Modelos Estadísticos para Clasificación y Regresión

Práctico 5 - Clasificación

IMERL - FIng

13 de setiembre de 2023

1 Repaso

2 Práctico 5

1 Repaso

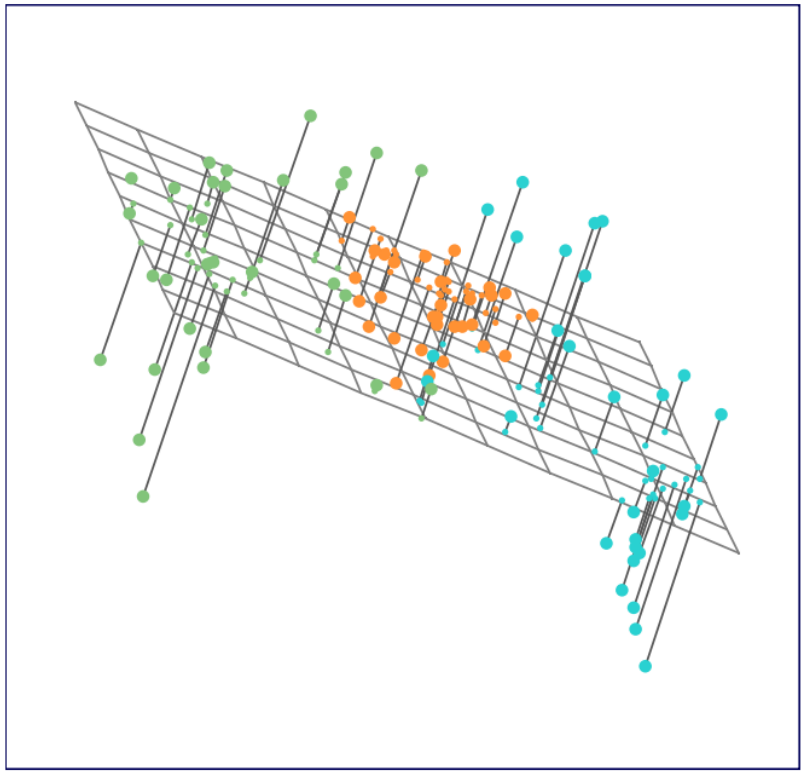
2 Práctico 5

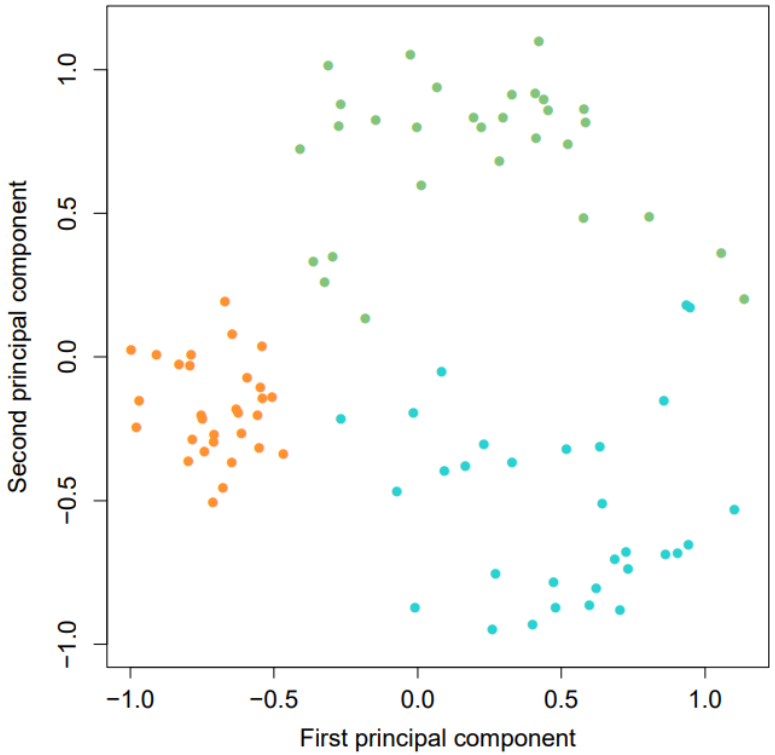
Análisis de Componentes Principales (*PCA*)

Cuando tenemos un dataset grande con atributos correlacionados, PCA nos permite 'resumirlo' utilizando una menor cantidad de atributos que colectivamente explican la mayor parte de la varianza del dataset.

Cada observación vive en un espacio p-dimensional, pero no todas las dimensiones son igual de 'interesantes'. PCA busca un número de dimensiones que sean lo más interesantes posible, en el sentido de la varianza de los datos.[1]

Análisis de Componentes Principales (*PCA*)





1

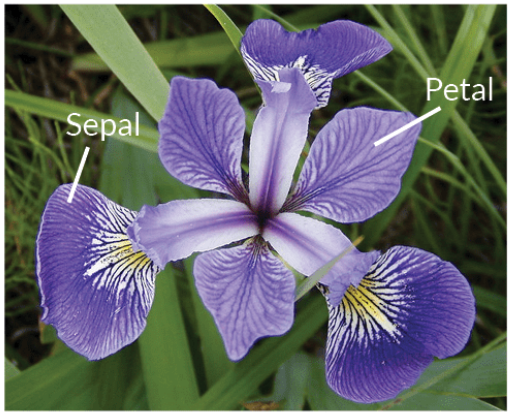
[1] Imagen extraída de [1].

