Universidad del País Vasco Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA
Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

INGENIERÍA COMPUTACIONAL Y SISTEMAS INTELIGENTES

EXPLORACIÓN Y ANÁLISIS DE DATOS

Entrega 2

Estudiante: Eneko Perez **Estudiante:** Alan García

Curso: 2024-2025 **Fecha:** 7 de octubre de 2024

Índice

1. Coordenadas principales

A la hora de elegir un espacio de representación reducido para un conjunto de datos es esencial determinar el número de coordenadas de este nuevo espacio. Es por ello que se ha realizado un análisis de los valores propios de la matriz de datos proporcionada en un *scree graph*.

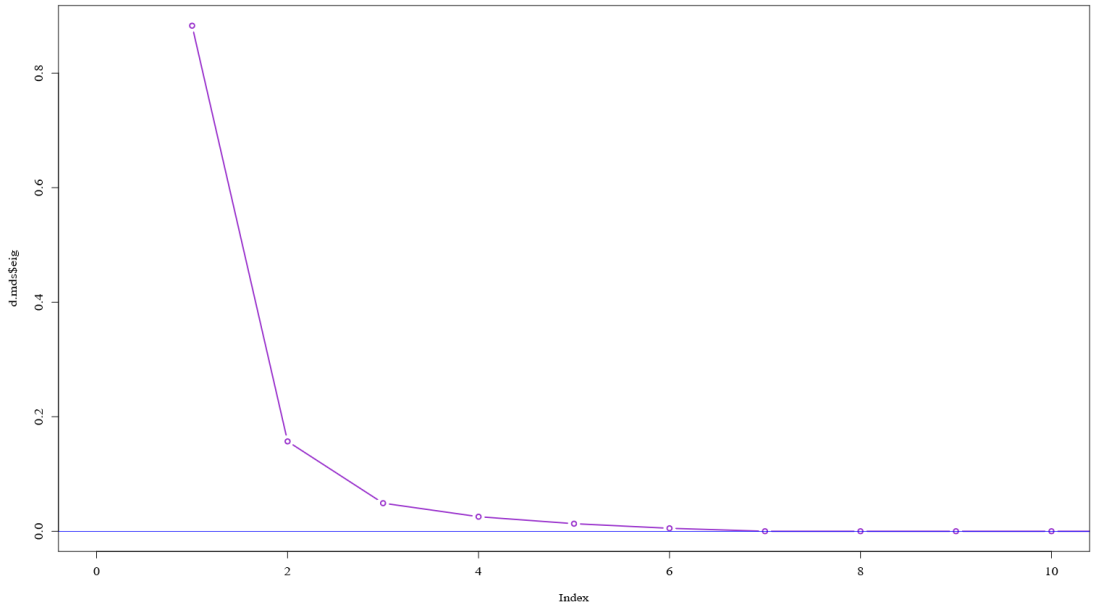

Figura 1: Scree Graph para PCA

En este gráfico sólo se muestran 10 de las 211 componentes, pero se puede apreciar cómo la mayor parte de la información está representada por las 6 primeras variables, llegando a una variabilidad acumulada del 100 %. Sin embargo, se ha decidido contar con un espacio formado por 2 variables con el fin de facilitar la representación y la obtención de conclusiones. Además, y como se puede observar en la tabla 1, con las 2 variables más significativas se mantiene un $91,81\,\%$ de la variabilidad de los datos, lo cual se considera suficiente para el desarrollo de esta práctica.

q	GOF
2	0.918088
3	0.9613737
4	0.9837884
5	0.9954371
6	1

Tabla 1: Goodness Of Fitness para distintas dimensiones

Con el espacio reducido a 2 coordenadas, se consigue la siguiente representación de los datos originales clasificados.

