frac{\left(n_+^2 \right)}{n} + n_+ \right) c_+ c_+' - \left(\frac{\left(2n_+ n_- \right)}{n} \right) c\_c_+' + \left(-\frac{\left(n_-^2 \right)}{n} + n_- \right) c\_c\_'$$

$$= S_W + \frac{n_+}{n} \left(-n_+ + n \right) c_+ c_+' - \left(\frac{\left(2n_+ n_- \right)}{n} \right) c\_c_+' + \frac{n_-}{n} \left(-n_- + n \right) c\_c\_'$$

$$= S_W + \frac{n_+ n_-}{n} c_+ c_+' - \left(\frac{\left(2n_+ n_- \right)}{n} \right) c\_c_+' + \frac{n\_n_+}{n} c\_c\_' = S_W + \left(\frac{n_+ n_-}{n} \right) \left(c_+ c_+' - 2c\_c_+' + c\_c\_' \right)$$

$$= S_W + \left(\frac{n_+ n_-}{n} \right) \left(c_+ - c_- \right) \left(c_+ - c_- \right)' = S_W + \left(\frac{n_+ n_-}{n} \right) S_B$$

Finalmente:

$$\left(S_W + \left(\frac{n_+ n_-}{n} \right) S_B \right) \beta = n \left(c_+ - c_- \right) \tag{1.5}$$

d) Y tomando el resultado obtenido previamente $S_B \beta = \lambda v$, donde $\lambda = v' \beta$, tenemos

$$S_W \beta + \left(\frac{n_+ n_-}{n} \right) S_B \beta = nv \rightarrow S_W \beta + \left(\frac{n_+ n_-}{n} \right) \lambda v = nv$$

$$= S_W \beta = nv - \left(\frac{n_+ n_-}{n} \right) \lambda v = \left(n - \frac{n_+ n\_\lambda}{n} \right) v$$

$$\beta = cte S_W^{-1} v \tag{1.6}$$

Entonces β también se encuentra en la dirección de $\left(c_+ - c_- \right) = v$, que coincide con la de Fisher (FDA).

e) Lo anterior permite implementar FDA mediante el algoritmo de mínimos cuadrados. Por ejemplo si elegimos un conjunto de datos sintéticos en 2D como los de la Figura 1.1 a), donde X , es la matriz de datos (o de observaciones y variables), en este caso las coordenadas en el plano (x1,x2) para todas las observaciones, es decir una matriz de tamaño 100x2, con la misma cantidad de observaciones una clase que de la otra (50,50), y el vector Y de tamaño 100x1, cada entrada representa según la clase a la que corresponda:

$$\theta(y_i = 1) = \frac{n}{n_+} = \frac{100}{50} = 2, \theta(y_i = 0) = -\frac{n}{n_-} = -2$$

Entonces resolvemos el problema de mínimos cuadrados $X'X\beta_v = X'Y \rightarrow \beta_v = (X'X)^{-1} X'Y$ para encontrar el plano que separa ambos conjuntos $f(X) = \beta_0 + \beta' X = 0$, que en este caso es la línea negra que observa en la Figura 1.1 b). Note que, para poder graficarla, necesitamos despejar x_2, $\beta_0 + \beta' X = 0 \rightarrow x_2 = (\beta_0 + \beta_1 x_1) / -\beta_2$.

Y para realizar la clasificación necesitamos conocer las distancias de los datos al plano, esto mediante $d = \frac{\beta_0 + \beta' X}{\|\beta\|}$ y dependiendo del signo de esta distancia se le asigna la clase correspondiente, es decir si $d \geq 0, clase = 0$, o $d < 0, clase = 1$.

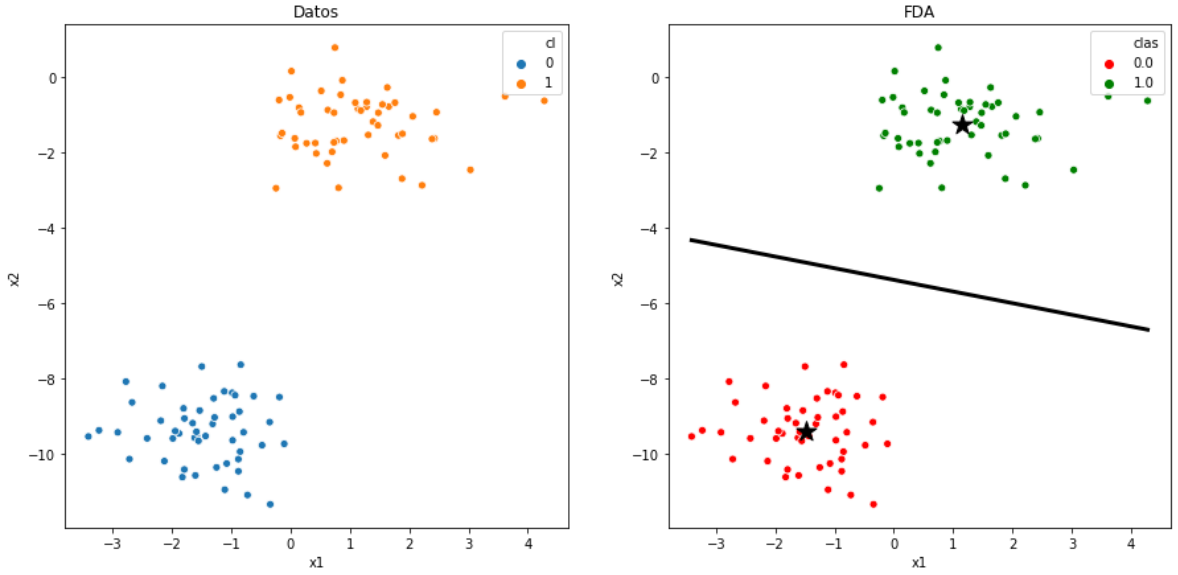

Figura 1.1: a) Datos sintéticos, b) FDA aplicado a los datos sintéticos para clasificación, la línea negra representa el plano separador $f(X) = 0$.

f) Pero ahora si utilizamos un conjunto de datos sintéticos con datos atípicos (Figura 1.2 a), podremos observar que el plano separador es influenciado por los datos atípicos, los centroides se encuentran ligeramente desplazados en la dirección de los datos atípicos (Figura 1.2 b).

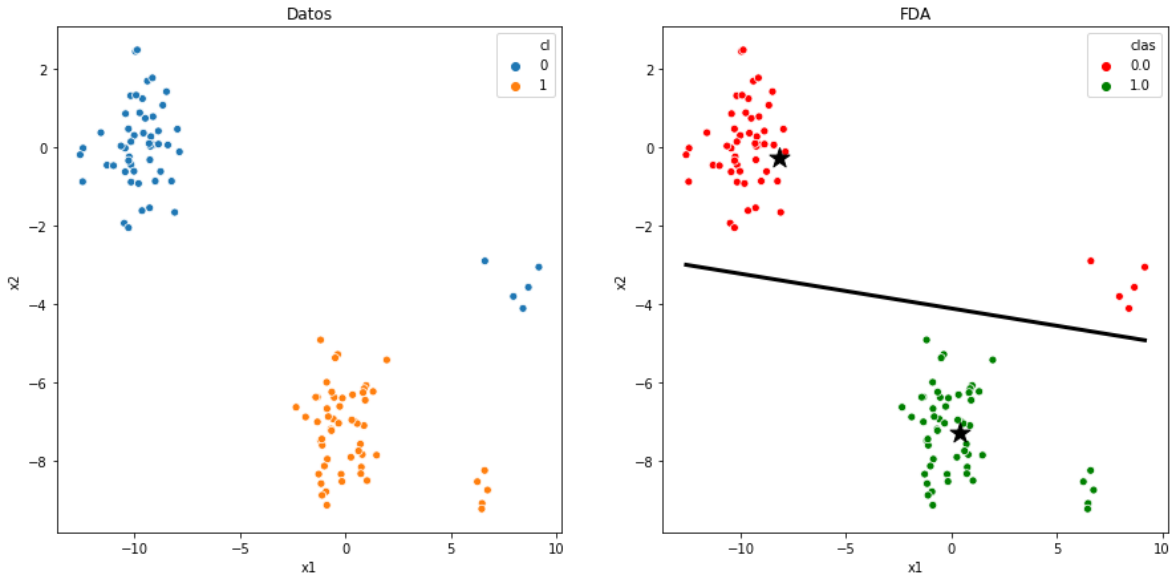

Figura 1.2: a) Datos sintéticos con datos 6típicos, b) FDA aplicado a los datos sintéticos para clasificación, la línea negra representa el plano separador $f(X)=0$ obtenido mediante mínimos cuadrados.

