Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Identificación de números escritos a mano

Trabajo Final integrador de la Carrera de Especialización en Estadística

Autor:
Mauro Esteban Lioy

Febrero 2022

Índice

1. Introducción

Este trabajo profundiza uno de los problemas estudiados en el *Taller de Estadística 2021* en el cual se pretende resolver un problema de clasificación a través de distintos enfoques. En esta ampliación se incorporó el tratamiento y análisis de enfoques como `Linear and Quadratic Discriminant Analysis` (LDA y QDA), además de los previamente explorados: `baseline CART`, `RandomForest` y `Regresión logística multinomial regularizada` aplicados con las librerías *tree* [1] *RandomForest* [2] y *glmnet* [3], respectivamente.

En la siguiente sección se expone el problema a tratar y la descripción de los datos fuente. En la SECCIÓN 3 se postulan las métricas con las que se evalúan los modelos. La SECCIÓN 4 contiene una descripción de los modelos utilizados y las librerías implementadas. En la SECCIÓN 5 se presenta una discusión sobre las condiciones para aplicar LDA y QDA en relación al comportamiento de los datos utilizados. Por último en la SECCIÓN 6 se resumen los resultados obtenidos.

2. El problema

La situación a resolver es la siguiente: Para cada uno de los 10 dígitos (0,...,9) se tienen 200 imágenes digitalizadas del dígito escrito a mano. De cada imagen se obtuvieron 649 características ("features"). El objetivo del trabajo es predecir el dígito correspondiente a una imagen, en función de sus características.
Este problema se encuadra dentro de los problemas de clasificación y al disponerse de los datos tagueados, es decir de una muestra que enlaza las observaciones con la clase, se constituye un problema supervizado.

2.1. Datos

Los datos se encuentran disponibles en el siguiente link y están conformados por seis conjuntos de características.

1. mfeat-fou: 76 Fourier coefficients of the character shapes,
2. mfeat-fac: 216 profile correlations,
3. mfeat-kar: 64 Karhunen-Love coefficients
4. mfeat-pix: 240 pixel averages in 2 x 3 windows
5. mfeat-zer: 47 Zernike moments
6. mfeat-mor: 6 morphological features

En total el set se conforma de 2000 observaciones y 649 features. Para cada clase (número) hay 200 observaciones. Entre el conjunto de características *mfeat-mor* se encuentras 3 variables categóricas de 3,7 y 6 niveles, que fueron tratadas según el modelo utilizado.

set	0	1	2	3	4	5	6	7	8	9
$\mathcal{D}$	129	147	144	136	146	143	148	134	132	141
$\mathcal{T}$	71	53	56	64	54	57	52	66	68	59

Cuadro 1: Cantidad de observaciones para cada clase en los conjuntos de entrenamiento y test.

A partir del dataset se conformó de forma aleatoria un conjunto de entrenamiento $\mathcal{D}$ (70 %) y se dejó el resto para el conjunto de test $\mathcal{T}$ garantizando la distribución de las clases. La tabla 1 muestra la cantidad de individuos, separados por clase en cada conjunto.

3. Métricas

Para evaluar la performance de los modelos se tomaron dos métricas globales sobre el set $\mathcal{T}$. Dado que los sets están balanceados se midió el error de clasificación definido de acuerdo a la ecuación 1:

$$EC_{\mathcal{T}} = \frac{1}{N_{\mathcal{T}}} \sum_{i=1}^{N_{\mathcal{T}}} I_{G_{\mathcal{D}}}(X_i^{\mathcal{T}}) \neq y_i^{\mathcal{T}} \tag{1}$$

donde $G_{\mathcal{D}}$ es el clasificador entrenado en $\mathcal{D}$.También se tomó la medida del F1-score global (o macro-F1) [4] calculado como el promedio del F1-score de cada clase, considerando una clase contra todas las demás (ecuación 2).

$$F1_{\text{global}} = \frac{1}{K} \sum_{k=0}^{K} F1_k \qquad F1_k = \frac{TP_k}{(TP_k + \frac{1}{2}(FP_k + FN_k))} \tag{2}$$

donde el índice k indica el cálculo sobre la clase k del los valores de True-Positive, False-Positive y False-Negative.
Adicionalmente se construyeron y observaron las matrices de confusión de cada modelo con el fin de identificar la existencia de alguna clase más perjudicada en la clasificación.

