Clasificación física de estrellas usando PCA y clustering: Caso práctico basado en datos reales de Gaia DR3 (ESA)

Enrique Garrido

Universidad Alfonso X el Sabio, Grado en Física

ABSTRACT

Gaia DR3 ofrece un catálogo amplio y consistente de datos estelares con múltiples parámetros físicos y observacionales, lo que hace especialmente adecuado el uso de técnicas de aprendizaje automático no supervisado para el estudio de poblaciones estelares. En este trabajo se analizan 200 000 estrellas de Gaia DR3 utilizando análisis de componentes principales (PCA) y algoritmos de clustering, sin imponer clasificaciones previas. A partir de un conjunto reducido de variables físicas fundamentales, el PCA permite identificar las direcciones dominantes del espacio de parámetros y reducir la dimensionalidad del problema, concentrando la mayor parte de la varianza en las primeras componentes. Sobre el espacio PCA se aplica clustering no supervisado, principalmente mediante *k-means*, identificando grupos con propiedades físicas diferenciadas. La proyección de los clusters en el diagrama de Hertzsprung–Russell muestra una correspondencia coherente con poblaciones estelares clásicas. Los resultados ponen de manifiesto el potencial de estas técnicas para el análisis exploratorio de grandes catálogos astronómicos.

1 Introduction

El estudio de las poblaciones estelares es un elemento central de la astrofísica, ya que permite comprender la estructura, formación y evolución de la Galaxia. Tradicionalmente, la clasificación de estrellas se ha basado en esquemas discretos, como las clases espectrales o las regiones del diagrama de Hertzsprung–Russell, definidos a partir de propiedades físicas y observacionales bien establecidas. Sin embargo, la disponibilidad actual de grandes catálogos astrofísicos plantea nuevos retos y oportunidades para el análisis de estas poblaciones.

La misión *Gaia* de la Agencia Espacial Europea ha proporcionado, en su tercera liberación de datos (Gaia DR3), información astrométrica, fotométrica y física de millones de estrellas con una precisión sin precedentes. Este volumen de datos, junto con la complejidad y continuidad inherente de las propiedades estelares, hace especialmente adecuado el uso de técnicas de aprendizaje automático no supervisado para explorar la estructura del espacio de parámetros sin imponer clasificaciones previas.

Entre estas técnicas, el análisis de componentes principales (PCA) permite identificar combinaciones lineales de variables físicas que concentran la mayor parte de la varianza del sistema, facilitando una representación de baja dimensión del conjunto de datos. Por su parte, los algoritmos de clustering no supervisado permiten agrupar objetos según su similitud estadística en dicho espacio reducido, ofreciendo una segmentación emergente basada únicamente en los datos.

Desde un punto de vista físico, este enfoque resulta especialmente interesante en el contexto de la clasificación estelar, ya que las propiedades fundamentales de las estrellas —como la temperatura efectiva, la gravedad superficial, la luminosidad o la composición química— varían de forma continua a lo largo de la evolución estelar. Por tanto, cualquier clasificación obtenida de forma no supervisada debe interpretarse como una aproximación estadística a estas transiciones físicas, más que como una división estrictamente discreta.

El objetivo de este trabajo es aplicar técnicas de aprendizaje automático no supervisado a un subconjunto representativo de datos de Gaia DR3, utilizando PCA y distintos métodos de clustering para explorar la estructura del espacio de parámetros estelares. Se busca evaluar hasta qué punto las principales secuencias y poblaciones estelares emergen de manera natural a partir de los datos, sin introducir etiquetas evolutivas ni clasificaciones preconcebidas, y analizar las diferencias entre métodos de clustering basados en particiones y en densidad.

2 Datos y metodología

2.1 Conjunto de datos

En este trabajo se emplea un subconjunto de datos procedente de *Gaia DR3*, centrado en estrellas para las que se dispone de parámetros astrofísicos fundamentales estimados de forma homogénea. El catálogo original contiene aproximadamente

6.3×10^5 fuentes y un total de 50 variables, incluyendo información astrométrica, fotométrica y parámetros físicos derivados.

Con el objetivo de aplicar técnicas de aprendizaje automático no supervisado de forma controlada y físicamente interpretable, se seleccionó un conjunto reducido de variables directamente relacionadas con la estructura y el estado físico de las estrellas. Las variables utilizadas en el análisis son:

- Temperatura efectiva ($T_{\rm eff}$)
- Gravedad superficial ($\log g$)
- Metalicidad ([Fe/H])
- Magnitud absoluta en banda G (M_G)
- Color fotométrico $BP - RP$
- Luminosidad estelar estimada (Lum-Flame)
- Radio estelar estimado (Rad-Flame)

El color $BP - RP$ se construyó explícitamente a partir de las magnitudes BP y RP y se priorizaron variables que caracterizan propiedades fundamentales de las estrellas, evitando el uso de etiquetas discretas o parámetros explícitamente supervisados.

La temperatura efectiva ($T_{\rm eff}$) y la gravedad superficial ($\log g$) permiten caracterizar el estado termodinámico y la estructura interna de las estrellas, mientras que la metalicidad ([Fe/H]) aporta información sobre su composición química. La magnitud absoluta en banda G (M_G) y el color fotométrico $BP - RP$ describen la posición de las estrellas en el diagrama de Hertzsprung–Russell y están directamente relacionados con su luminosidad y temperatura.

Adicionalmente, se incluyeron el radio y la luminosidad estelares estimados (Rad-Flame y Lum-Flame) con el fin de incorporar información sobre el tamaño y la energía emitida por las estrellas, complementando la descripción puramente fotométrica. Aunque estos parámetros dependen de modelos estelares, su inclusión permite enriquecer el espacio de características sin introducir etiquetas categóricas ni información evolutiva explícita.

No se utilizaron variables cinemáticas, probabilísticas ni parámetros evolutivos discretos en ninguna fase del entrenamiento, con el fin de evitar introducir información supervisada o dependiente de modelos complejos.