Si ahora utilizamos mínimos cuadrados ponderados para hacerlo más robusto, es decir resolver $X'(WX)\beta_v = X'(WY) \rightarrow \beta_v = (X'WX)^{-1} X'WY$, donde W es una matriz diagonal con pesos. Una forma de asignar estos pesos podría ser $w_{i,i}=0$ para los datos atípicos y $w_{i,i}=1$ para el resto, es decir, sería como removerlos y una forma de determinarlos, puede ser con el grafico chi-cuadrado.

En este caso, ya sabemos de antemano cuales son los datos atípicos, a estos se les asigno el peso $w_{i,i}=0$, se formó la matriz W y se resolvió el problema de mínimos cuadrados antes mencionado. Observe la Figura 1.3 b) y observe que ahora la dirección del plano y la posición de los centroides es diferente, ya que ya no se ve afectado por los datos atípicos.

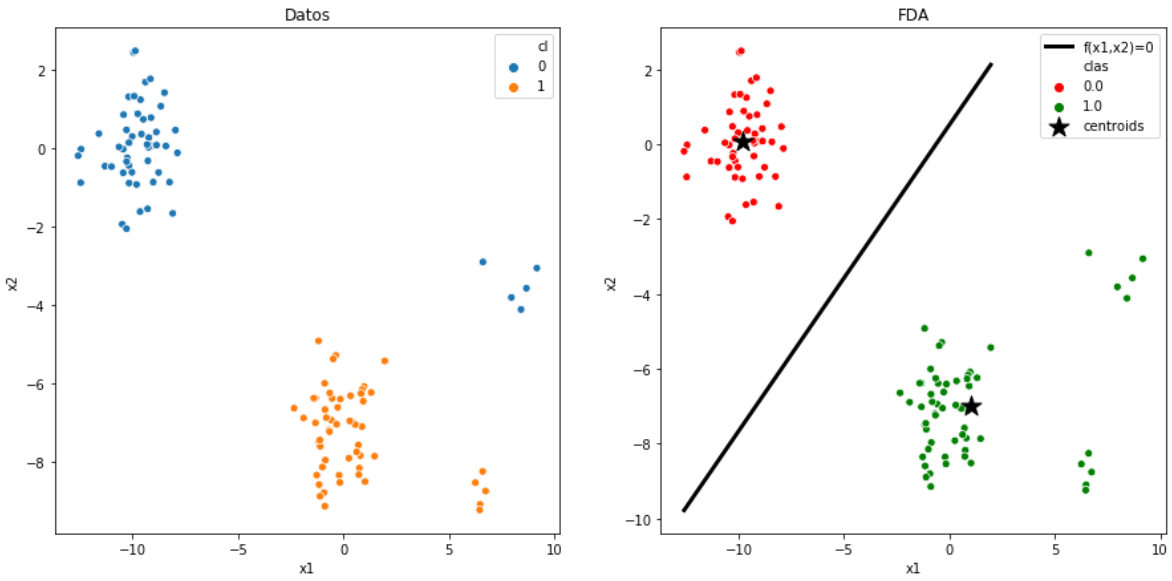

Figura 1.2: a) Datos sintéticos con datos atípicos, b) FDA aplicado a los datos sintéticos para clasificación, la línea negra representa el plano separador $f(X)=0$ obtenido mediante mínimos cuadrados ponderados.

2. Problema 2

Para este problema se utilizó el conjunto de datos MNIST que es una base de dígitos del 0 al 9 a mano, que son imágenes de 28x28 pixeles, en total se cuenta con 70,000 imágenes.

Se tomó un conjunto de entrenamiento (X_train, y_train) del 75% y otro de prueba del 25% restante (X_test, y_test), ambos balanceados (Figura 2.1). El conjunto de entrenamiento se utilizó para ajustar un modelo de regresión multivariada (baseline) $Y = X\hat{B}$ y este utilizarlo para predecir la respuesta o clasificación de las imágenes, es decir a la imagen asignarle una etiqueta del digito correspondiente, tanto con el conjunto de entrenamiento como con el de prueba.

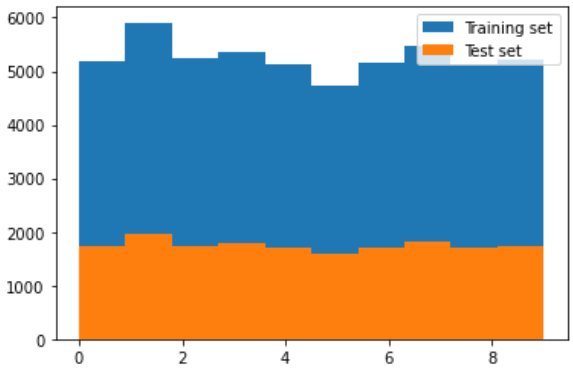

Figura 2.1: Conjuntos de datos, de entrenamiento y prueba.

De igual forma, con el mismo conjunto de datos, se aplicó por separado los modelos LDA y QDA y para evaluar los modelos se utilizaron dos métricas, la raíz de del promedio de los errores al cuadrado (RMSE) (1.7) y R^2 (1.8):

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{1.7}$$

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.8}$$

Observe la tabla 2.1 y note que el RMSE más pequeño se obtuvo mediante LDA y el más grande con QDA, esto mismo se observa en la Figura 2.2 y 2.3 donde LDA tiene nuevamente el R^2 más grande (y mayor precisión) y QDA el más pequeño, lo que nos podría señalar que el mejor modelo para este caso es justamente LDA y que entonces el supuesto de normalidad con varianzas iguales podría ser correcto y el error en estimación se podría deber a la alta dimensionalidad (28x28=784).

	Train set	*Test set*
Baseline	1.57	1.62
LDA	1.47	1.51
QDA	2.52	2.67

Tabla 2.1: RMSE para cada modelo (renglones) para cada conjunto de datos (columnas).

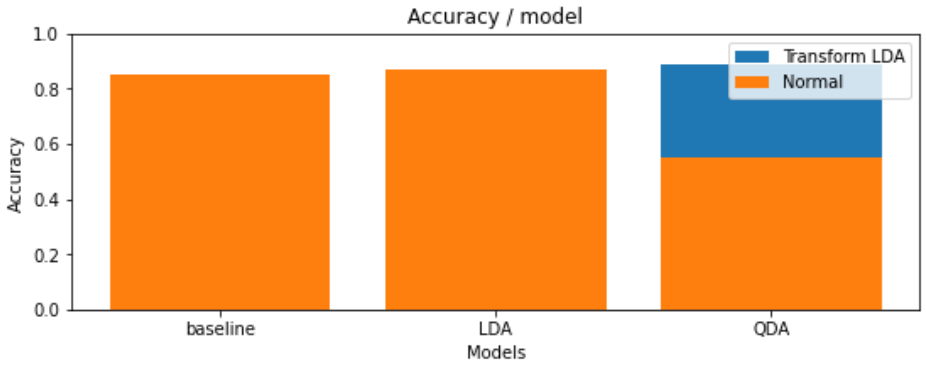

Figura 2.2: Precisión para cada modelo.