Problema de Clasificación

Tenemos un conjunto de N datos de entrenamiento $x_1, x_2, ..., x_N$ de dimensión D. Para cada dato tendremos su etiqueta correspondiente $y_1, y_2, ..., y_N$ que describe a cuál de las C clases pertenece. Típicamente estas etiquetas corresponden a valores enteros (por ej: $\{1, 2, 3\}, \{0, 1\}, \{-1, 1\}$).
Objetivo: dado un nuevo dato x_{nuevo} poder predecir su clase y_{nuevo}.

[4]

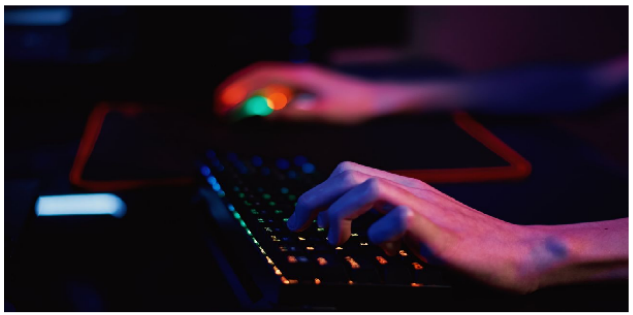
Ejemplos de clasificación



Iris Versicolor

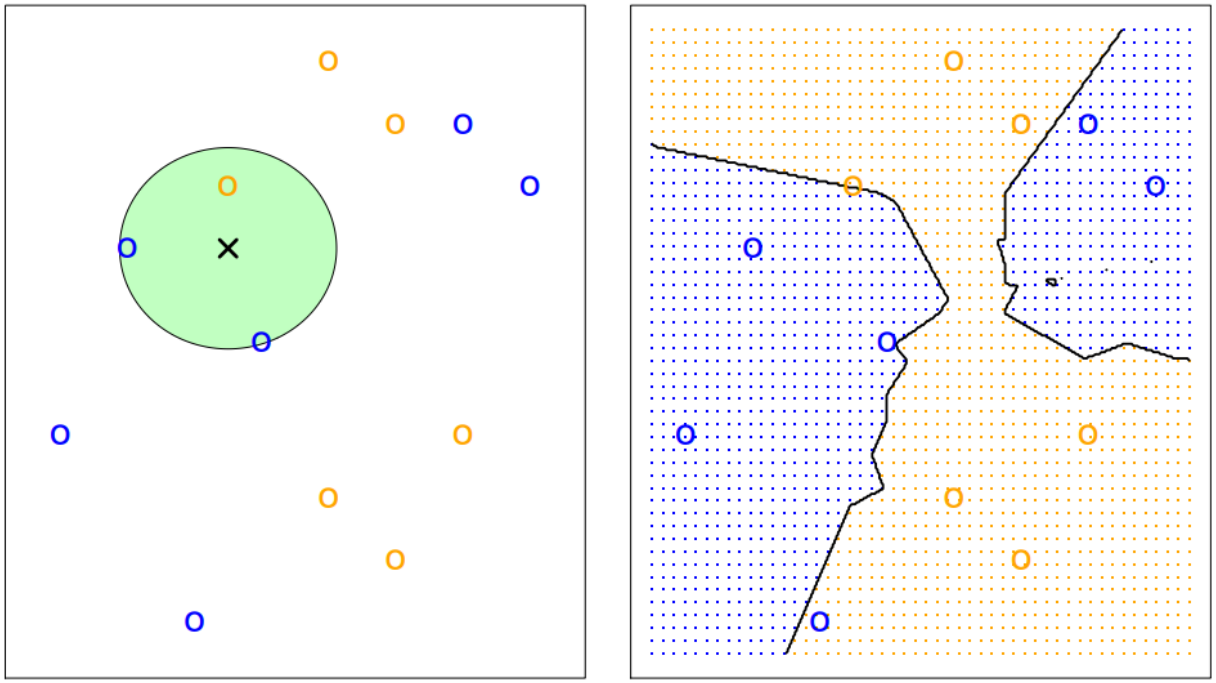
Iris Setosa

Iris Virginica



Vecinos más Cercanos (*kNN*)