2.2 Preprocesado y selección de la muestra

Antes del análisis se aplicaron varios criterios de calidad y consistencia física:

- Paralaje positiva (Plx > 0)
- Temperatura efectiva positiva ($T_{\rm eff} > 0$)
- Gravedad superficial no negativa ($\log g \geq 0$)
- Parámetro de calidad astrométrica RUWE ≤ 1.4

Tras aplicar estos filtros y eliminar valores faltantes en las variables seleccionadas, se obtuvo una muestra limpia y homogénea. Con el fin de reducir el coste computacional manteniendo la estructura estadística del conjunto, se extrajo una submuestra aleatoria de **200 000 estrellas**, fijando una semilla para garantizar la reproducibilidad.

El análisis exploratorio inicial incluyó la inspección de las distribuciones marginales de $T_{\rm eff}$, $\log g$, $BP - RP$ y [Fe/H], así como un estudio sistemático de la fracción de valores ausentes en el catálogo original. Este paso permitió confirmar que las variables seleccionadas presentan una cobertura adecuada y distribuciones coherentes con poblaciones estelares reales.

2.3 Estandarización de variables

Dado que las variables consideradas presentan escalas físicas y unidades muy diferentes, todas ellas fueron estandarizadas antes del análisis mediante una normalización tipo *z-score*, restando la media y dividiendo por la desviación típica de cada variable. Este paso es esencial tanto para el análisis de componentes principales como para los algoritmos de clustering basados en distancias, ya que evita que variables con mayor dispersión dominen artificialmente el resultado.

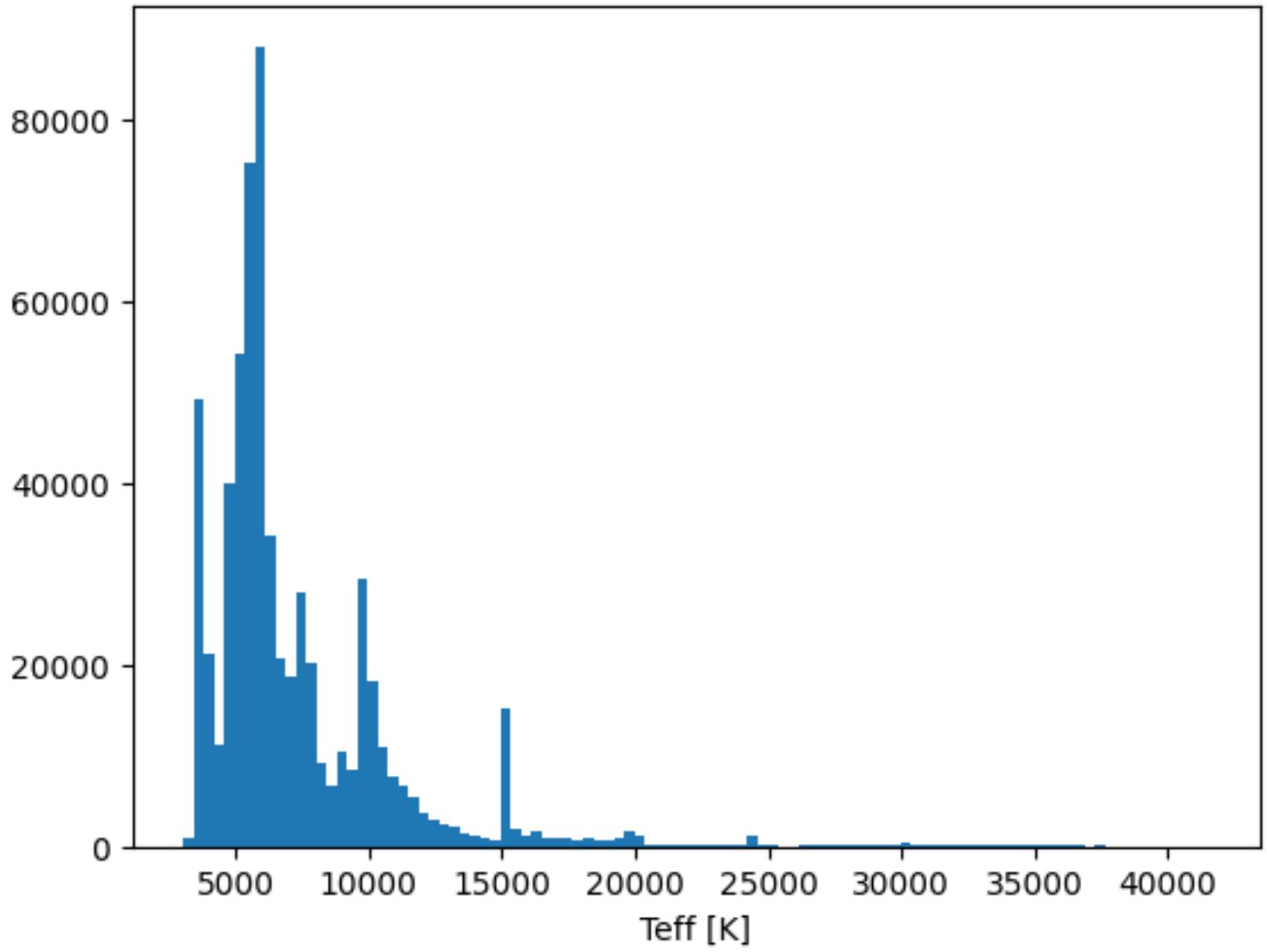

Figure 1. Distribución marginal de la temperatura efectiva, gravedad superficial, color fotométrico y metalicidad.

2.4 Análisis de Componentes Principales

El análisis de componentes principales (PCA) se utilizó como técnica de reducción de dimensionalidad y exploración del espacio de parámetros. El PCA se aplicó sobre las siete variables estandarizadas, sin imponer inicialmente un número fijo de componentes.