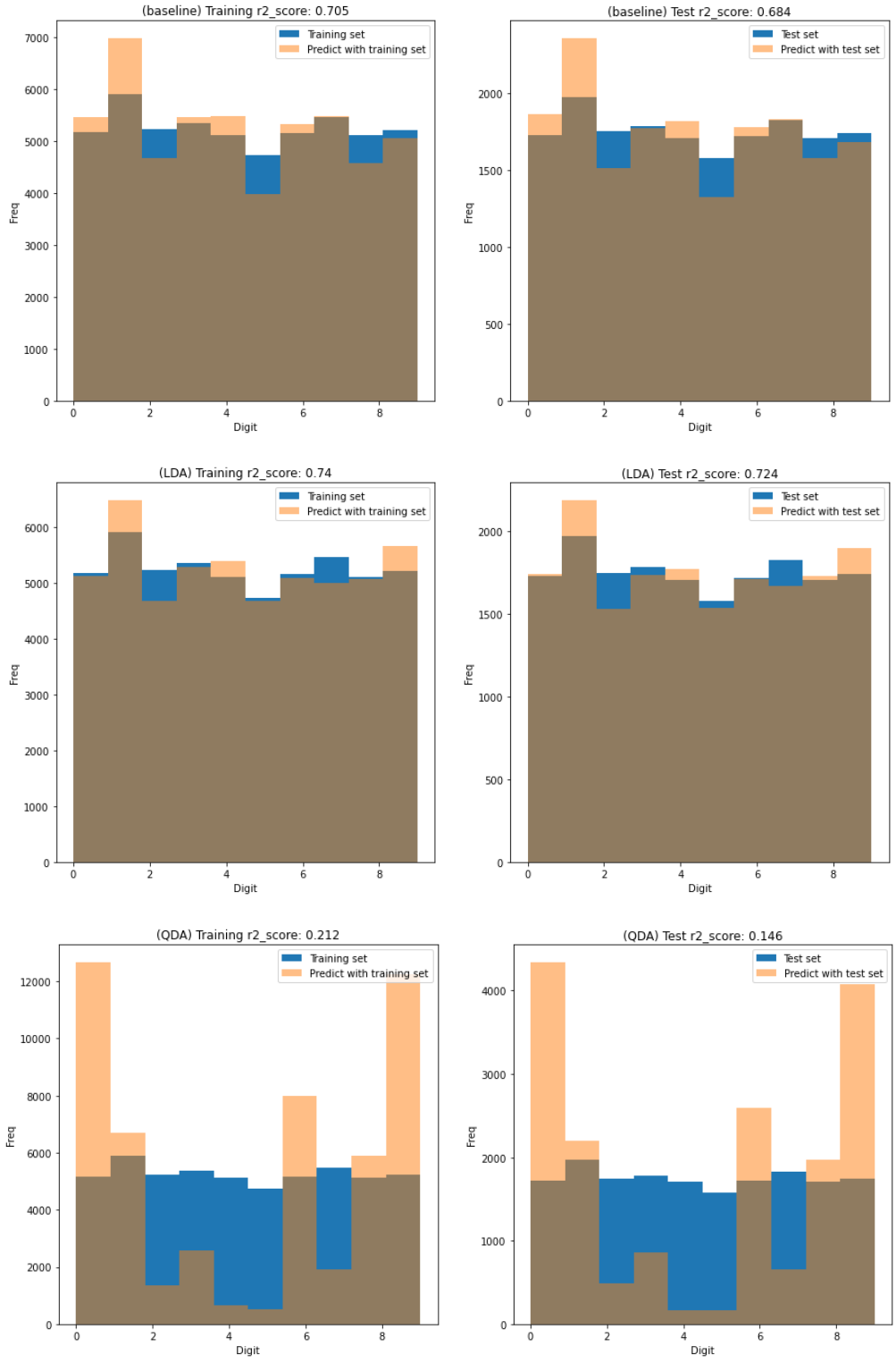

Figura 2.3: Histogramas de las clases reales y de las respuestas con su respectivo R^2, para cada modelo (renglones), para cada conjunto de datos (columnas).

Debido a las razones anteriores, se pensó que quizá una representación diferente de los dígitos en menor dimensión que permita, separar las clases y por lo tanto podría mejorar el modelo y la predicción. Para esto, se estandarizaron los conjuntos de datos, se aplicó una transformación a estos y se ajustó nuevamente los mismos 3 modelos.

Las transformaciones que se probaron fueron PCA con 200 componentes (basado en el criterio de varianza acumulada de al menos el 80%), Kernel PCA con conjuntos más pequeños (6000 y 1500 para el conjunto de entrenamiento y de prueba respectivamente y balanceados), con diferentes kernels encontrando que el de tipo coseno era el que mejor separaba las clases, y por último una transformación LDA con 9 componentes (identifica atributos que maximizan la varianza entre clases y minimiza la varianza entre clases).

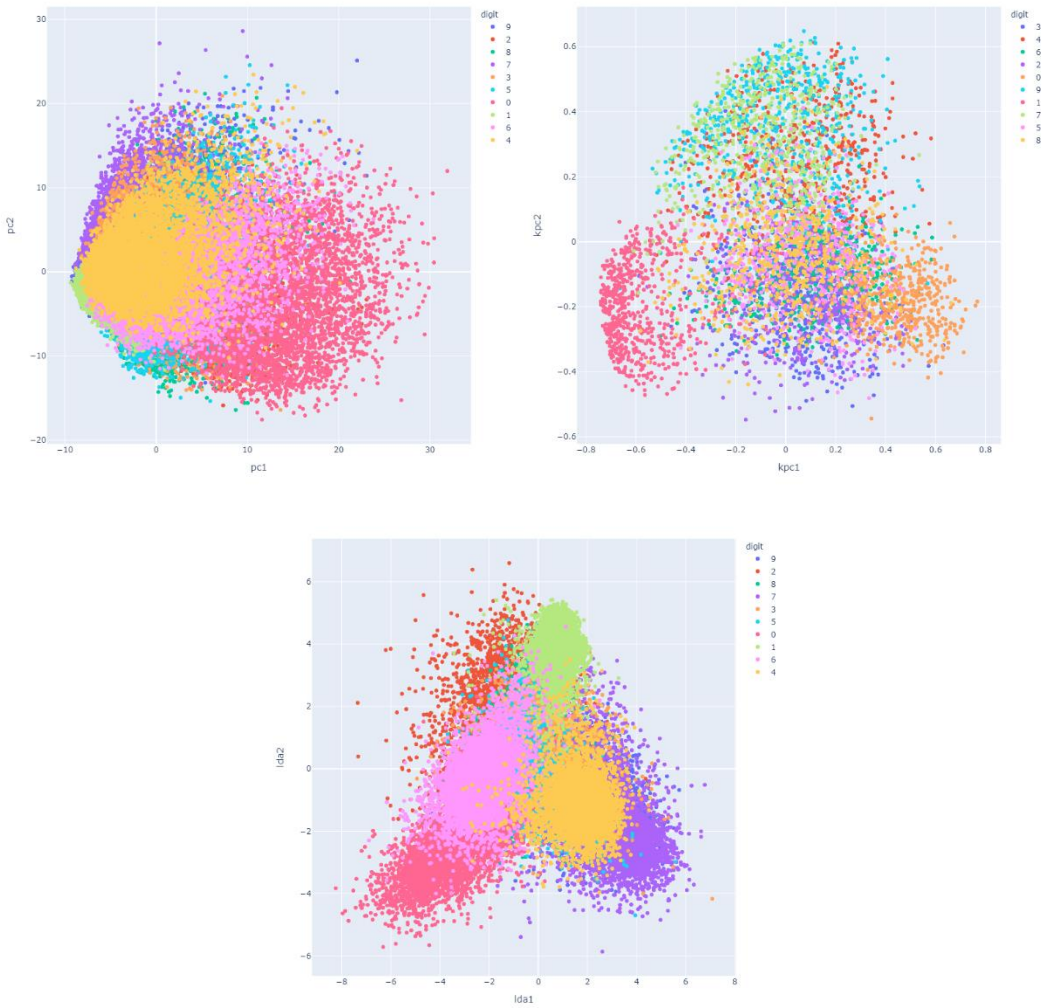

Figura 2.4: Otras representaciones de los dígitos en los primeros dos componentes, en la parte superior PCA y Kernel PCA con kernel tipo coseno, y en la parte inferior LDA.

Observe la Figura 2.4 y note que la mejor representación en menor dimensión es LDA, ya que algunos conjuntos de las clases se notan ligeramente separados del resto en los extremos, y más juntos entre si los que son del mismo tipo, por lo tanto, se eligió esta representación y se repitió el mismo procedimiento.

Observe la Tabla 2.2 y note que mientras que para el baseline y LDA se mantiene igual el RMSE, para QDA disminuyó y lo mismo se puede observar en la Figura 2.2 y 2.6 donde R^2 pasó de 0.146 a 0.776 (la precisión también mejoró considerablemente) y se tomó como el mejor modelo. Esto tiene sentido, si observamos en la representación en los primeros dos componentes que parecen normales, se nota una forma elíptica con tamaños y direcciones diferentes, por lo tanto, se cumple el supuesto de que las varianzas son diferentes para el modelo QDA.

	Train set	*Test set*
Baseline	1.57	1.62
LDA	1.47	1.51
QDA	1.34	1.36

Tabla 2.2: RMSE para cada modelo (renglones) para cada conjunto de datos (columnas) utilizando la representación de LDA.

Por último, estos resultados, se probaron utilizando una aplicación interactiva y con el 50% del conjunto de los datos balanceados para entrenamiento (Figura 2.5).