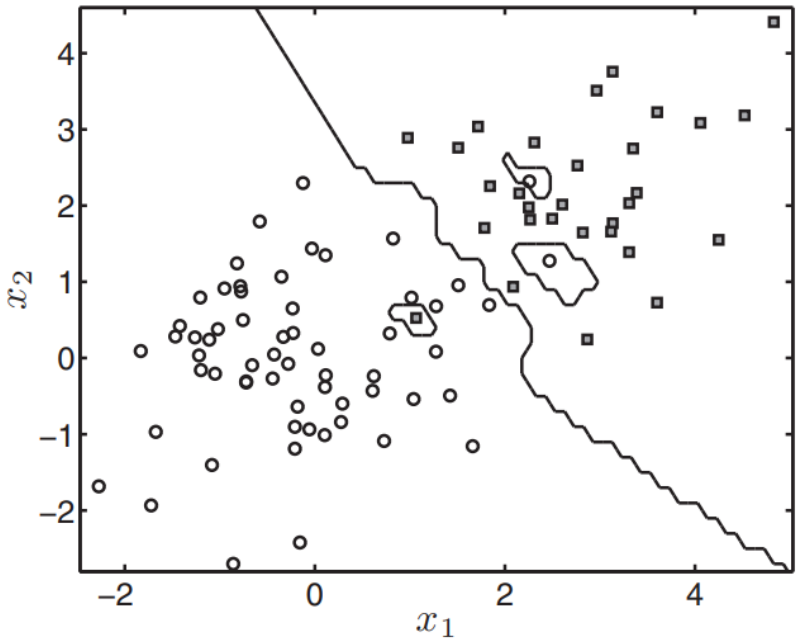
Para clasificar x_{nuevo} con *kNN*, hallamos los K puntos de entrenamiento que están más cerca de x_{nuevo}. Luego tomamos t_{nuevo} como la etiqueta de la mayoría de esos K puntos. [4].

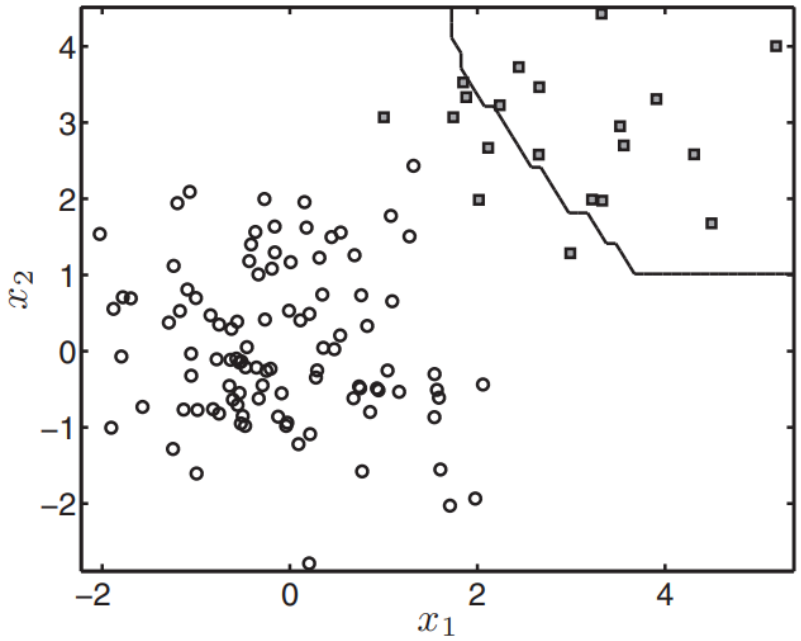


[2]Imagen extraída de [1].

kNN: elección de K

- ¿Qué pasa si tomamos K muy chico?
- ¿Y muy grande?



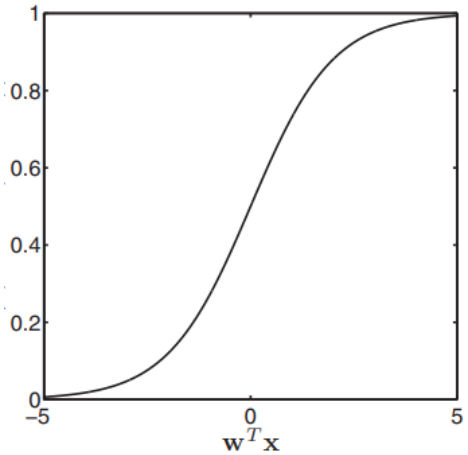


[3]Imagen extraída de [4].

Regresión Logística

El modelo lineal no permite modelar la probabilidad de que un dato X pertenezca a una cierta categoría, ya que podríamos obtener $p(X) > 1$ o $p(X) < 0$. Para solucionar esto, a la salida del modelo lineal $s = w^T X$ se aplica la función *sigmoide*:

$$p(X) = \frac{e^s}{1 + e^s}$$

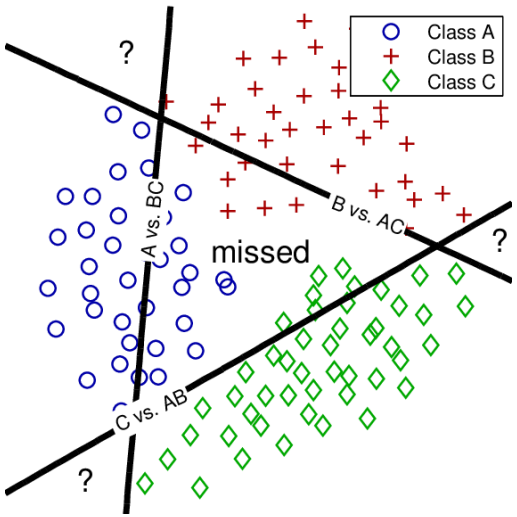


[4]

[4] Imagen extraída de [4].