La fracción de varianza explicada por cada componente principal se evaluó mediante el espectro de autovalores y la varianza acumulada. Este análisis mostró que las primeras componentes concentran la mayor parte de la información del sistema, permitiendo describir el conjunto de datos en un espacio de dimensión reducida sin una pérdida significativa de información.

Además de la varianza explicada, se analizaron las cargas de cada componente principal con el objetivo de identificar qué variables contribuyen de forma dominante a cada eje del espacio PCA. Para ello se emplearon tanto la matriz completa de cargas como representaciones gráficas ordenadas por valor absoluto, así como biplots en el plano PC1–PC2.

Las proyecciones de los datos en el espacio PCA se visualizaron mediante diagramas de dispersión y mapas de densidad (*hexbin*), y se validaron coloreando los puntos con variables físicas no utilizadas directamente en la definición de los ejes, como $\log g$, $T_{\rm eff}$ y $[\mathrm{Fe/H}]$. Este procedimiento permitió evaluar la coherencia física de las componentes principales obtenidas.

2.5 Clustering no supervisado

Una vez reducido el espacio de parámetros mediante PCA, se aplicaron técnicas de clustering no supervisado con el objetivo de identificar agrupaciones naturales en los datos. El clustering se realizó en el subespacio definido por las primeras cuatro componentes principales, que concentran la mayor parte de la varianza total.

Se utilizó el algoritmo *k-means*, ampliamente empleado en problemas de clasificación no supervisada por su simplicidad y estabilidad. Para determinar un número adecuado de clusters se evaluaron distintos valores de k mediante métricas internas

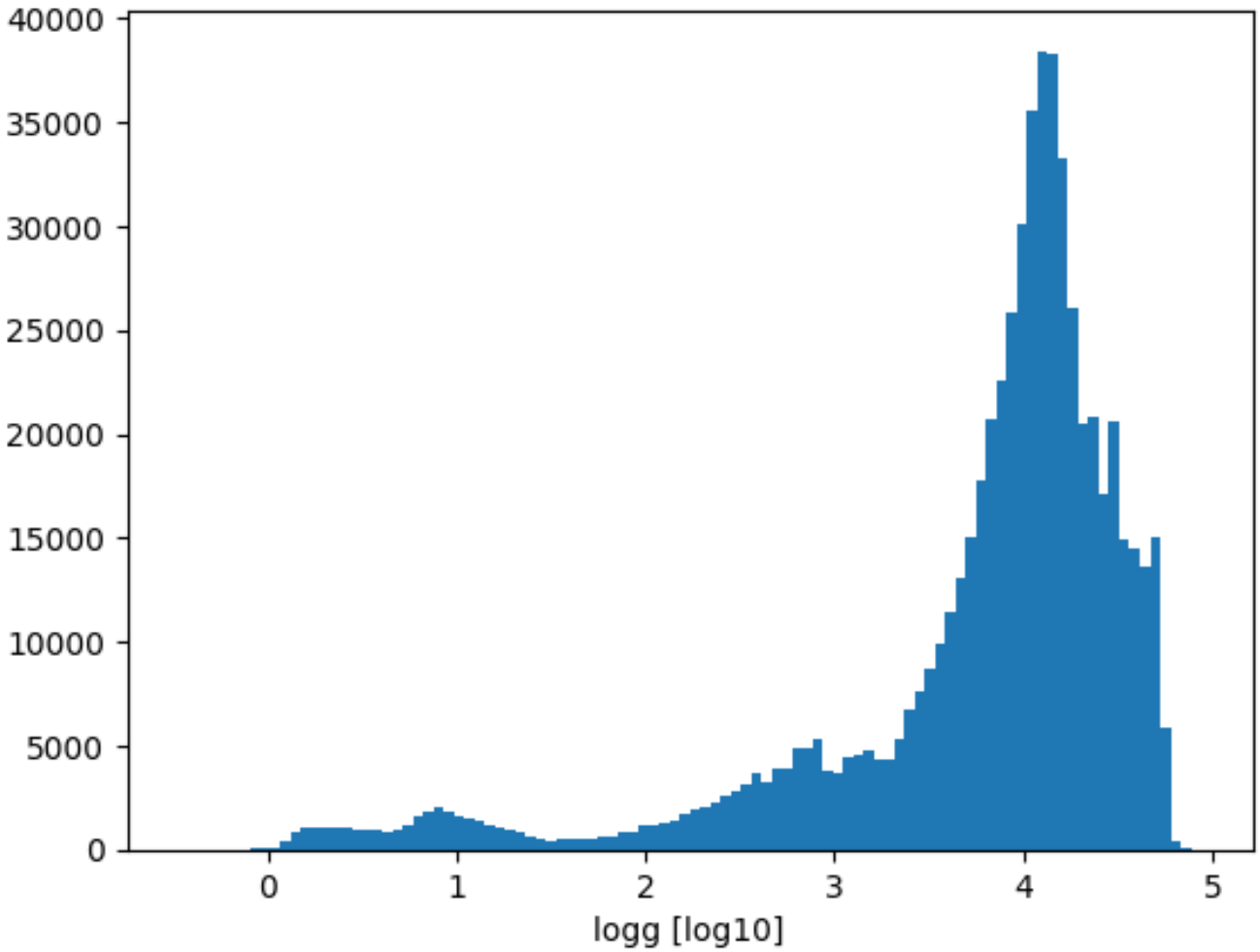

Figure 2. Distribución marginal de la temperatura efectiva, gravedad superficial, color fotométrico y metalicidad.

de validación, concretamente el coeficiente *silhouette* y el índice de Davies–Bouldin. Estas métricas se calcularon de forma sistemática para varios valores de k, permitiendo comparar la compacidad y separación de los grupos obtenidos.