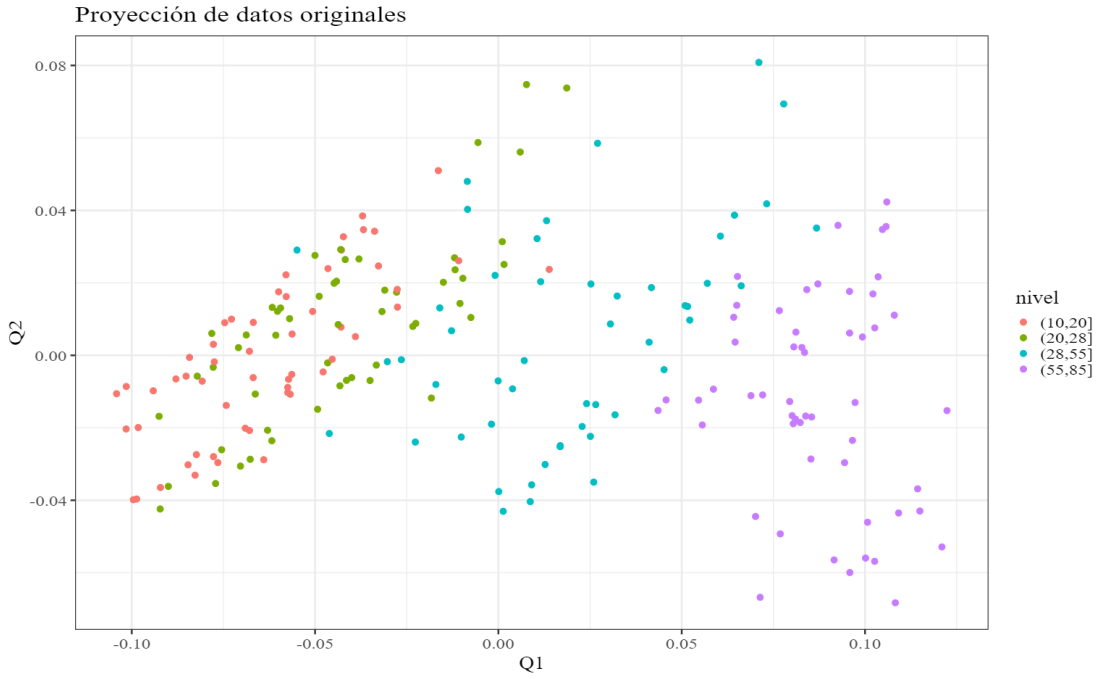

Figura 2: Proyección de los datos originales

En el eje horizontal se recoge la primera componente más significativa $Q1$. Atendiendo a la naturaleza de los datos, se puede interpretar como la cantidad de materia seca presente en las muestras. A mayores valores de $Q1$, más niveles de materia seca presentan las muestras. Sin embargo, esta relación no parece ser lineal, ya que con valores pequeños de $Q1$ las muestras se aglomeran y es complicado distinguir a qué nivel pertenece cada una de ellas, pero conforme aumenta $Q1$ resulta más sencillo diferenciar los niveles.

A continuación, se nos presenta un subconjunto de datos que no está clasificado en función del nivel de materia seca presente. Para intentar contextualizar estos datos, se han proyectado en el espacio reducido de los datos originales.