Regresión Logística - esquema *OvR*

Si bien la Regresión Logística está pensada para problemas de clasificación binaria, su uso puede extenderse a clasificación multiclase usando esquemas *OvO* o *OvR*.
SkLearn implementa *OvR*:

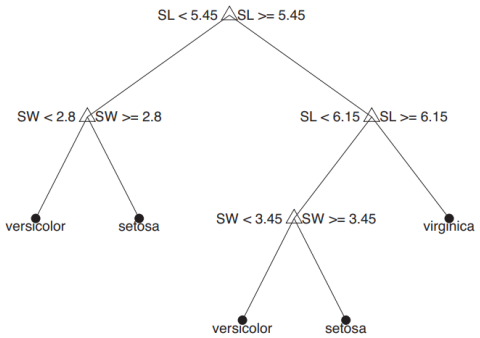


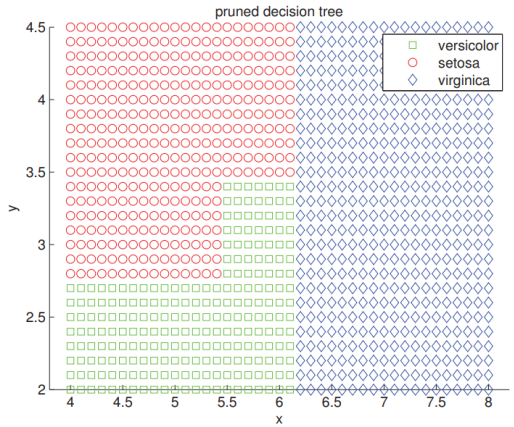
5

[5] Imagen extraída de [3].

Árboles de Decisión

Los árboles de decisión permiten particionar el espacio en regiones simples. El nombre proviene de que las reglas que se usan para segmentar el espacio se pueden resumir en un árbol.





Una vez entrenado el árbol, se predice la clase de un nuevo dato x_{nuevo} a partir de la clase mayoritaria de los datos de entrenamiento que caen en la misma hoja que x_{nuevo}.
[6]

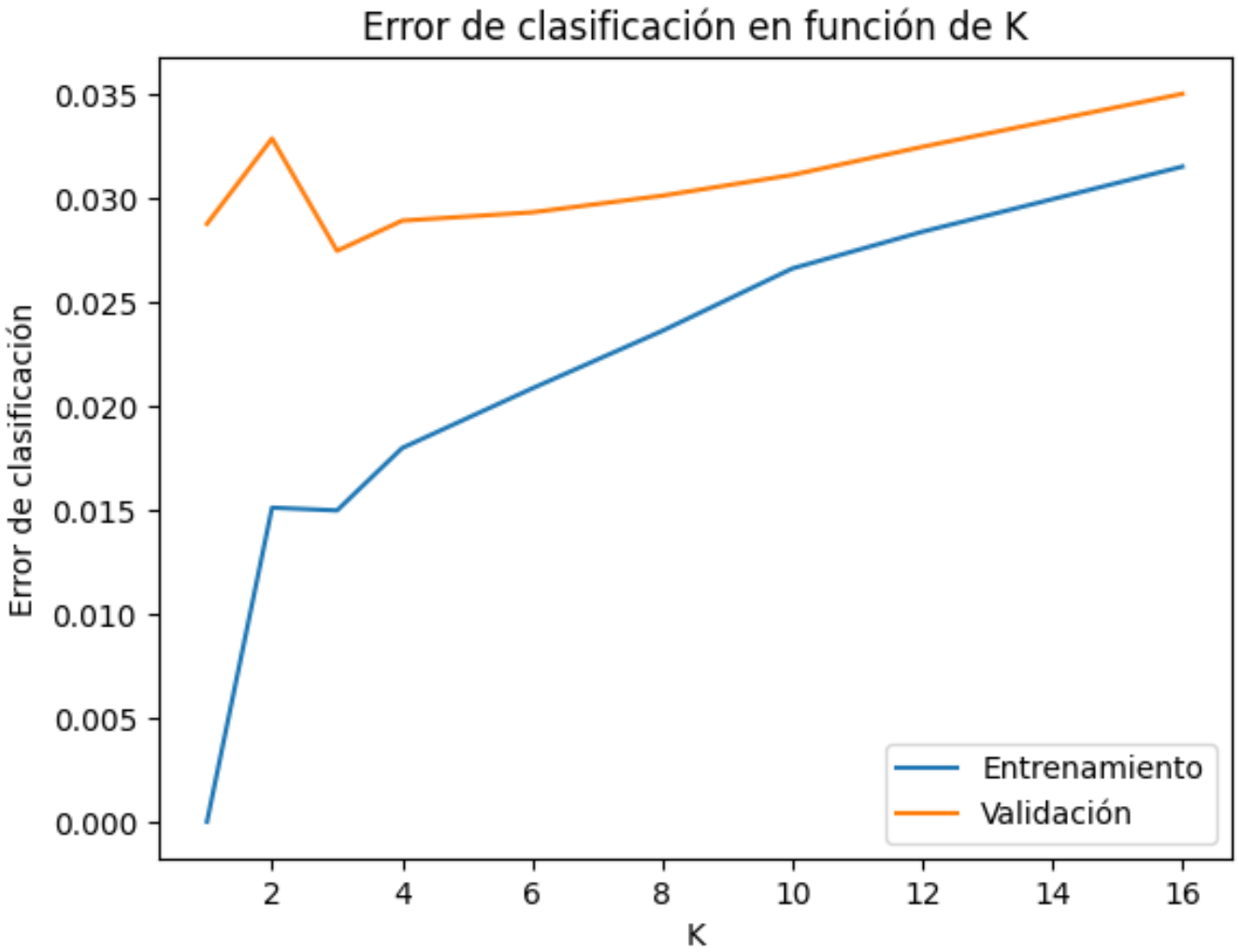
[6] Imagen extraída de [2].