Una vez fijado el número de clusters, se analizaron las proyecciones de los resultados tanto en el espacio PCA como en el diagrama de Hertzsprung–Russell clásico (M_G frente a $BP - RP$). Asimismo, se calcularon estadísticas descriptivas (media y desviación típica) de las variables físicas originales para cada cluster, con el fin de facilitar su interpretación posterior.

Adicionalmente, se exploró un método de clustering basado en densidad (DBSCAN) con el objetivo de comparar su comportamiento con el algoritmo *k-means*. Para ello, se evaluó un rango de valores de los parámetros *eps* y *min_samples*, analizando el número de clusters detectados y la fracción de puntos clasificados como ruido. Esta comparación permite evaluar la robustez de los resultados frente a la elección del algoritmo de clustering.

3 Resultados

3.1 Resultados del Análisis de Componentes Principales

La Figura correspondiente al espectro de varianza explicada muestra que la primera componente principal (PC1) explica aproximadamente el 48.6% de la varianza total del conjunto de datos. Las dos primeras componentes acumulan alrededor del 73.4%, mientras que las cuatro primeras componentes concentran aproximadamente el 93.7% de la varianza total. A partir de la quinta componente, la contribución adicional a la varianza explicada es marginal.

La matriz de cargas del PCA revela que todas las variables originales contribuyen de forma no nula a las distintas componentes principales, con pesos relativos diferentes según el eje considerado. Las representaciones gráficas de las cargas ordenadas permiten visualizar qué variables dominan cada componente principal, mostrando patrones diferenciados entre PC1,

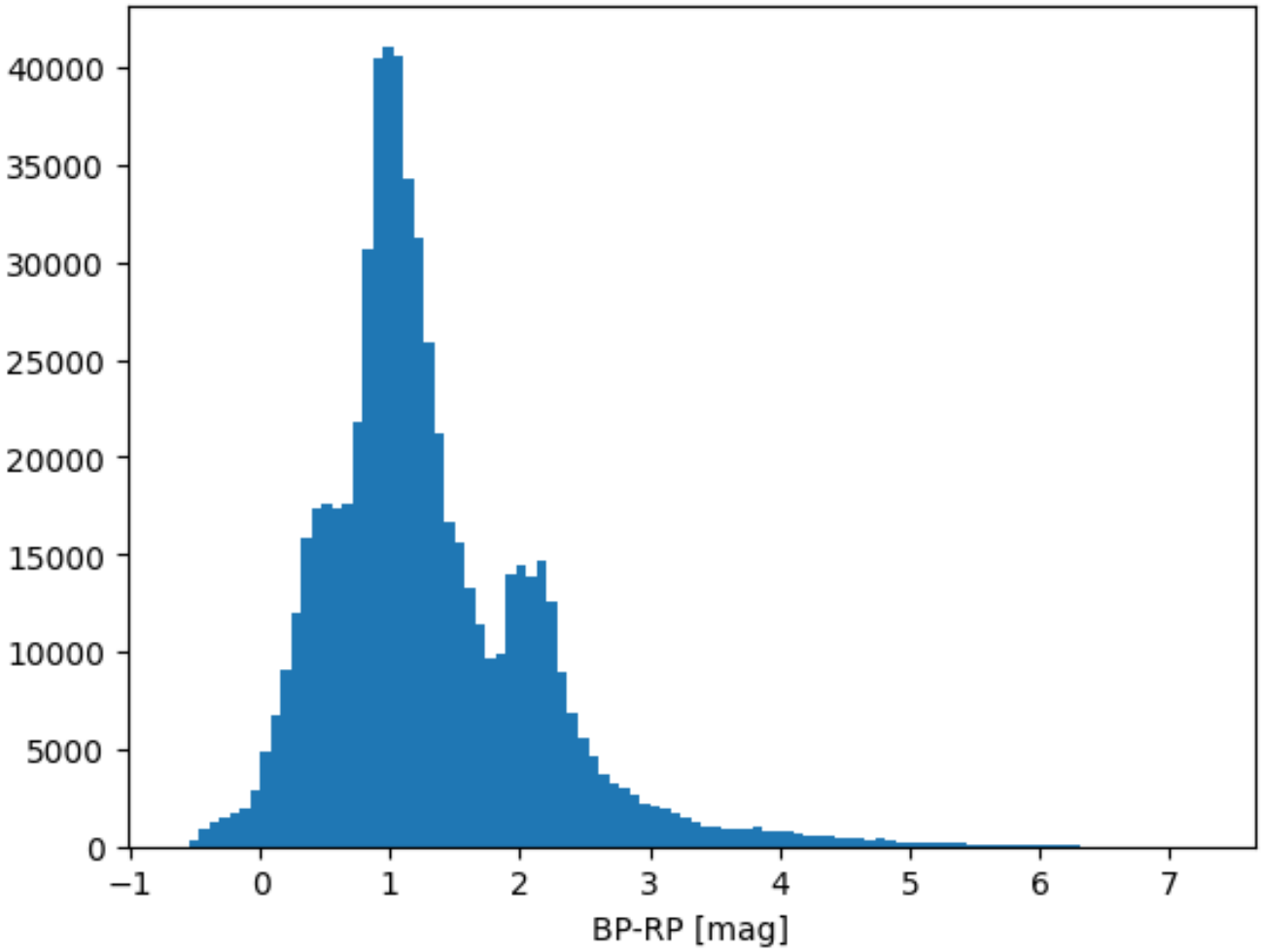

Figure 3. Distribución marginal de la temperatura efectiva, gravedad superficial, color fotométrico y metalicidad.

PC2, PC3 y PC4.