4. Modelos

4.1. Baseline-CART

Como baseline se tomó un modelo sencillo a partir de Classification And Regression Trees. Este modelo se consiguió con la librería `tree` en R y se utilizó Cross Validation para el parámetro óptimo de la función de Costo Complejidad. Se encontró un árbol de 13 hojas basado en 12 variables (figura 1).

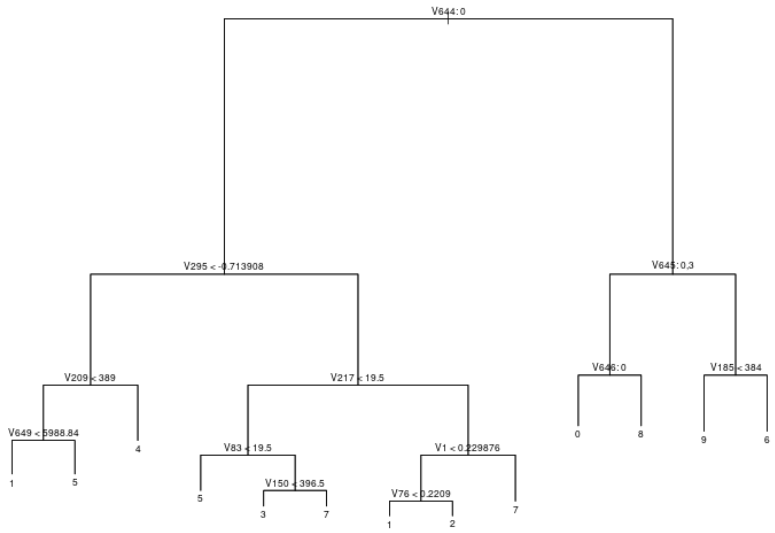

Figura 1: Esquema del modelo de árbol generado

4.2. RandomForest

RandomForest es un algoritmo que plantea una mejora con respecto a un modelo de árbol (CART) ya que se basa en un ensamble de árboles no correlacionados construidos a partir de los datos. Si bien el proceso de emsamble puede que pierda interpretabilidad, dado el máximo objetivo de predecir la clase, este algoritmo se presenta un modelo adecuado para el problema.
Para generar el modelo se utilizó la librería `randomForest` en R.

4.3. Regresión Logística Multinomial Regularizada

Fuera del enfoque de los árboles se decidió probar con un modelo lineal generalizado de la familia multinomial. Sobre este modelo se aplicó una regularización Lasso y una Ridge con el valor de lambda óptimo hallado a partir de Cross-Validation. Para generar el modelo se utilizó la librería `glmnet` en R.

4.4. Linear and Quadratic Discriminant Functions (LDA-QDA)

El proceso de clasificación pide asignar de forma óptima el valor de una clase k a una observación x, es decir que se quiere un modelos para $P(G = k|X = x)$.

Si se supone que $f_k(x)$ es la densidad de X condicionada a una clase k, y que π_k es la probabilidad a priori de la clase k, con $\sum_k = 1^k = 1$, luego por la aplicación del teorema de Bayes:

$$Pr(G = k|X = x) = \frac{P(X = x|G = k)P(G = k}{\sum_k P(X = x|G = k)P(G = k)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{3}$$

Por lo que una mayor probabilidad posterior a una clase k decide la pertenencia de la observación a la clase. Se observa que en términos de la habilidad para clasificar $f_k(x)$ se comporta igual que la cantidad $Pr(G = k|X = x)$. Muchos modelos de clasificación parten de esta configuración para modelar la probabilidad posterior[5]. LDA y QDA surgen de considerar una densidad gaussiana para $P(X = x|G = k)$. Entonces se supone para cada clase una densidad Normal Multivariada:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{\frac{1}{2}}} \exp -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) \tag{4}$$

La frontera de decisión entre dos clases k y l se construye al considerar la condición de igualdad de sus posteriores, es decir $Pr(G = k|X = x) = Pr(G = l|X = x)$. En el caso especial en que se asume que las covarianzas de cada clase son iguales, es decir $\Sigma_k = \Sigma \quad \forall k$ resulta un discriminante lineal en x (LDA - Ecuación 5), de los contrario se encuentra un discrimitante cuadrático (QDA - Ecuación 6).

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \tag{5}$$

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma^{-1}| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - mu_k) + \log \pi_k \tag{6}$$

En la práctica nos e conocen los parámetros de la distribución, por lo que se estiman a partir de los datos.

- $\hat{\pi}_k = N_k/N$ donde N_k es el número de observaciones de la clase k.
- $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N - K)$

La implementación de LDA y QDA se hizo a través de la librería `MASS` en R.