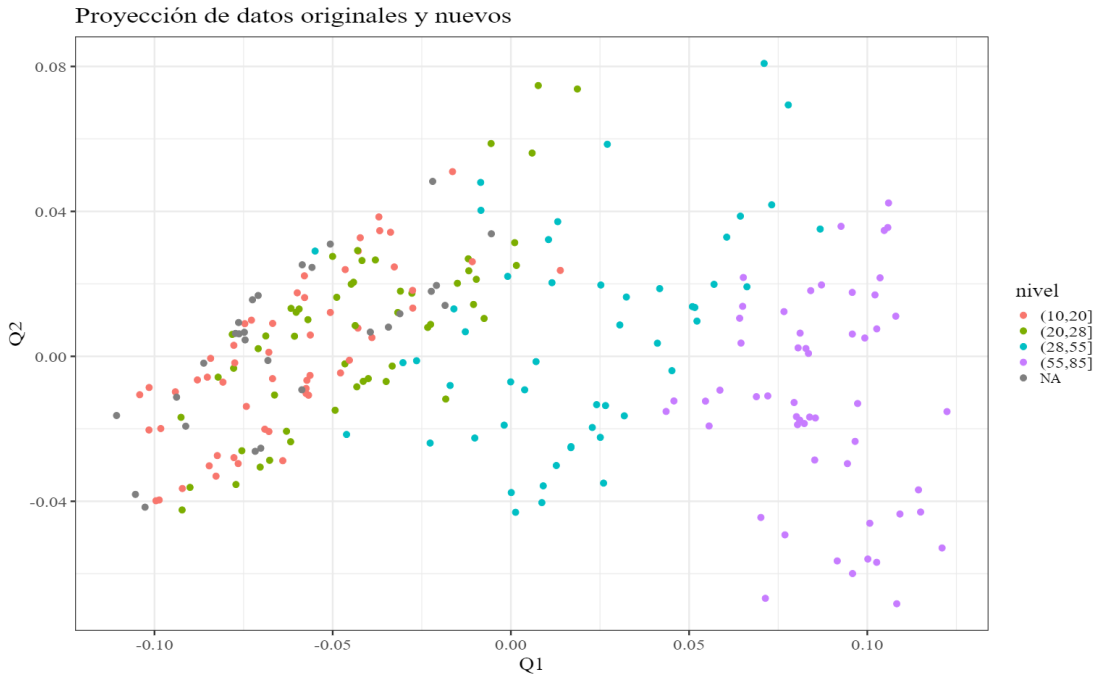

Figura 3: Proyección de los datos sin clasificar

Estas proyecciones están representadas por puntos grises en el gráfico y, como se puede apreciar, parecen corresponderse con muestras de un nivel medio-bajo de materia seca.

2. Kernel-componentes principales

Para este apartado se ha conseguido una representación según kernel-componentes principales en el espacio de dimensión 2. Se ha utilizado el kernel gaussiano con los valores $\sigma_1 = 0,05$ y $\sigma_2 = 0,5$.

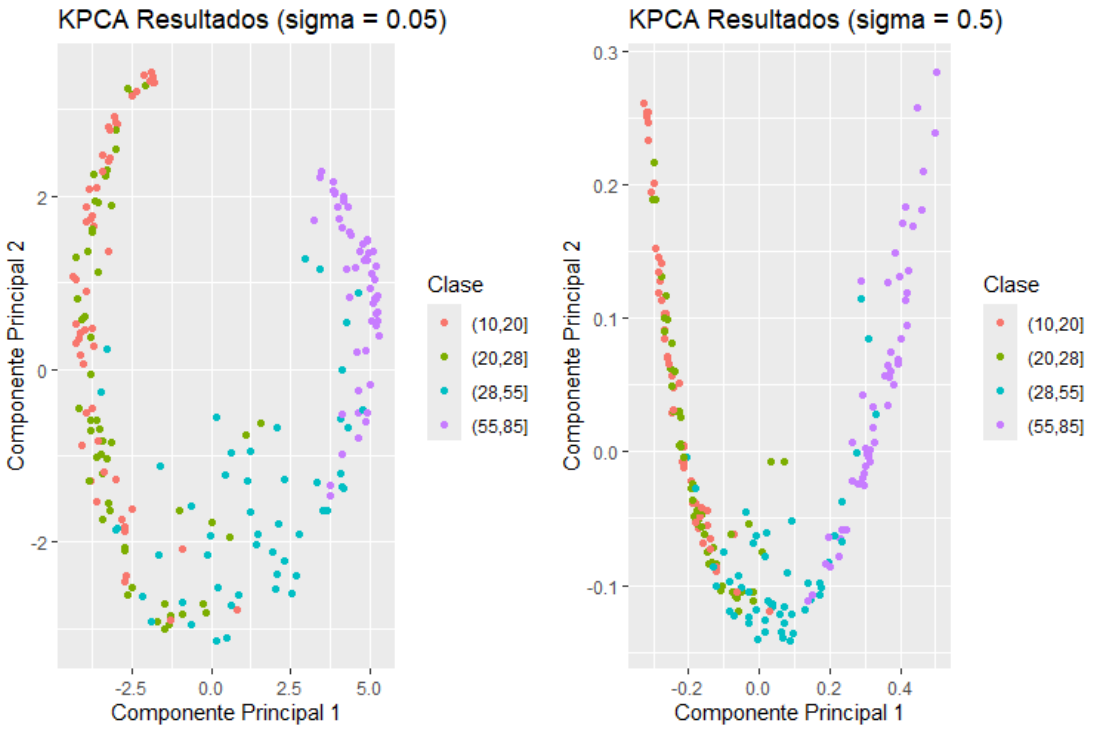

Figura 4: Proyección de los datos originales

Estos gráficos muestran los resultados de un KPCA sobre un conjunto de datos proyectados en dos dimensiones usando diferentes valores de sigma (parámetro del núcleo RBF).