Las proyecciones de los datos en el plano PC1–PC2, mostradas mediante diagramas de dispersión y mapas de densidad (*hexbin*), presentan una distribución no uniforme, con regiones de alta densidad claramente diferenciadas de zonas más dispersas. Este patrón se mantiene, con variaciones en la morfología, en las proyecciones PC1–PC3 y PC2–PC3. Al colorear las proyecciones PC1–PC2 por diferentes parámetros físicos, como $\log g$, $T_{\rm eff}$ y $[{\rm Fe}/{\rm H}]$, se observa una variación sistemática del color a lo largo del espacio PCA. Estas transiciones son suaves y continuas, sin fronteras abruptas, y se reproducen de forma consistente en distintas proyecciones bidimensionales.

El biplot PC1–PC2 muestra la orientación de las variables originales en el espacio definido por las dos primeras componentes principales. Las flechas asociadas a cada variable presentan direcciones y longitudes diferentes, indicando contribuciones distintas de las variables al plano PC1–PC2. Este diagrama permite visualizar simultáneamente la distribución de los datos y la relación geométrica entre las variables originales y las componentes principales.

Finalmente, se analizó la distribución radial de los datos en el espacio definido por las primeras cuatro componentes principales, identificándose un pequeño subconjunto de puntos situados a grandes distancias del origen. Estos puntos representan una fracción muy reducida del total de la muestra.

3.2 Resultados del clustering no supervisado

El clustering se llevó a cabo en el subespacio definido por las primeras cuatro componentes principales. Para evaluar la dependencia de los resultados con el número de clusters, se calcularon el coeficiente *silhouette* y el índice de Davies–Bouldin para distintos valores de k.

Los valores obtenidos para ambas métricas muestran una variación sistemática con k, sin la aparición de un mínimo o

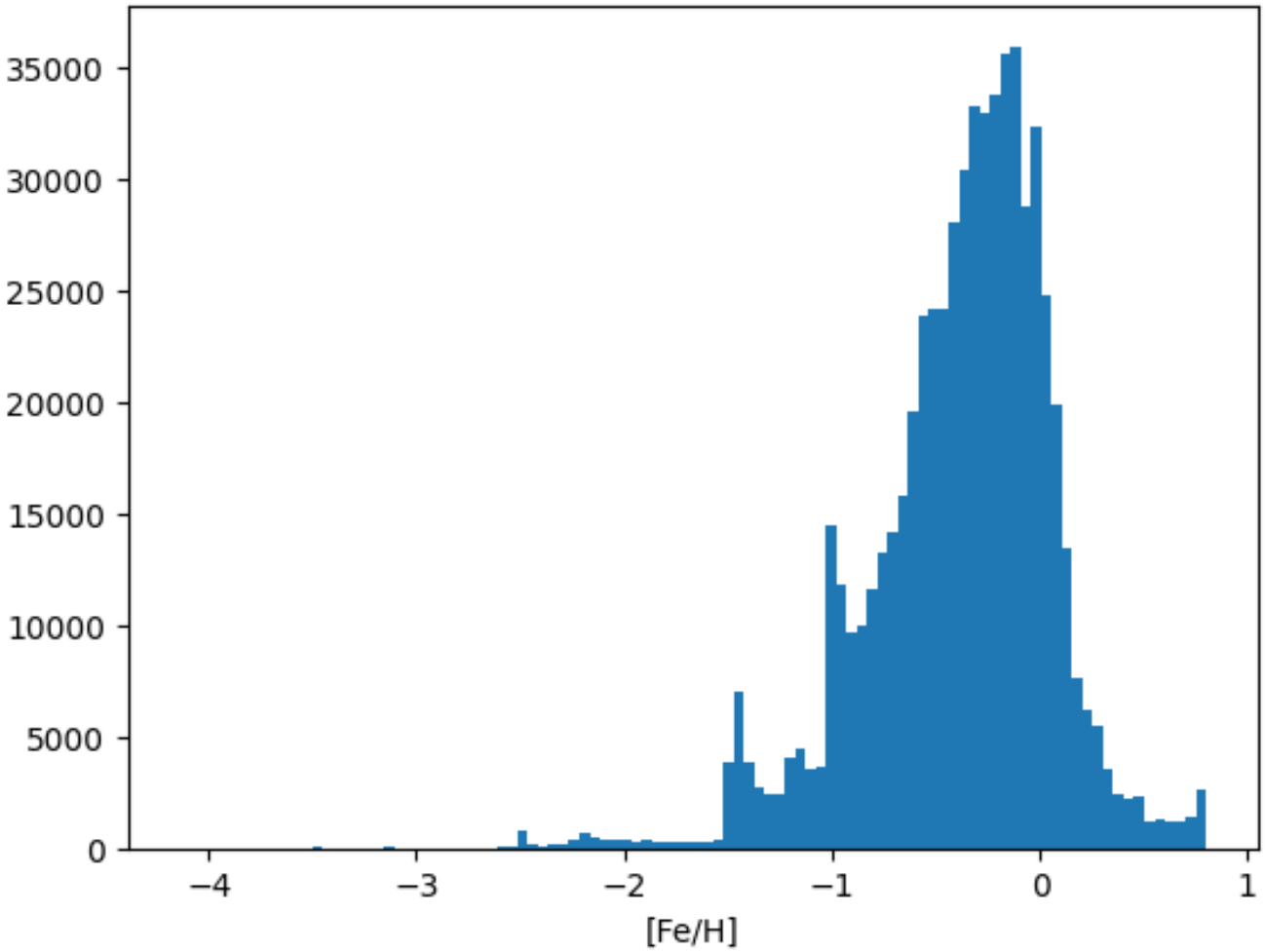

Figure 4. Distribución marginal de la temperatura efectiva, gravedad superficial, color fotométrico y metalicidad.

máximo extremadamente pronunciado. En función de estas métricas, se seleccionó un valor final de $k = 5$ para el análisis posterior.

La proyección de los clusters obtenidos en el plano PC1–PC2 muestra regiones ocupadas predominantemente por un único cluster, junto con zonas de transición donde varios clusters se solapan parcialmente. Esta estructura se mantiene al representar los clusters en otras proyecciones del espacio PCA.