Preguntas

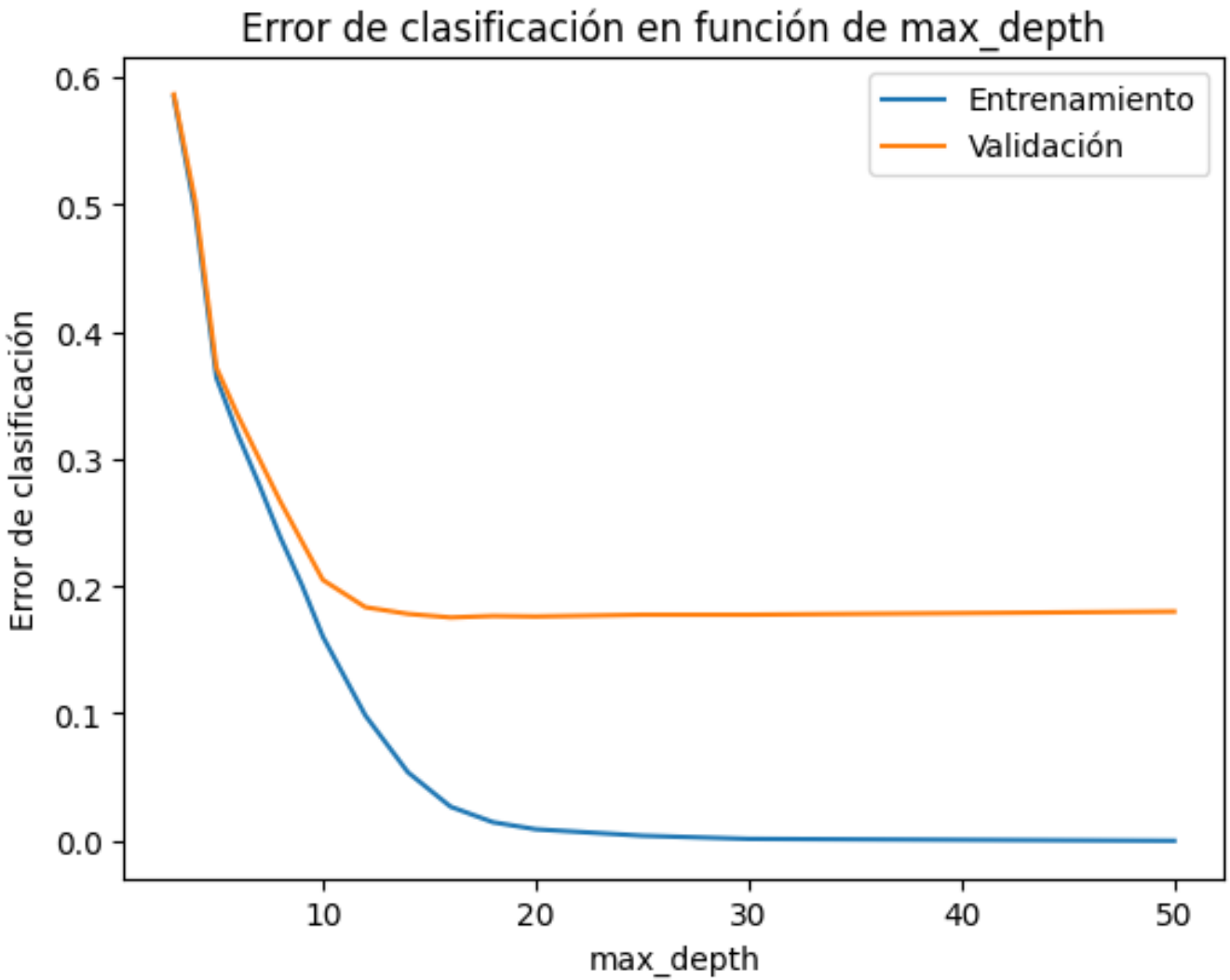
1 Repaso

2 Práctico 5

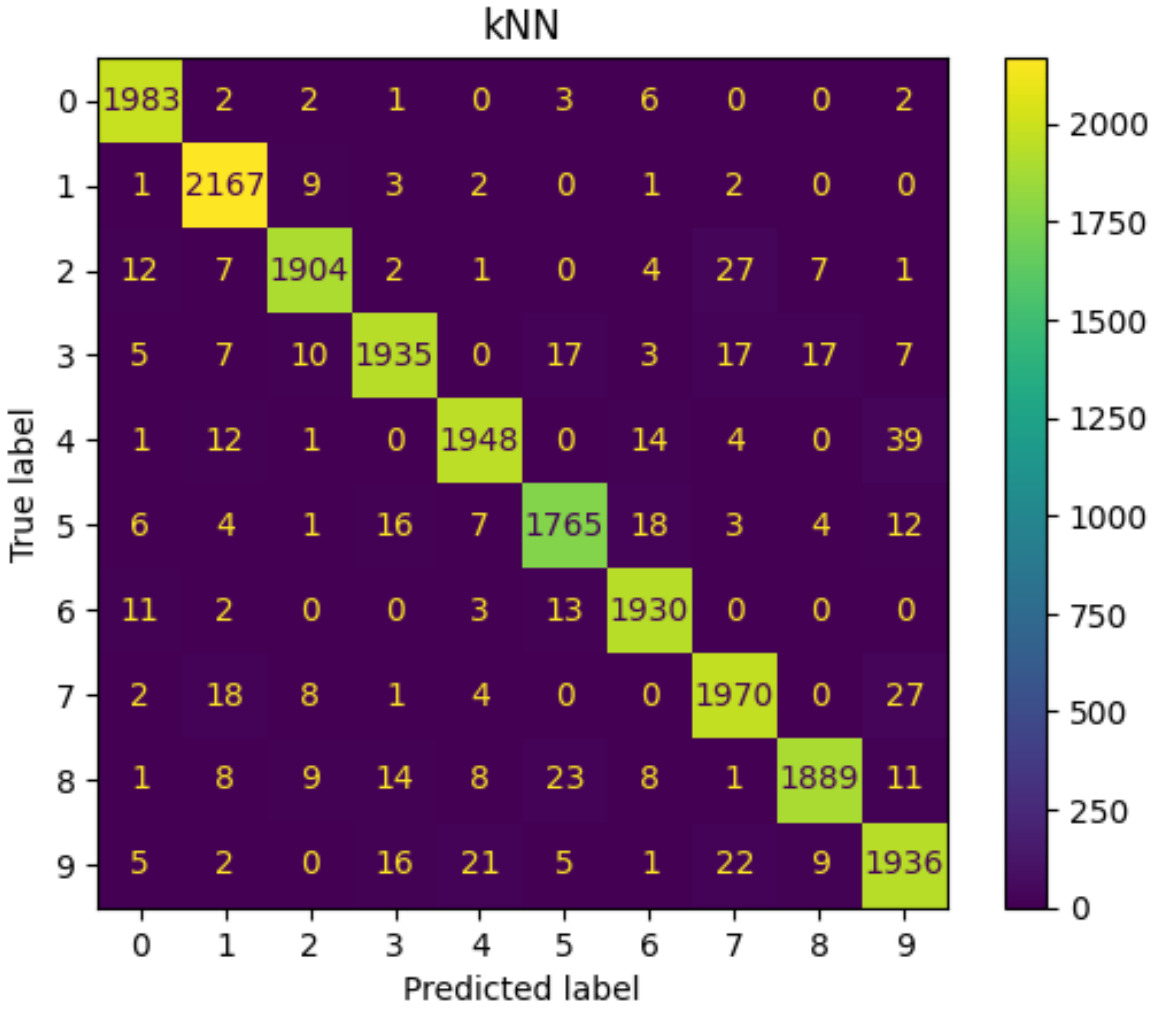
Práctico 5


Error de clasificación en función de K
Error de clasificación
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
2
4
6
8
10
12
14
16
K
Entrenamiento
Validación