5. Discusión

Como se describió en la sección anterior el enfoque de LDA y QDA conlleva asunciones más restrictivas que los otros modelos mencionados. Se asume que para cada clase k los predictores de dimensión p se distribuyen de forma normal-multivariada, es decir $X_p^{(k)} \sim N_p(\mu_k, \Sigma_k)$; y en el caso de LDA se pide también

que $\Sigma_k = \Sigma \quad \forall k$.

Estas condiciones no son descriptivas del set de datos utilizado. Sin embargo, en la práctica se puede esperar de LDA un comportamiento aceptable incluso frente a la carencia de estas condiciones. Si se trabaja sobre un espacio que maximiza la distancia entre grupos en términos de varianza, el discriminante lineal puede ser óptimo para la clasificación.[5][6]

Por definición si un vector aleatorio de dimensión p sigue una distribución normal-multivariada $X_p^{(k)} \sim N_p(\mu_k, \Sigma_k)$ cada una de sus marginales se comportan con distribución normal univariada [7]. En este caso se pide que para cada clase k el vector de predictores $X_p^{(k)} \sim N_p(\mu_k, \Sigma_k)$ por lo tanto sus marginales se distribuyen según $X^(k)_i \sim N(\mu_{ki}, \sigma_i)$ donde μ_{ki} es el elemento i del vector de medias μ_k y σ_i es el elemento ii de la matriz de covarianzas $Sigma_k$. En la figura 2 se muestra la distribución de los p-valores de la prueba de normalidad univariada de Shapiro sobre las marginales condicionadas a cada clase. La prueba de Shapiro tiene como hipótesis-nula la normalidad de los datos, y si bien éste es un test de baja potencia, es decir con baja probabilidad de rechazo de la hipótesis-nula bajo una hipótesis alternativa cierta, se observa que la mayoría de las variables en cada clase tienen p-valores menores 0,05 (linea vertical azul en la figura).

Por completitud también se comprobó la falta de normalidad-mutivariada con los test Multivariados de Royston y Henze-Zirkler a través de la libreria *MVN* [8]. En ambos casos se obtuvieron p-valores del orden del cero [9]. La condición de homocedasticidad $\Sigma_k = \Sigma \quad \forall k$ se analizó con el BOX'S-M-TEST [10] implementado en la librería *biotools* y se verificó que no se cumple. Este test es sensible a la falta de normalidad, por lo que su resultado puede estar alterado por dicha condición.

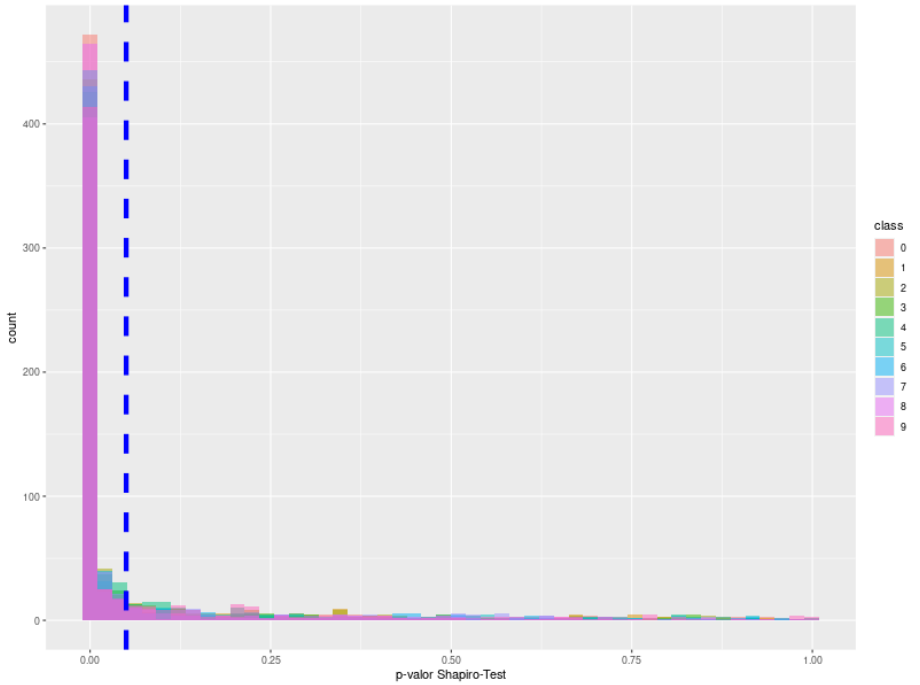

Figura 2: Distribución de los p-valores de la prueba de normalidad de Shapiro para las variables condicionadas a cada clase k. Se observa que el test rechaza la hipótesis nula (de normalidad) en la gran mayoría de las variables. La linea vertical azul remarca el valor 0.05.