La Figura correspondiente al diagrama de Hertzsprung–Russell coloreado por clusters muestra que los distintos grupos identificados ocupan regiones diferenciadas del plano M_G frente a $BP - RP$. Aunque existe solapamiento entre clusters, cada uno de ellos presenta una distribución característica en el diagrama HR.

Para cada cluster se calcularon estadísticas descriptivas de las variables físicas originales, incluyendo medias y desviaciones típicas de $T_{\rm eff}$, $\log g$, $[{\rm Fe/H}]$, M_G, radio estelar y luminosidad. Estas estadísticas muestran diferencias cuantitativas entre clusters en todas las variables consideradas, así como dispersiones internas apreciables dentro de cada grupo.

Adicionalmente, se repitió el análisis de clustering en subespacios PCA de distinta dimensionalidad, comparando los resultados obtenidos al utilizar dos, tres y cuatro componentes principales. Las métricas internas y las distribuciones de los clusters en el espacio PCA presentan variaciones entre estos casos, aunque la estructura global del espacio permanece similar.

3.2.1 Comparación con un método basado en densidad

Además del algoritmo *k-means*, se aplicó el método DBSCAN en el mismo subespacio PCA con el fin de comparar los patrones de agrupamiento obtenidos. Para DBSCAN se exploraron distintas combinaciones de los parámetros *eps* y *min_samples*, analizando el número de clusters identificados y la fracción de puntos clasificados como ruido.

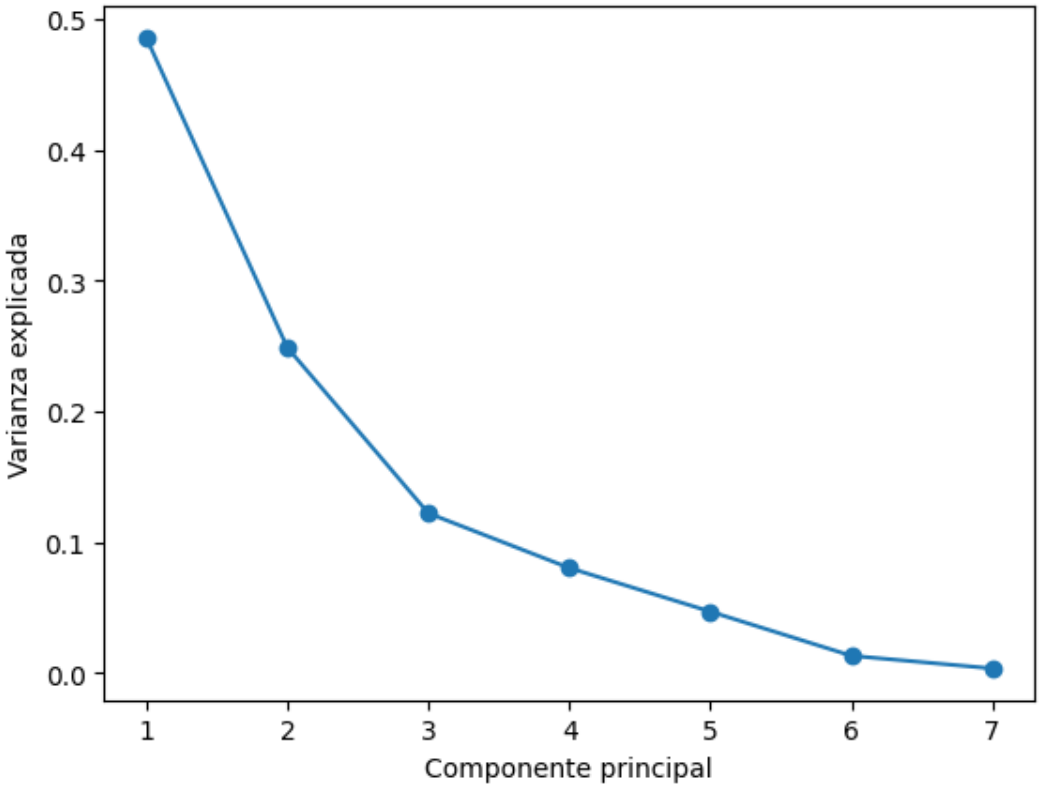

Figure 5. Fracción de varianza explicada por cada componente principal y varianza acumulada.

4 Discusión

El análisis muestra que las técnicas de aprendizaje automático no supervisado aplicadas a un conjunto reducido de variables físicas permiten recuperar de forma emergente estructuras coherentes con la clasificación estelar clásica.

4.1 Interpretación física de las componentes principales

La primera componente principal concentra una fracción significativa de la varianza total del sistema y está dominada por variables relacionadas con el tamaño y la estructura estelar, como el radio, la luminosidad y la gravedad superficial. Esta componente organiza el conjunto de datos a lo largo de una dirección que separa estrellas compactas de baja luminosidad de estrellas de gran radio y alta luminosidad, reflejando una distinción estructural fundamental.

La segunda componente principal presenta una contribución dominante de la temperatura efectiva y de las variables fotométricas, describiendo una secuencia térmica que es claramente visible en las proyecciones del espacio PCA. Esta componente está estrechamente relacionada con la distribución de las estrellas en el diagrama de Hertzsprung–Russell y reproduce una transición continua a lo largo de la secuencia principal y hacia fases evolutivas más avanzadas.

La tercera componente principal está fuertemente asociada a la metalicidad, introduciendo una dimensión adicional que discrimina poblaciones con distinta composición química. Esta información no domina la estructura global del espacio, pero contribuye a refinar la separación entre estrellas que ocupan regiones similares en términos de temperatura y estructura.

La cuarta componente principal aporta una contribución adicional menor, asociada a combinaciones más sutiles de variables físicas, y puede interpretarse como un refinamiento del estado evolutivo dentro de poblaciones ya separadas por las componentes principales anteriores.