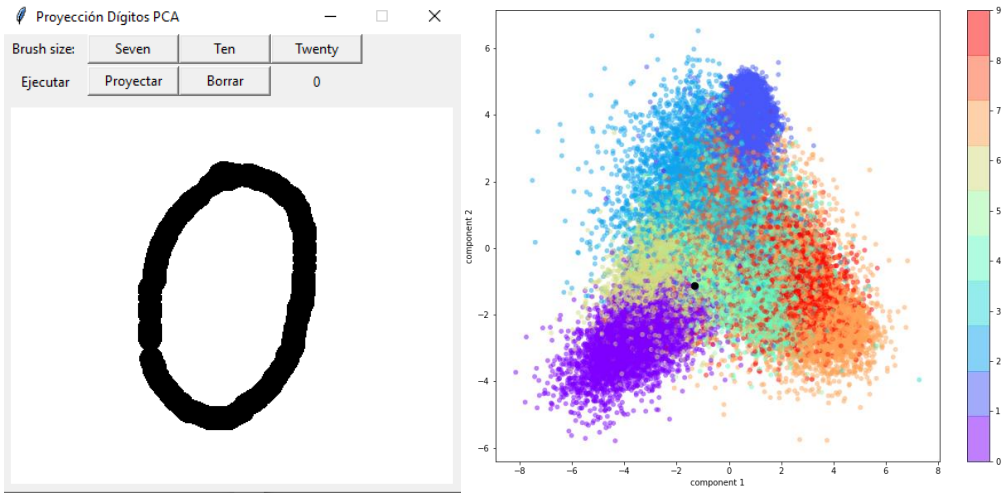

Figura 2.5: a) Aplicación interactiva para dibujar un digito y te dice cuál es, b) la proyección del digito (punto negro) en los primeros 2 componentes de la representación LDA.

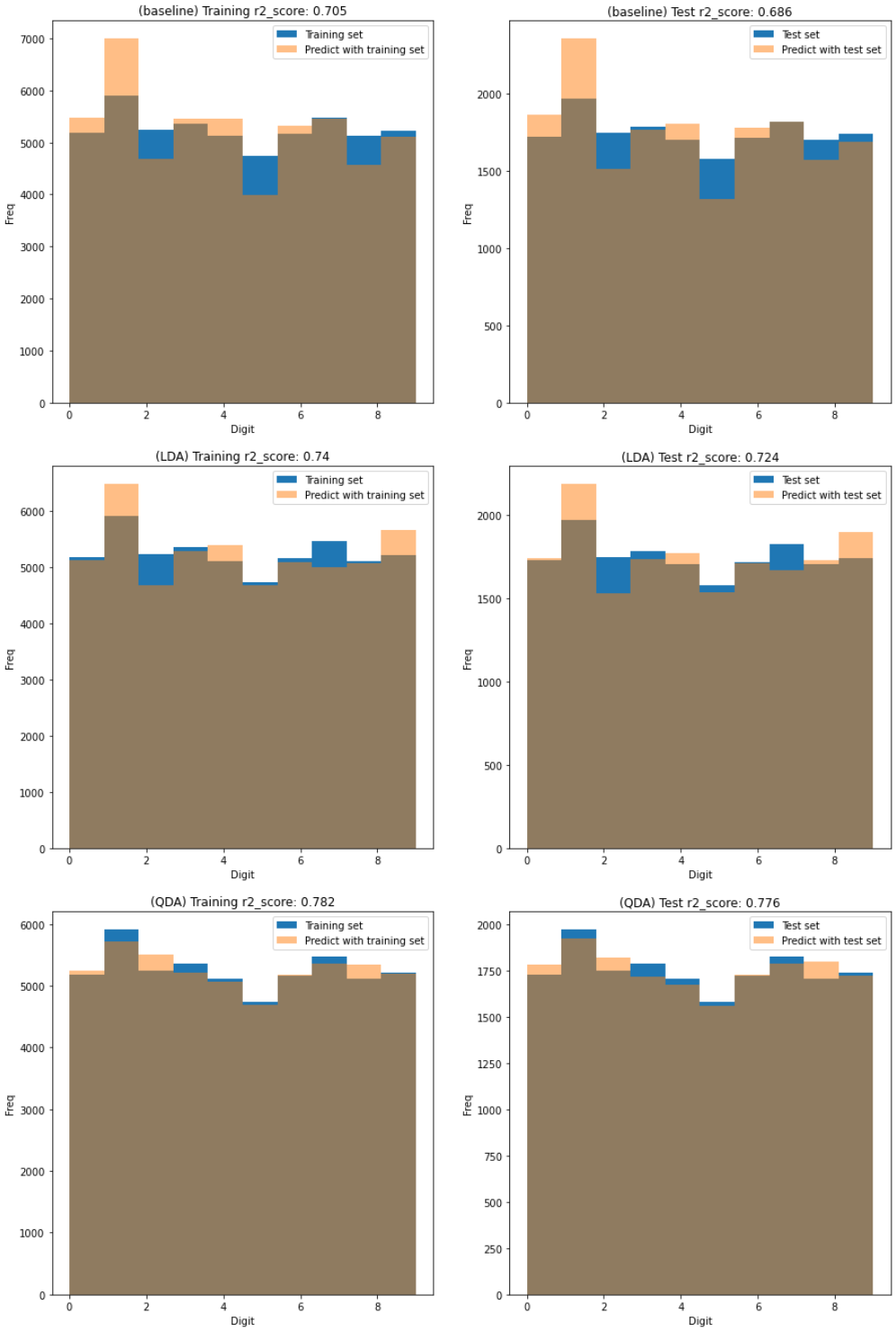

Figura 2.6: Histogramas de las clases reales y de las respuestas con su respectivo R^2, para cada modelo utilizando la representación LDA (renglones), para cada conjunto de datos (columnas).

3. Problema 3

En este ejercicio se emplearon los datos de 400 opiniones para 9 productos: automóviles, hoteles, lavadoras, teléfonos celulares, computadoras, libros, música y películas, que se dividieron en 2 conjuntos uno de entrenamiento y el otro de prueba (80% y 20% respectivamente), ambos conjuntos están bien balanceados, además ambos, tienen asociado un sentimiento, positivo o negativo (yes, no). De lo anterior podemos notar que claramente hay 2 categorías en los productos, una la llamaremos *things*, que abarca el conjunto de automóviles, hoteles, lavadoras, teléfonos celulares, computadoras y al otro *hobby,* para libros, música y películas.

La primera etapa para analizar esta información consistió en cargar ambos conjuntos, y hacer un preprocesado a los textos de ambos, quitar acentos, signos o elementos raros, convertir a minúsculas, o dejar la raíz de las palabras o quitar conectores. Posteriormente, se obtuvo una representación vectorial con el conjunto de entrenamiento, es decir se obtuvo un diccionario de las palabras que había en esos textos y la frecuencia con la que cada palabra aparece en cada texto tanto del conjunto de entrenamiento como al de prueba.

Una vez obtenidas las matrices de términos por documento, se pasó a hacer un filtrado solo de las palabras más representativas, para disminuir la dimensión de las matrices. Originalmente las matrices tenían ambas 10,000 palabras, pero haciendo un análisis similar al de PCA, donde nuestra información es la frecuencia de estas palabras entonces podemos obtener un diagrama similar al de la Figura 3.1, donde notará que con 2000 palabras tenemos aproximadamente el 80% de la información más representativa.

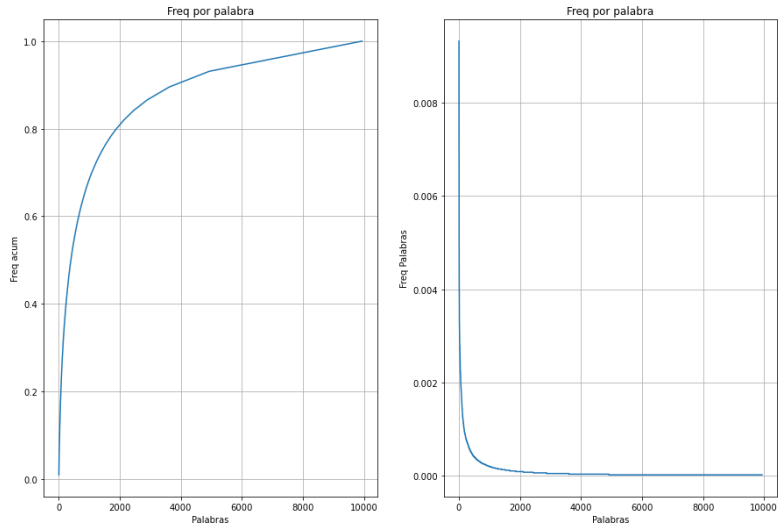

Figura 3.1: Frecuencia acumulada y frecuencia por palabra.