En términos de implementación se encontró que los datos de entrenamiento $\mathcal{D}$ condicionados a una clase presentaban colinealidad entre variables, lo cual provocó problemas de singularidad para operar con la estimación de la matriz de covarianzas de la clase. Para tratar dicha situación y aportar un espacio con mejor representación de la varianza de los datos, se aplicó un proceso de rotación a través del análisis de componentes principales (PCA) y se redujo la dimensionalidad de 649 a 70 variables (CP). La cantidad de componentes principales surgió de la selección de componentes con varianza mayor que 1, ya que el PCA se aplicó sobre las variables reescaldas y se tomó el criterio de quitar las componentes que aportasen menos información que una variable original. Las 70 componentes acumulaban un 90 % del total de la varianza. Una vez encontradas las coordenadas de proyección, se aplicó la misma rotación al set de test $\mathcal{T}$. Sobre estos datos se aplicaron LDA y QDA.

En la figura 3 se muestran las regiones generadas por la intersección de los hiperplanos discriminantes obtenidos por LDA sobre las primeras 3 componentes (que acumulan el 34 % de la varianza). En la figura 4 se muestran las regiones con límites cuadráticos generadas por QDA sobre las mismas componentes. En

ambas figuras, si se observa con atención se puede notar la acumulación de puntos de una clase en las regiones generadas.

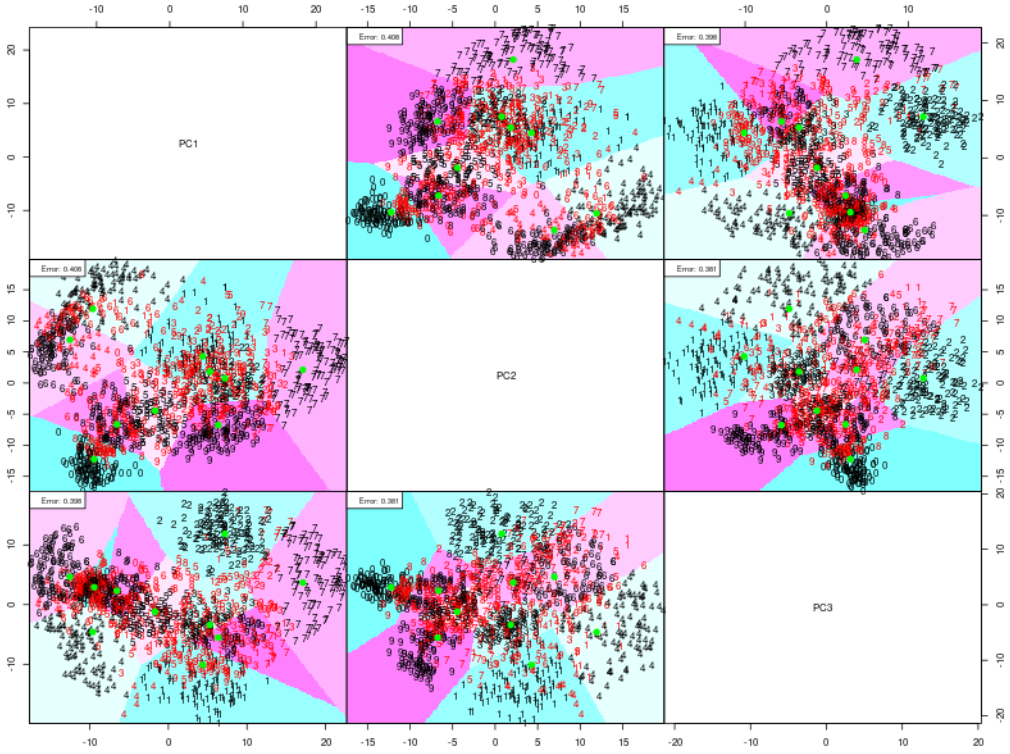

Figura 3: Regiones generadas por los discriminantes lineales LDA sobre las 3 primeras componentes principales que explican el 34 % de la varianza. Se observa la acumulación de puntos de la misma clase sobre las distintas regiones. En verde se plotea el centroide de cada grupos en esas componentes.