En conjunto, las primeras cuatro componentes principales permiten describir de forma compacta la mayor parte de la información física relevante del sistema, separando progresivamente las estrellas según estructura, temperatura y composición química.

4.2 Clustering en el espacio PCA y relación con el diagrama HR

El clustering aplicado en el subespacio definido por las primeras componentes principales identifica grupos de estrellas que ocupan regiones diferenciadas tanto en el espacio PCA como en el diagrama de Hertzsprung–Russell. Aunque los límites entre

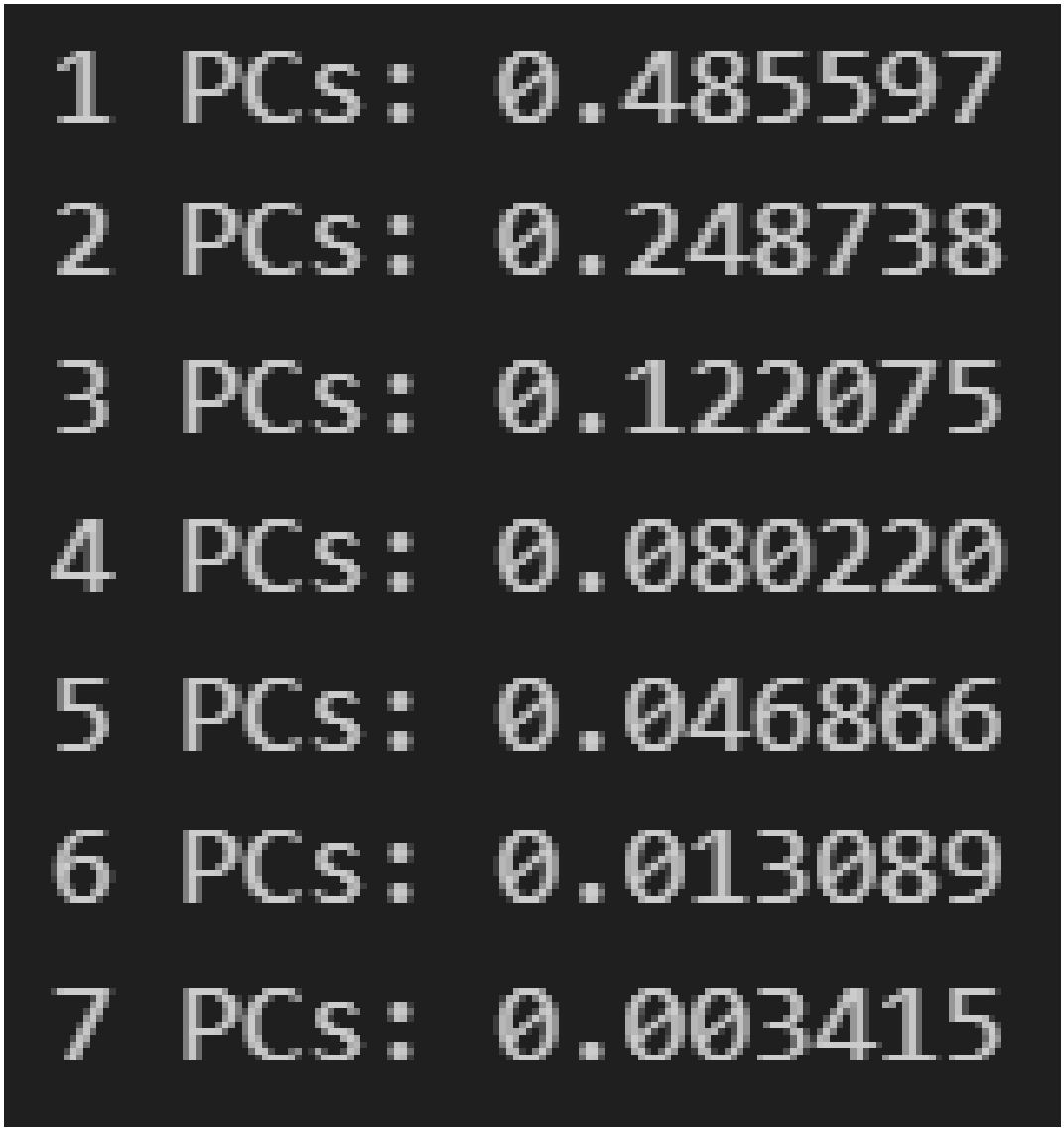

Figure 6. Fracción de varianza explicada por cada componente principal y varianza acumulada.

clusters no son abruptos, los grupos obtenidos presentan propiedades físicas medias claramente diferenciadas.

La proyección de los clusters sobre el diagrama HR muestra que estos corresponden a distintas regiones clásicas del mismo, como la secuencia principal, la rama de gigantes y poblaciones de estrellas compactas y poco luminosas. El solapamiento parcial entre clusters refleja la naturaleza continua de la evolución estelar y pone de manifiesto que el clustering no impone divisiones artificiales, sino que segmenta el espacio en regiones de mayor densidad.

El hecho de que esta separación emerja sin utilizar explícitamente etiquetas evolutivas ni parámetros discretos refuerza la idea de que las variables físicas seleccionadas contienen información suficiente para reconstruir las principales fases evolutivas de las estrellas.

La comparación con métodos de clustering alternativos permite evaluar hasta qué punto estas estructuras son robustas frente a la elección del algoritmo, aspecto que se discute a continuación.

4.3 Comparación entre métodos de clustering

La aplicación de distintos métodos de clustering no supervisado permite evaluar no sólo la estructura del espacio de parámetros, sino también la adecuación de cada algoritmo al objetivo físico del estudio.