Una vez hecho esto para ambos conjuntos, se procedió a estandarizar y aplicar Kernel PCA, con un kernel de tipo coseno con 5 componentes principales para posteriormente explorar las proyecciones de estos 5 componentes por categoría de producto, la nueva categoría que señalamos previamente y por sentimiento, donde se encontraron principalmente 2.

Observe la Figura 3.2 y note que las dos categorías se separan muy bien (things y hobby), mientras que para sentimiento no, hay una gran dispersión de ambos en todo el espacio.

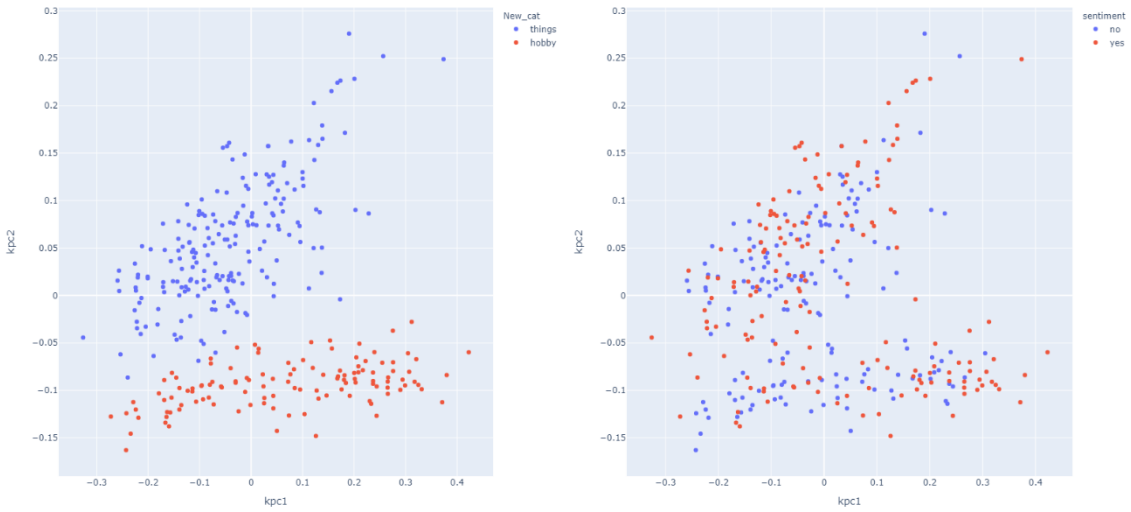

Figura 3.2: Proyección en los primeros 2 componentes, a) nueva categoría (cosas y hobbies), b) sentimiento.

Ahora observe la Figura 3.3, y note que, aunque para el sentimiento tenemos el mismo inconveniente, para los productos se hay una separación, en la parte del centro están los de tipo hobbies y en los extremos los de tipo de cosas.

Debido a lo anterior es que se decidió utilizar estas dos proyecciones en posteriores análisis, la proyección en los primeros dos componentes para hacer el análisis de sentimientos y la de los componentes 4 y 5 para el query.

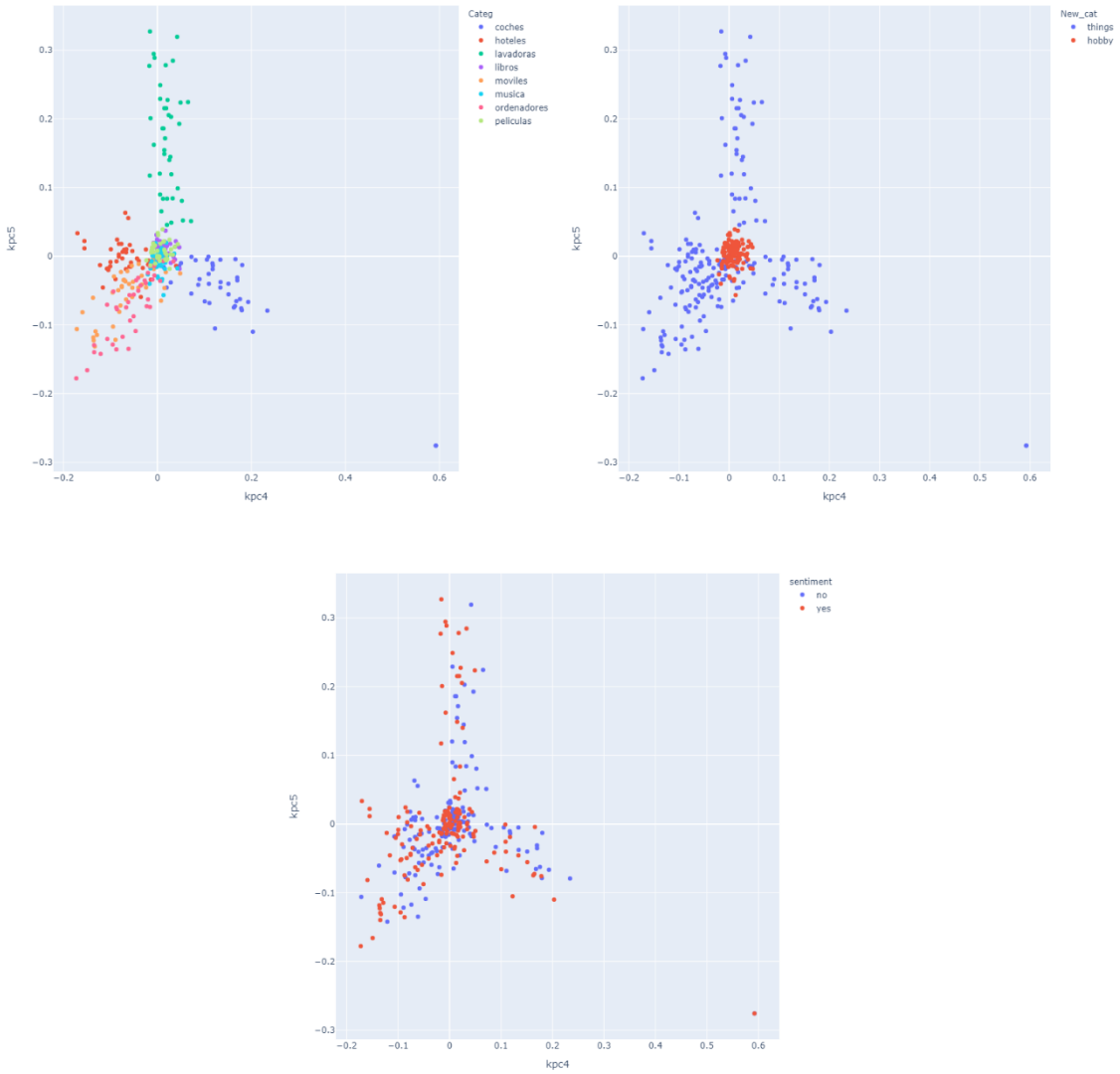

Figura 3.3: Proyección en los componentes 4 y 5, parte superior a) categorías de producto, b) nueva categoría, parte inferior c) sentimiento.

Ahora aplicamos algunos algoritmos de clustering sobre estas proyecciones para tratar de identificar categorías en ellas. Observe la Figura 3.5 y note que ninguno de los métodos de clustering separa bien alguna de las dos categorías. Observe también la Figura 3.6 y note ahora que, para la mayoría de los métodos, si se encuentra algo similar, algo bastante parecido a lo que se observa en la Figura 3.3 a), donde en el centro tenemos una categoría principal (hobbies) y en los extremos la de cosas.