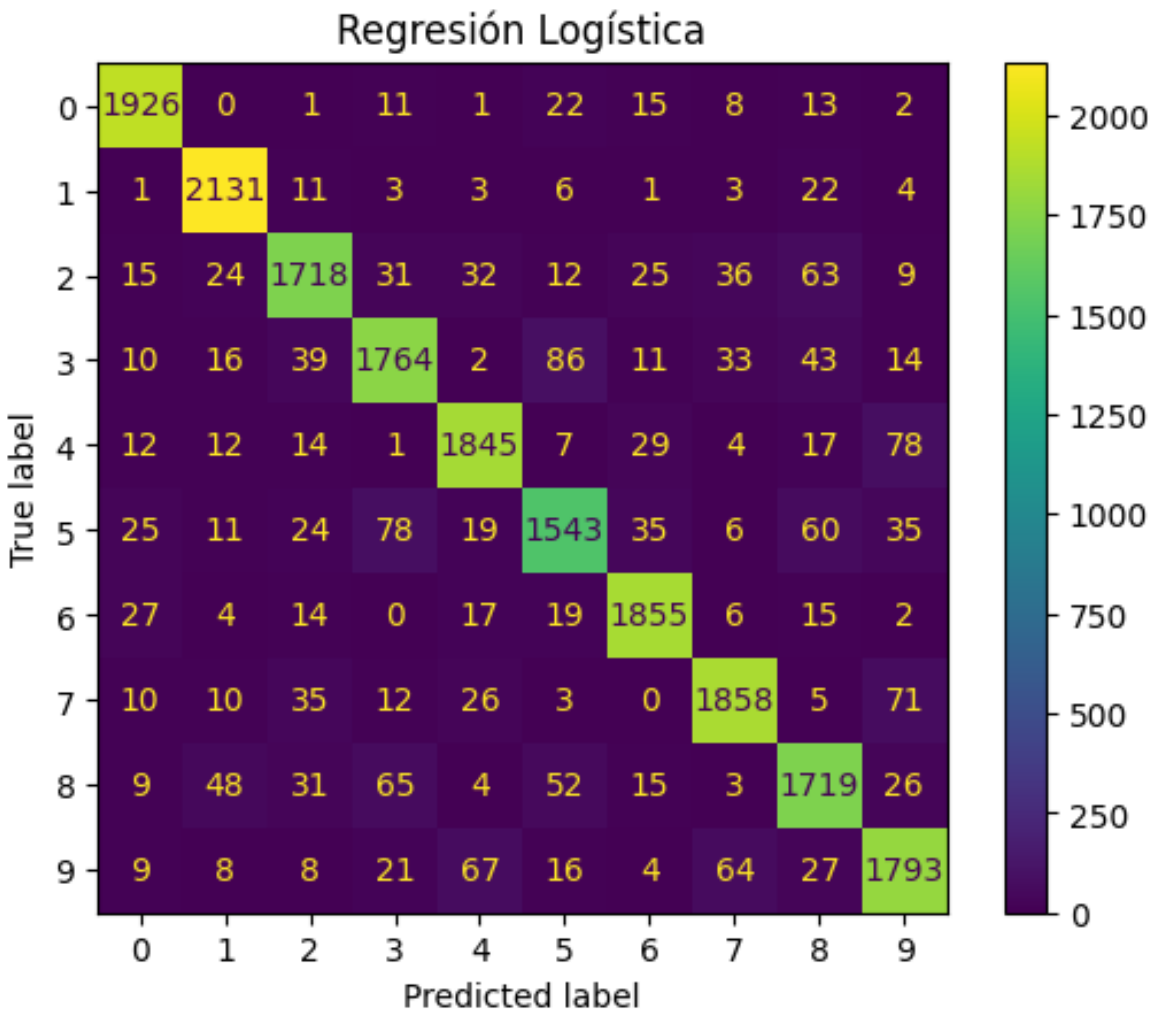
Práctico 5


Error de clasificación en función de max_depth
Entrenamiento
Validación
Error de clasificación
0.0
0.1
0.2
0.3
0.4
0.5
0.6
10
20
30
40
50
max_depth