En el eje horizontal se encuentra el primer componente principal (PC1), mientras que el eje vertical corresponde al segundo componente principal (PC2). Cada punto está coloreado según la clase a la que pertenece. A pesar de haber utilizado diferentes valores de σ, los gráficos no presentan variaciones significativas.

En ambos gráficos, se vuelve a observar que el PC1 está directamente relacionado con la cantidad de materia seca. Cuando el PC1 tiene valores más negativos, la cantidad de materia seca tiende a ser menor; en cambio, valores más positivos en el PC1 están asociados con una mayor cantidad de materia seca.

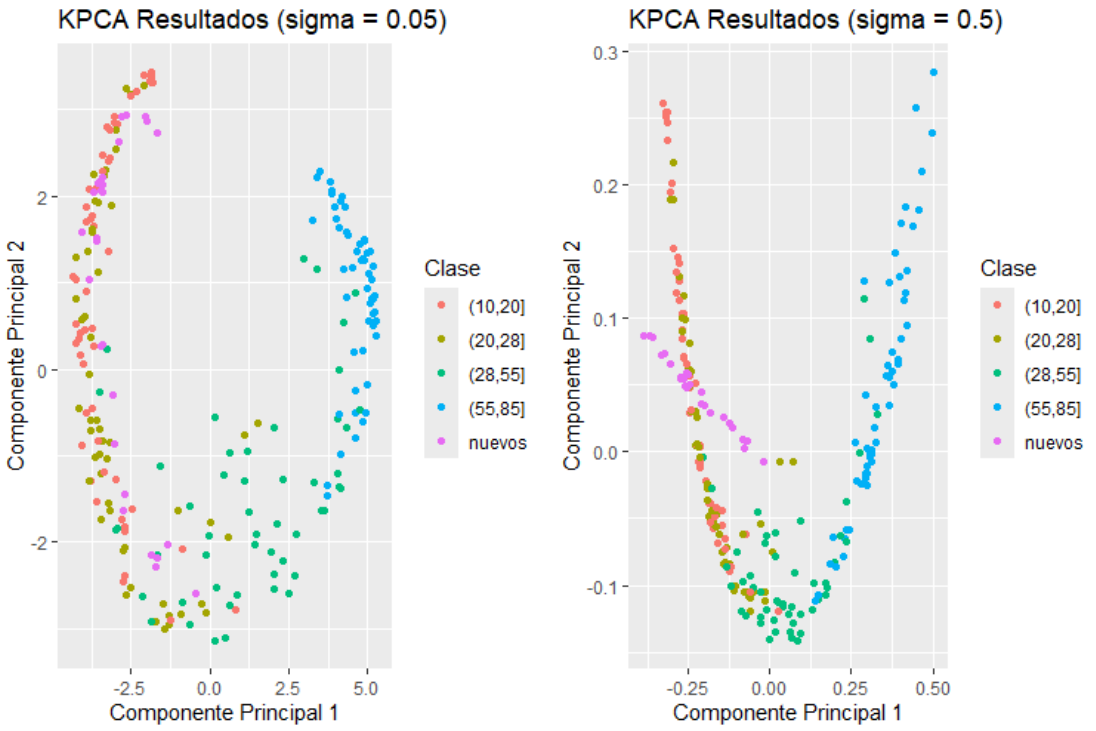

Figura 5: Proyección de los datos nuevos

Finalmente, en este gráfico se puede observar como las nuevas proyecciones están relacionadas con la clase (10, 20] y (20, 28]. Lo que sugiere que los nuevos datos tienen un índice de materia seca bajo.