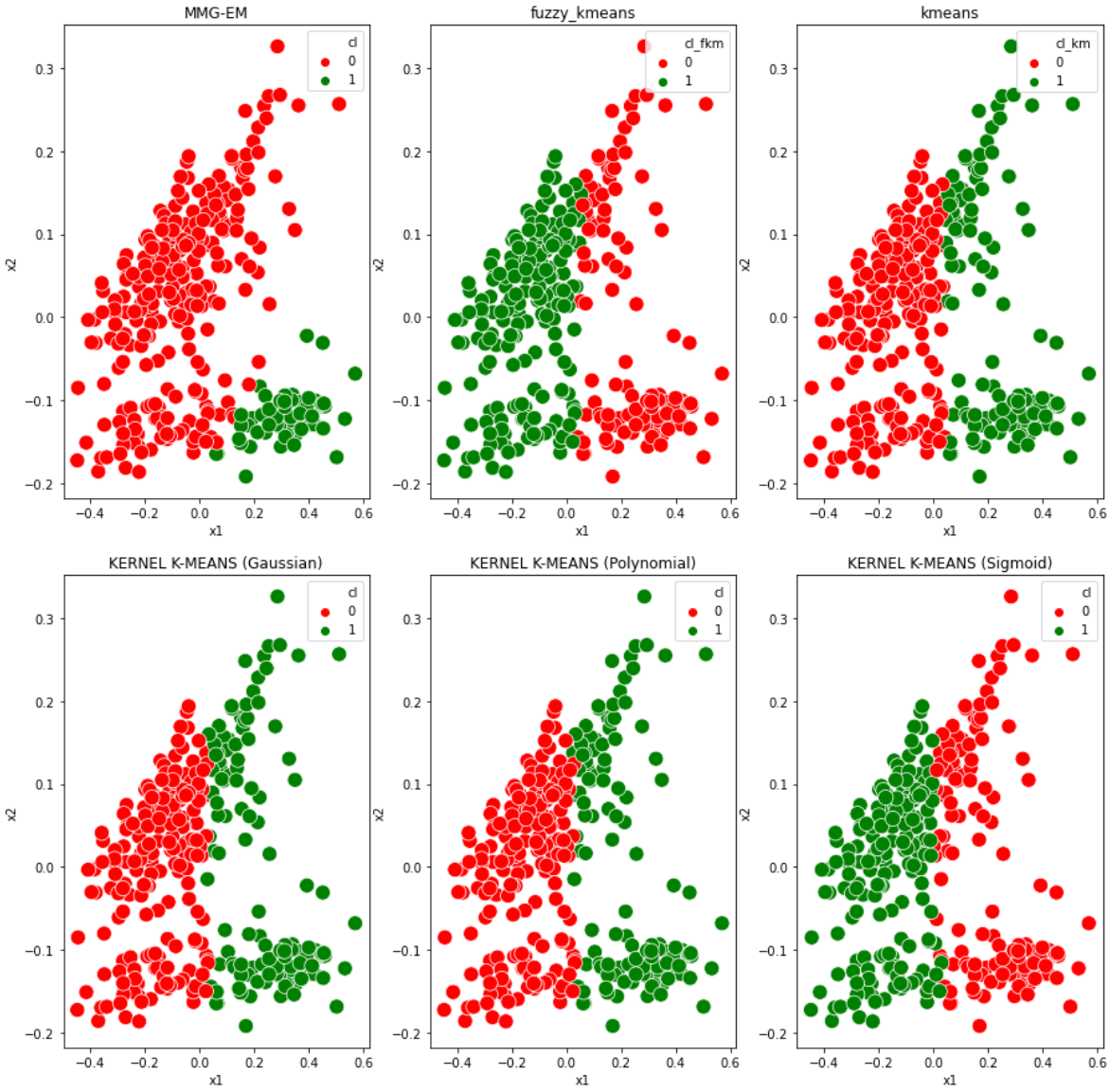

Figura 3.4: Proyección en los primeros dos componentes, con diferentes tipos de clustering.

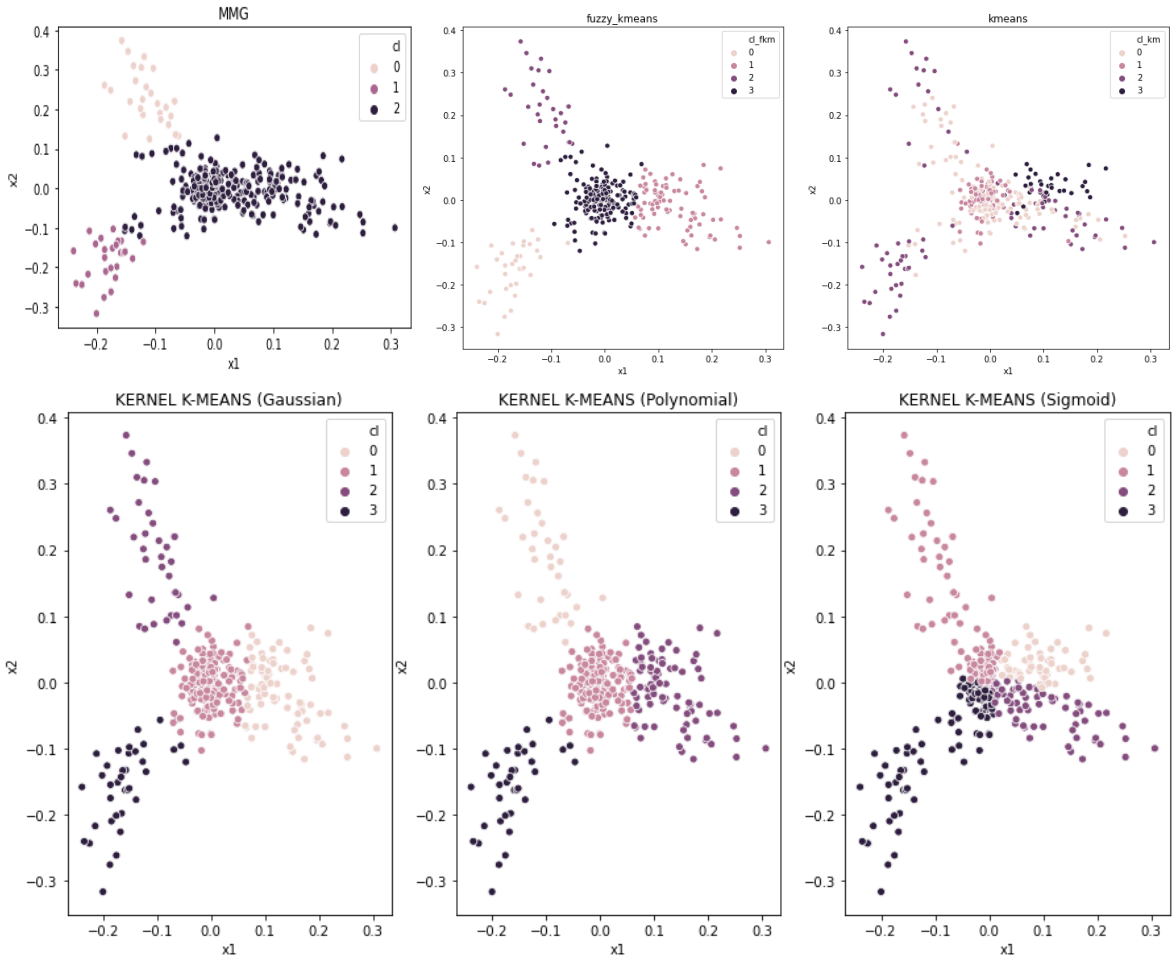

Figura 3.5: Proyección en los componentes 4 y 5, con diferentes tipos de clúster.

Para hacer la elección de cuantos clústers buscar por proyección (que se observan en las Figuras 3.4 y 3.5) se utilizó el grafico de codo, de la inercia por número de clústers para cada proyección, con el método de K-means (Figura 3.6).

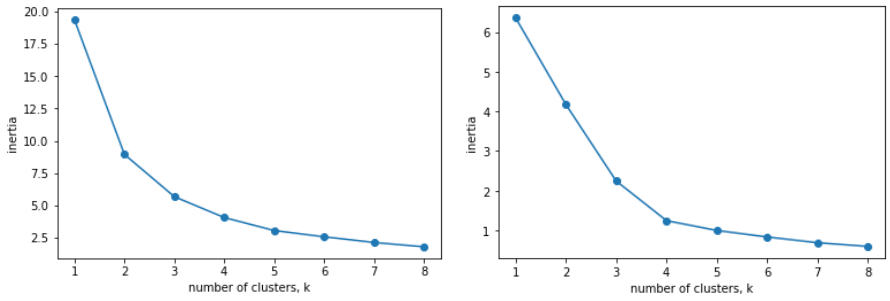

Figura 3.6: Diagrama de inercia por cluster para decidir numero de clusters para a) la proyeccion en los primeros dos componentes, b) la proyeccion en los componentes 4 y 5.

Se implementó un método de recuperación de información, para que, dado un texto de prueba, devolviera automáticamente los 5 documentos del conjunto de entrenamiento más relacionados.