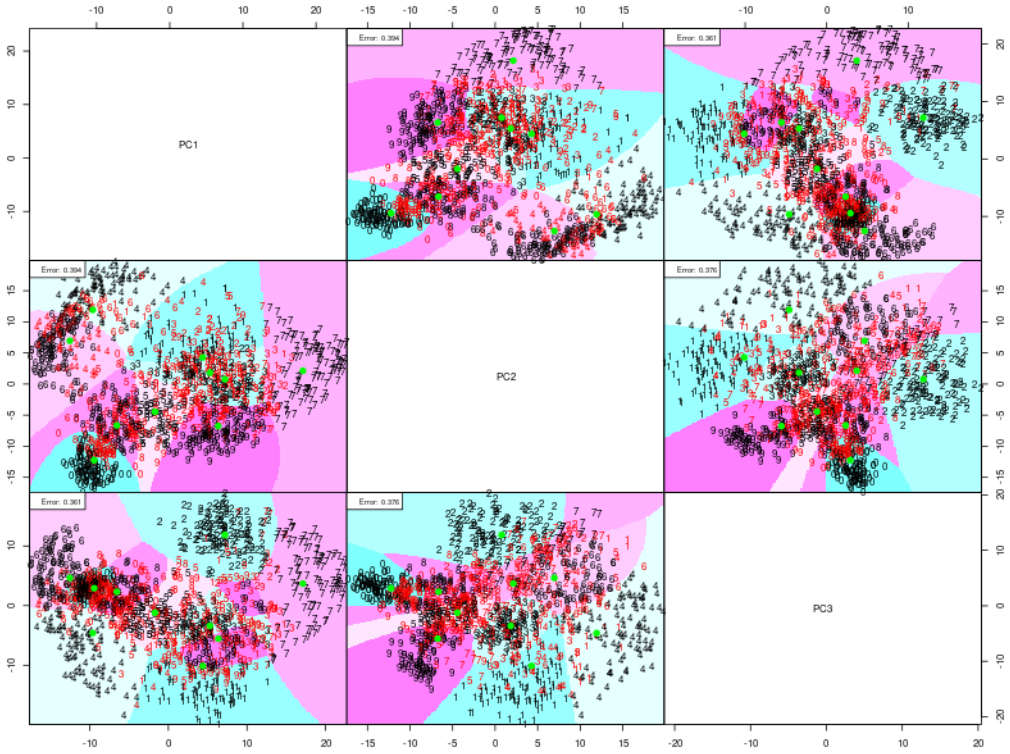

Figura 4: Regiones generadas por los discriminantes cuadráticos QDA sobre las 3 primeras componentes principales que explican el 34 % de la varianza. Se observa la acumulación de puntos de la misma clase sobre las distintas regiones. En verde se dibuja el centroide de cada grupo en esas componentes.

6. Resumen de resultados

La tabla 2 resume las métricas de los modelos estudiados. Como puede observarse, todos los modelos presentaron un performance aceptable muy superior al baseline. Si bien `RandomForest` siendo un modelo con mucha flexibilidad presentó numéricamente la mejor métrica, el modelo `PCA+LDA` con su enfoque lineal presentó un buen valor predictivo.

Modelo	$EC_{\mathcal{T}}$	$F1_{global}$
Baseline-CART	0.090	0.910
RandomForest	0.015	0.984
Regresión Logística Multinomial Lasso	0.018	0.980
Regresión Logística Multinomial Ridge	0.016	0.982
PCA+LDA	0.016	0.983
PCA+QDA	0.026	0.973

Cuadro 2: Métricas comparativas

A continuación, se muestran las matrices de confusión para cada uno de los modelos evaluados sobre el set $\mathcal{T}$. Desde el análisis del error no se evidencia cualitativamente la existencia de una clase especialmente perjudicada en la clasificación.

Clase	0	1	2	3	4	5	6	7	8	9
0	70	0	0	0	0	0	0	0	0	0
1	0	46	0	1	0	3	0	1	0	0
2	0	0	51	0	0	0	0	2	0	0
3	0	1	1	48	0	0	0	0	0	0
4	0	2	0	0	48	0	0	0	0	0
5	1	1	1	11	4	48	1	0	0	0
6	0	0	0	0	0	0	51	0	0	0
7	0	2	3	4	0	1	0	62	0	1
8	0	0	0	0	0	0	0	0	65	0
9	0	1	0	0	0	0	0	0	3	57

Cuadro 3: Matriz de confusión para el modelo **baseline-CART**. En columnas la clase verdadera y en filas la predicha.