3. Medidas de bondad

Para evaluar la calidad de la proyección, se han utilizado las siguientes medidas:

- Medida global: Esta medida indica la correlación entre las distancias originales y las distancias en el espacio de dimensión reducida del nuevo individuo en relación al resto.
- Medida local: Esta medida indica el porcentaje de coinciden entre los $K = 5$ vecinos mas cercanos a cada punto, considerando las distancias originales y las distancias en el espacio de dimensión reducida del nuevo individuo.

	PCA	KPCA $\sigma_1 = 0,05$	KPCA $\sigma_2 = 0,5$
Medida Global	0.357	0.062	0.357
Medida Local	0.707	0.278	0.135

Tabla 2: Medidas de Bondad

4. Conclusiones

Como ya se ha mencionado, las primeras coordenadas más significativas parecen estar estrechamente relacionadas con la cantidad de materia seca presente en las muestras. Además, las proyecciones de los datos sin clasificar al espacio reducido parecen corresponderse con muestras de un nivel medio-bajo de materia seca, manteniendo una consistencia entre las proyecciones PCA y KPCA.

Por otro lado, atendiendo a las medidas de bondad presentadas, se pueden extraer las siguientes conclusiones:

- Las correlaciones obtenidas por las bonanzas globales sugieren que existe una relación moderada entre las distancias originales y las distancias en el espacio de dimensión reducida. Esto implica que, aunque la proyección mantiene cierta estructura de los datos, hay un margen significativo de variabilidad que no se captura en la reducción de dimensionalidad.
- La bonanza local parece indicar que el método PCA mantiene mejor las inter-distancias después de realizar las proyecciones al nuevo espacio.

A. Anexo: Desarrollo alternativo de Kernel-PCA

En esta sección, se presenta una manera alternativa de implementar el Kernel-Componentes Principales, el código de este apartado está en el documento 'Kernel-PCA.R'. Para este enfoque, se ha utilizado la librería 'kernlab', empleando las funciones 'kpca' y 'predict'. Los resultados obtenidos difieren ligeramente de los mostrados anteriormente.