El funcionamiento de este método consistió en repetir el procedimiento anteriormente explicado, es decir, dado un texto nuevo de prueba, se procesó (o limpió), se llevó al mismo espacio que el conjunto de entrenamiento, es decir, se vectoriza mediante la frecuencia de las palabras en base al mismo diccionario del conjunto de entrenamiento, se estandariza y se le aplica Kernel PCA, para llevarlo al mismo espacio. En este caso se proyecta en los componentes 4 y 5, pues es donde más claramente hay una distinción entre las categorías y por último se busca los 5 textos más cercanos que se encuentren a su alrededor con el método de vecinos cercanos (NearestNeighbors). Algunos ejemplos:

Texto1: Los saltos con breves reflexiones, de lectura ágil y sin problemas de comprensión de una historia que se cuenta hacia atrás y en pequeñas dosis. Al libro le sobran páginas, divagaciones que no aportan demasiado, sin embargo, me gusto bastante.

Respuesta1: peliculas_yes_5_11.txt', 'libros_yes_4_23.txt', 'libros_yes_5_10.txt', 'libros _yes_5_6.txt', 'musica_no_2_11.txt

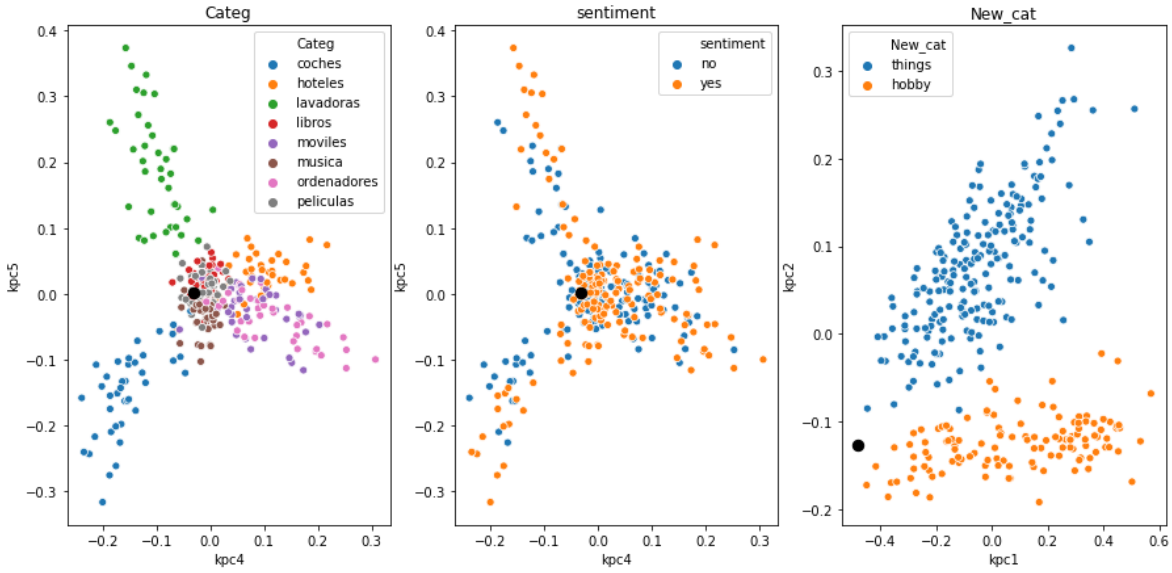

Figura 3.7: Proyecciones del query de libros del ejemplo 1.

Texto2: Hace dos meses que lo compré y estoy totalmente decepcionado. El cambio de mar chas que venden como automático es en realidad un coche con marchas. Cuando cambia de una a otra parece que se ahogue. En las subidas si sueltas el freno se va para atrás y como cambia a muy bajas vueltas tiene tan poca fuerza que parece que se para. Vamos que cualq uier parecido con el manejo de un coche automático es pura coincidencia. Si comparas con c ualquier coche de unos 70 CV te darás cuenta de que este ni se mueve. Conduciendo yo sol

o tiene dificultad para pasar de 110 km/h. Si utilizas el cambio en modo anual parece que v a bastante mejor, pero si yo supiese conducir un coche manual no me lo habría comprado a utomático. Eso sí, gasta poco y se aparca fácilmente. Por cierto, tiembla una barbaridad y hace un ruido de mil demonios. Vamos, por 11900 euros yo no me lo compraba. Por cierto, no lo probé por que en ningún sitio lo tenían. Ya saben lo que hacen, ya.

Respuesta2: 'coches_no_1_11.txt', 'coches_no_1_18.txt', 'coches_no_1_4.txt','coches_n o_2_9.txt', 'coches_no_2_20.txt'

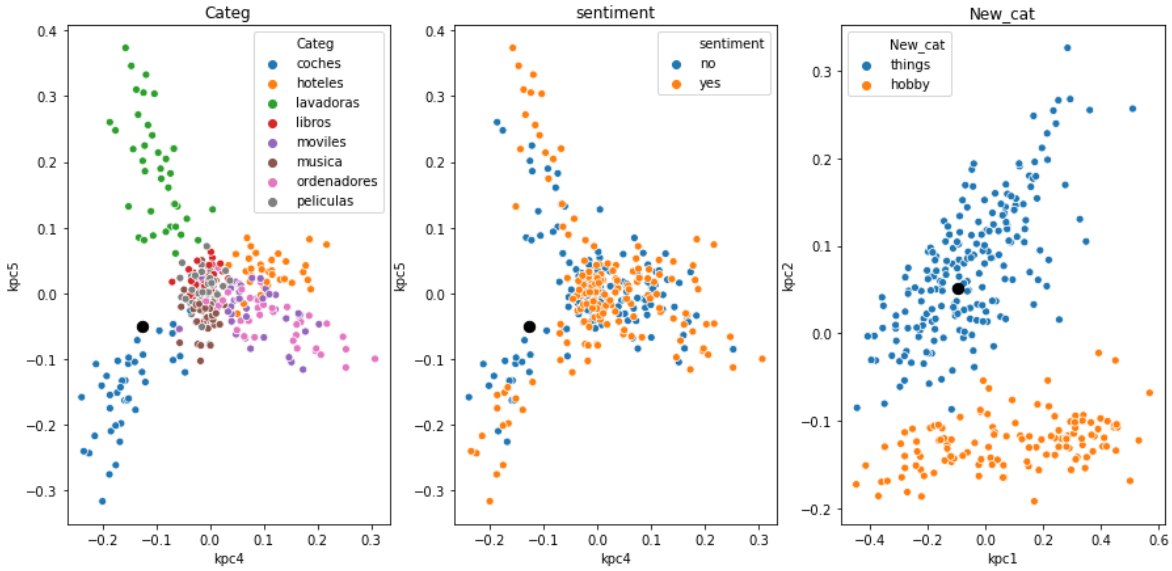

Figura 3.8: Proyecciones del query de coches del ejemplo 2.

Texto3: Ciertamente en los años ochenta apareció en nuestro mundo musical una pareja de hermanos Joaquín y Lucia que cual pareja nos da ciertas lecciones y pelas como si fuesen es posos mostrándonos realmente fuera de tener unas letras muy conflictivas entre dos "esposo s", "novios" o "enamorados", una gran dicción de voz especialmente el de Lucia galán que nos da una voz de soprano. En el único tema en que participa una tercera persona que recu erdo fue Dyango (para los que no saben padre de Jordi y Marcos Llunas) que canto "Ese ho mbre" donde se da una especie de enfrentamiento entre dos hombres creando un triángulo a moroso. En fin, en función a ese tipo de música o mejor dicho de letras y con música bajo e l ritmo de balada nos presenta este nuevo trabajo con el mismo estilo que les caracteriza. C iertamente esto es de hace mucho tiempo, ahora creo no se oye nada de ellos creo que han d ejado la música y se dedican a sus parejas no lo sé, ya no se les oye.