Clase	0	1	2	3	4	5	6	7	8	9
0	71	0	0	0	0	0	0	0	0	0
1	0	52	0	1	0	0	0	0	2	1
2	0	0	55	1	0	1	0	0	0	0
3	0	0	0	62	0	0	0	0	0	0
4	0	0	0	0	54	0	1	0	0	0
5	0	0	0	0	0	56	0	0	0	0
6	0	1	0	0	0	0	51	0	0	0
7	0	0	1	0	0	0	0	66	0	0
8	0	0	0	0	0	0	0	0	66	0
9	0	0	0	0	0	0	0	0	0	58

Cuadro 4: Matriz de confusión del modelo por **RandomForest**. En columnas la clase verdadera y en filas la predicha.

Clase	0	1	2	3	4	5	6	7	8	9
0	70	0	0	0	0	0	0	0	0	0
1	0	51	0	0	0	0	0	0	1	1
2	0	0	56	1	0	0	0	2	0	0
3	0	0	0	62	0	0	0	0	0	0
4	0	1	0	0	53	0	0	0	0	0
5	0	0	0	0	0	57	0	0	0	0
6	0	1	0	0	1	0	52	0	0	0
7	0	0	0	1	0	0	0	65	0	1
8	1	0	0	0	0	0	0	0	67	0
9	0	0	0	0	0	0	0	0	0	57

Cuadro 5: Matriz de confusión del modelo por regresión **logística multinomial con regularización Ridge**. En columnas la clase verdadera y en filas la predicha.

Clase	0	1	2	3	4	5	6	7	8	9
0	71	0	0	0	0	0	0	0	0	0
1	0	52	0	1	1	0	0	0	0	1
2	0	0	55	1	0	0	0	0	0	0
3	0	0	1	62	0	0	0	0	0	0
4	0	0	0	0	52	0	1	1	0	0
5	0	1	0	0	0	57	0	0	0	0
6	0	0	0	0	1	0	51	0	1	0
7	0	0	0	0	0	0	0	65	0	0
8	0	0	0	0	0	0	0	0	66	0
9	0	0	0	0	0	0	0	0	1	58

Cuadro 6: Matriz de confusión del modelo por regresión **logística multinomial con regularización Lasso**. En columnas la clase verdadera y en filas la predicha.

Clase	0	1	2	3	4	5	6	7	8	9
0	70	0	0	0	0	0	0	0	1	0
1	0	51	0	0	0	1	0	1	0	0
2	0	0	56	0	0	0	0	0	0	0
3	0	1	1	62	0	0	0	0	0	0
4	0	0	0	0	53	0	0	0	0	0
5	0	1	0	0	0	56	0	0	0	0
6	0	0	0	0	1	0	52	0	1	0
7	0	0	1	0	0	0	0	65	0	0
8	0	0	0	0	0	0	1	0	67	0
9	0	1	0	0	0	0	0	0	0	58

Cuadro 7: Matriz de confusión del modelo por regresión **PCA+LDA**. En columnas la clase verdadera y en filas la predicha.

Clase	0	1	2	3	4	5	6	7	8	9
0	69	0	0	0	0	0	1	0	1	0
1	0	51	0	0	0	2	0	1	0	0
2	0	0	54	0	0	0	0	0	0	2
3	0	0	1	63	0	0	0	0	0	0
4	0	1	0	0	52	0	1	0	0	0
5	0	0	0	0	0	57	0	0	0	0
6	0	0	0	0	0	1	51	0	1	0
7	0	0	0	0	0	0	0	63	0	3
8	0	0	0	0	0	0	1	0	66	0
9	0	1	0	0	0	0	0	0	0	58

Cuadro 8: Matriz de confusión del modelo por regresión **PCA+QDA**. En columnas la clase verdadera y en filas la predicha.

Referencias

[1] Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[4] Juri Opitz and Sebastian Burst. Macro f1 and macro f1, 2021.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, chapter 4. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[6] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowl. Inf. Syst.*, 10(4):453–472, 2006.

[7] George A. F. Seber. *Multivariate observations*. Wily-Interscience, 1984.

[8] Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162, 2014.

[9] Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162, 2014.

[10] Knavoot Jiamwattanapong, Nisand Ingadapa, and Bandhita Plubin. On testing homogeneity of covariance matrices with box's m and the approximate tests for multivariate data. *European Journal of Applied Sciences*, 9(5):426–436, Nov. 2021.