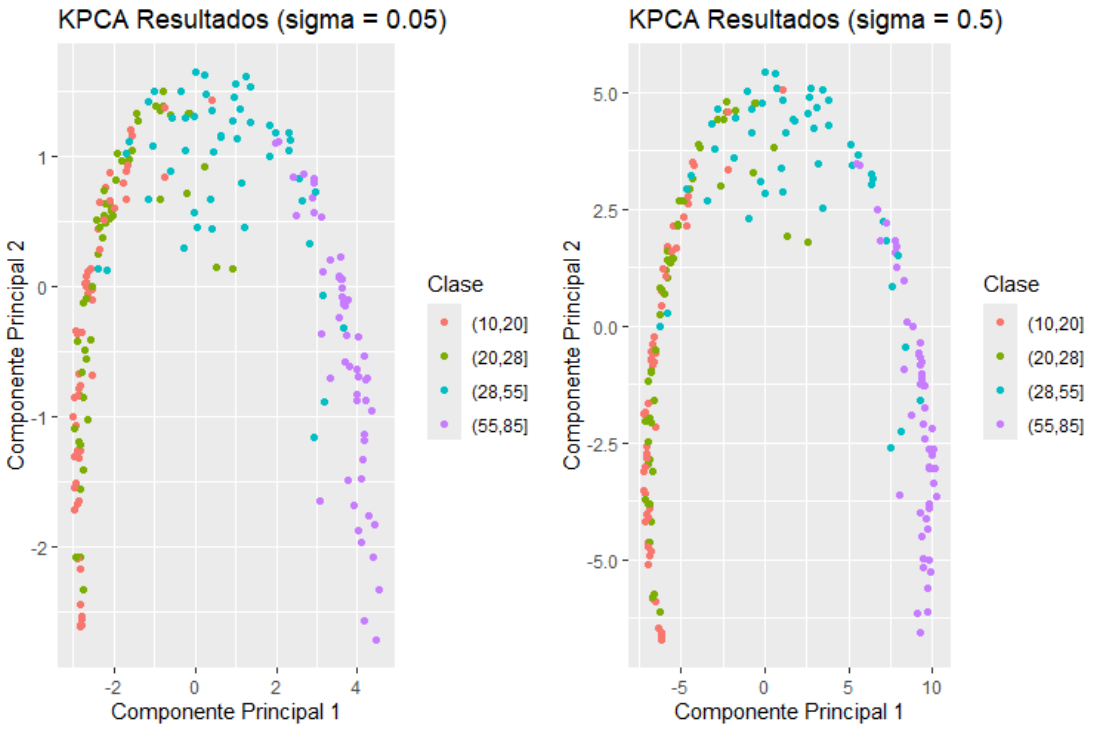

Figura 6: Proyección de los datos originales

El gráfico anterior es bastante similar al de la figura 4, aunque el vector PC2 está al invertido.

	Global $\sigma_1 = 0,05$	Global $\sigma_2 = 0,5$	Local σ_1	Local σ_2
Anterior	0.062	0.357	0.278	0.135
Anexo	0.314	0.234	0.328	0.314

Tabla 3: Medidas de Bondad

Como se puede observar en la tabla, el método alternativo proporciona índices ligeramente superiores en todos los casos.

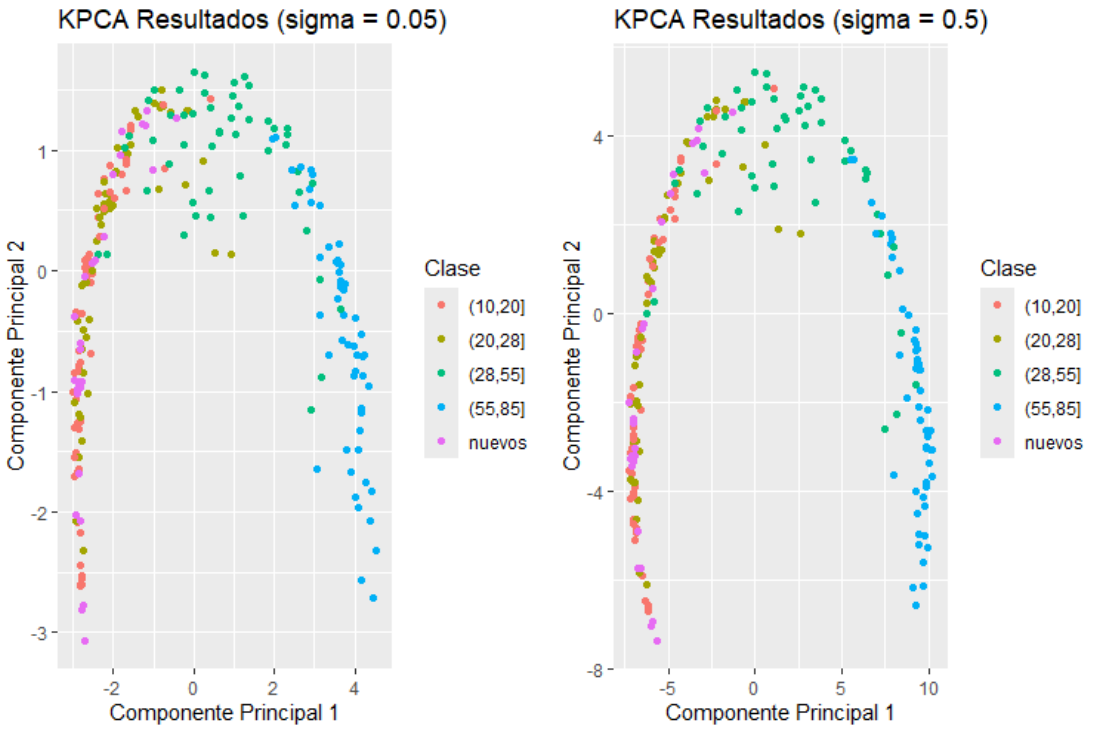

Figura 7: Proyección de los datos nuevos

Por último, la diferencia más notable entre las dos formas en las que se ha desarrollado esta tarea es este último gráfico. Como se puede observar las nuevas proyecciones están mejor representadas que en la figura 5.