Respuesta3: 'musica_yes_5_24.txt', 'musica_yes_5_14.txt', 'musica_no_2_12.txt','music a_yes_5_25.txt', 'musica_no_2_25.txt'

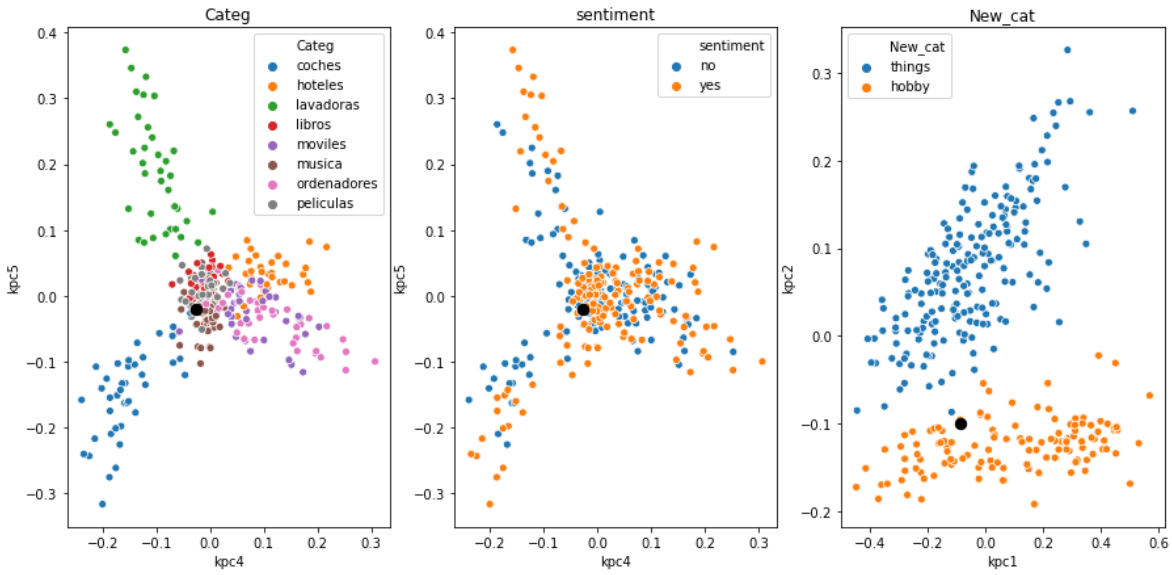

Figura 3.9: Proyecciones del query de música del ejemplo 3.

Note que, en los 3 ejemplos, se regresa también la proyección del query en las 2 principales representaciones que vimos anteriormente. Observe para el primer ejemplo sobre libros, positivo, en la Figura 3.7 la proyección cae muy cerca del centro y cerca de los positivos, pero en medio entre las categorías hobbies y cosas. Para el ejemplo 2 de query de coches, negativa, la proyección cae en la categoría de coches, muy cerca de los negativos y en la categoría de cosas (Figura 3.8). Para el último ejemplo que es de música, opinión positiva, la proyección cae muy cerca del centro y rodeado de los positivos, y cae en la categoría de hobbies.

Por último, se aplicó LDA y QDA para clasificar las opiniones positivas y negativas utilizando los conjuntos de entrenamiento y de prueba anteriormente mencionados y se utilizaron las mismas métricas que en el problema anterior. A continuación, se muestran los resultados:

	Train set	*Test set*	*Accuracy (Test set)*
LDA	0.055	0.5	0.75
QDA	0.055	0.680	0.54

Tabla 3.1: RMSE para cada modelo (renglones) para cada conjunto de datos (columnas) utilizando la representación de LDA. Precisión para cada modelo del Test set.

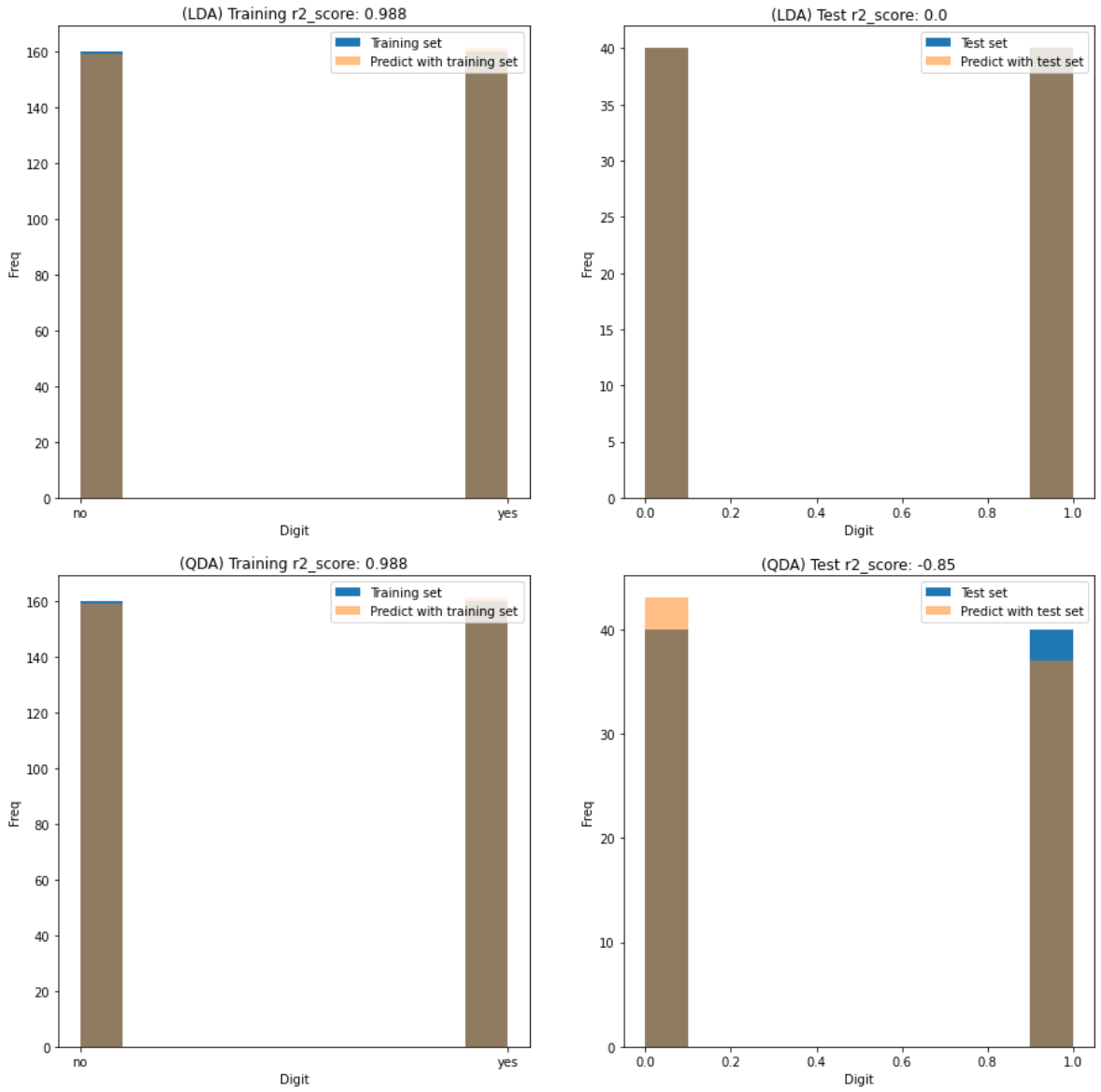

Figura 3.10: Histogramas de las clases reales y de las respuestas con su respectivo R^2, para cada modelo (renglones), para cada conjunto de datos (columnas).

	LDA		**QDA**	
REAL/PRED	Pred: 0	Pred: 1	Pred: 0	Pred: 1
REAL: 0	30	10	23	17
REAL: 1	10	30	20	20

Tabla 3.2: Tabla de Falsos positivos y Falsos negativos (Tabla de confusión) por modelo para el conjunto de entrenamiento. 0 para sentimiento negativo y 1 para positivo.

En los resultados anteriores podemos notar que el mejor modelo es el LDA, de forma global tiene una precisión del 0.75%, sin embargo los resultados del RMSE para los conjuntos de entrenamiento y de prueba son muy diferentes (Tabla 3.1), mientras que para el de entrenamiento es muy bajo para ambos modelos, el de prueba es muy alto, incluso el R^2, sale negativo o casi cero para el conjunto de prueba (Figura 3.10), esto quiere decir que ninguno de los dos modelos predice bien para muestras desconocidas o fuera del conjunto de entrenamiento.

Una idea para explicar lo anterior, es que por un lado ambos conjuntos son muy pequeños y quizá el modelo necesita más información para ajustarlo, y también una mayor cantidad para el conjunto de prueba. Por otro lado, quizá también lo afecte la calidad de los textos de entrenamiento, por ejemplo, puede que no sean muestras de calidad, o que la representación empleada no fue la mejor. Respecto a lo último, eso podría tener sentido ya que como pudimos ver en las representaciones (Figura 3.3) los sentimientos están muy dispersos entre sí, parecen una mezcla, no hay una representación adecuada que los separe y si recordamos que LDA y QDA se basan en el supuesto de normalidad con matriz de varianzas y covarianzas iguales o diferentes, respectivamente, si estas clases se encuentran mezcladas, entonces es complicado que logren los modelos hacer una clasificación adecuada (Tabla 3.2).