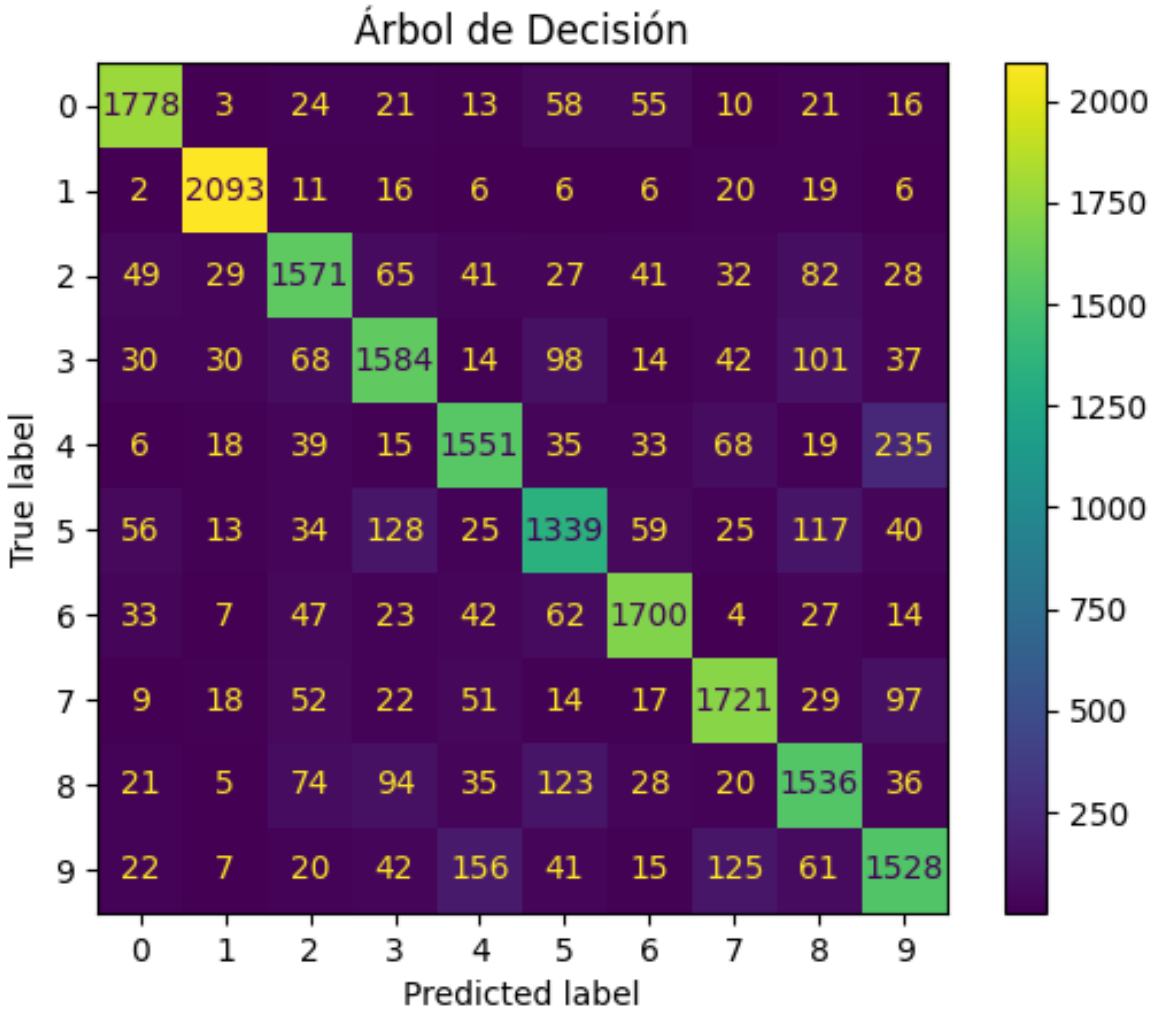
Práctico 5


kNN
True label
Predicted label
0 1983 2 2 1 0 3 6 0 0 2
1 1 2167 9 3 2 0 1 2 0 0
2 12 7 1904 2 1 0 4 27 7 1
3 5 7 10 1935 0 17 3 17 17 7
4 1 12 1 0 1948 0 14 4 0 39
5 6 4 1 16 7 1765 18 3 4 12
6 11 2 0 0 3 13 1930 0 0 0
7 2 18 8 1 4 0 0 1970 0 27
8 1 8 9 14 8 23 8 1 1889 11
9 5 2 0 16 21 5 1 22 9 1936
0 1 2 3 4 5 6 7 8 9
2000
1750
1500
1250
1000
750
500
250
0

Práctico 5


Regresión Logística
True label
Predicted label
0 1926 0 1 11 1 22 15 8 13 2
1 1 2131 11 3 3 6 1 3 22 4
2 15 24 1718 31 32 12 25 36 63 9
3 10 16 39 1764 2 86 11 33 43 14
4 12 12 14 1 1845 7 29 4 17 78
5 25 11 24 78 19 1543 35 6 60 35
6 27 4 14 0 17 19 1855 6 15 2
7 10 10 35 12 26 3 0 1858 5 71
8 9 48 31 65 4 52 15 3 1719 26
9 9 8 8 21 67 16 4 64 27 1793
0 1 2 3 4 5 6 7 8 9
2000 1750 1500 1250 1000 750 500 250 0

Práctico 5


Árbol de Decisión
True label
Predicted label
0 1778 3 24 21 13 58 55 10 21 16
1 2 2093 11 16 6 6 6 20 19 6
2 49 29 1571 65 41 27 41 32 82 28
3 30 30 68 1584 14 98 14 42 101 37
4 6 18 39 15 1551 35 33 68 19 235
5 56 13 34 128 25 1339 59 25 117 40
6 33 7 47 23 42 62 1700 4 27 14
7 9 18 52 22 51 14 17 1721 29 97
8 21 5 74 94 35 123 28 20 1536 36
9 22 7 20 42 156 41 15 125 61 1528
0 1 2 3 4 5 6 7 8 9
2000
1750
1500
1250
1000
750
500
250

Referencias

[1] Gareth James et al. *An Introduction to Statistical Learning*. Springer International Publishing, 2023. DOI: 10.1007/978-3-031-38747-0. URL: https://doi.org/10.1007/978-3-031-38747-0.

[2] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN: 9780262018029 0262018020.

[3] Sebastian Nusser. "Robust Learning in Safety-Related Domains : machine learning methods for solving safety-related application problems". Tesis doct. Jul. de 2009.

[4] Simon Rogers y Mark Girolami. *A First Course in Machine Learning*. Chapman y Hall/CRC, oct. de 2016. DOI: 10.1201/9781315382159. URL: https://doi.org/10.1201/9781315382159.