El algoritmo *k-means*, al requerir la fijación previa del número de clusters, impone una partición explícita del espacio PCA. En este trabajo, esta característica resulta ventajosa, ya que permite segmentar de forma controlada un espacio de parámetros continuo en un número finito de grupos con propiedades físicas medias diferenciadas. La segmentación obtenida mediante

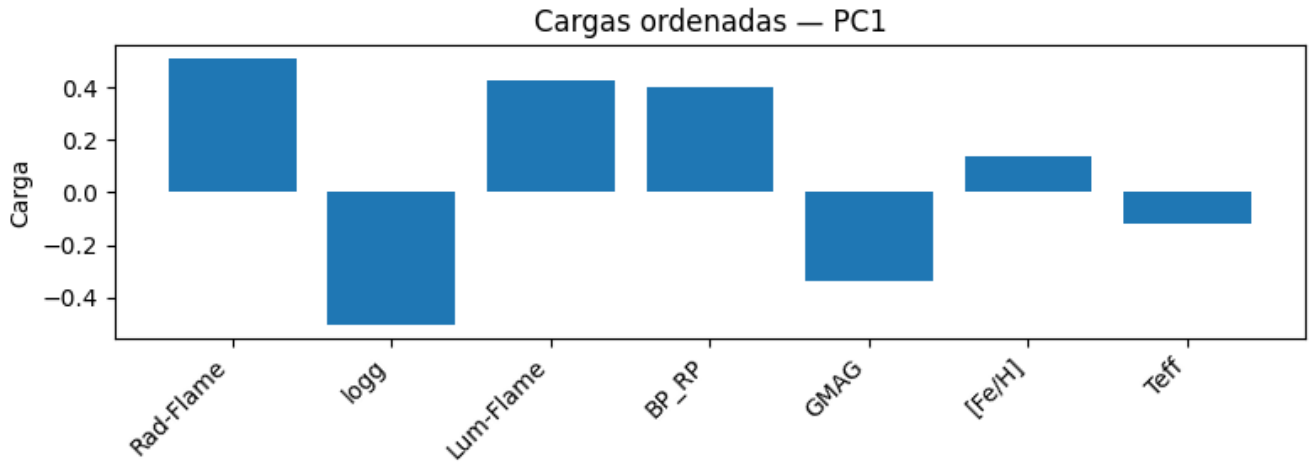

Figure 7. CargasPC1.

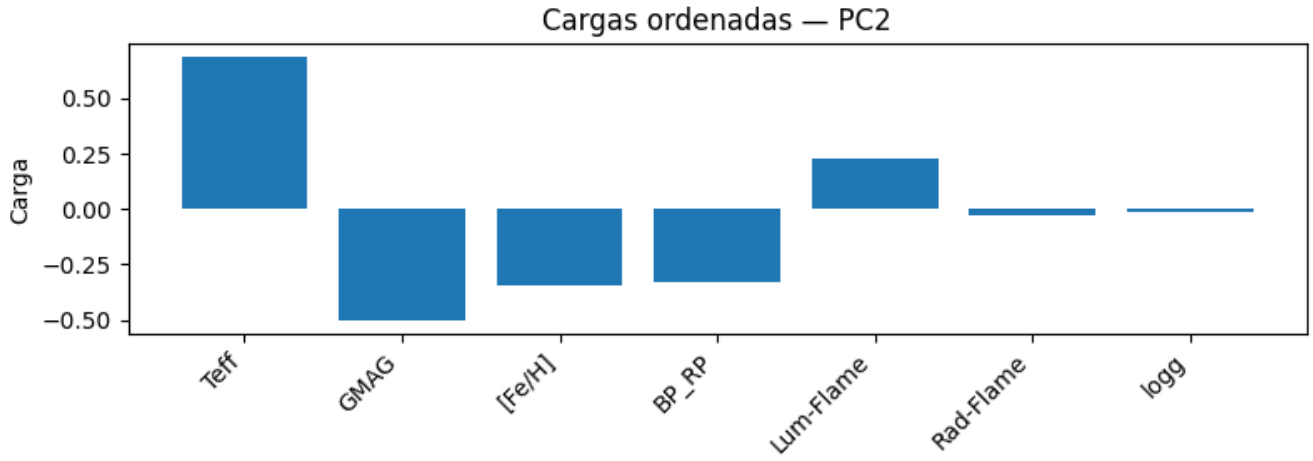

Figure 8. CargasPC2.

k-means produce clusters bien definidos tanto en el espacio PCA como en el diagrama de Hertzsprung–Russell, facilitando una interpretación física clara y una comparación directa entre grupos.

Por el contrario, el método DBSCAN identifica regiones de alta densidad sin imponer un número de clusters a priori, clasificando además una fracción de los datos como ruido. En el espacio PCA analizado, DBSCAN detecta únicamente un número reducido de clusters principales, correspondientes a grandes poblaciones estelares, sin capturar subdivisiones internas dentro de estas regiones densas. Aunque este comportamiento es coherente desde un punto de vista estadístico, limita la capacidad del método para distinguir subpoblaciones estelares con propiedades físicas diferenciadas.

Desde el punto de vista del objetivo de este trabajo —la exploración y clasificación física de estrellas a partir de sus parámetros fundamentales—, el clustering mediante *k-means* resulta claramente más adecuado. Este método permite identificar grupos con diferencias cuantificables en temperatura efectiva, gravedad superficial, luminosidad y radio estelar, proporcionando una clasificación más rica y detallada que la obtenida mediante DBSCAN.

Las métricas internas de validación reflejan estas diferencias metodológicas. Mientras que DBSCAN presenta una elevada compacidad para los clusters identificados, lo hace a costa de una segmentación excesivamente global. En cambio, *k-means* ofrece un equilibrio más apropiado entre compacidad y capacidad discriminante, alineándose mejor con la interpretación física de las poblaciones estelares clásicas.

En conjunto, aunque DBSCAN resulta útil como herramienta exploratoria para identificar grandes estructuras de densidad en el espacio de parámetros, los resultados obtenidos indican que *k-means* proporciona una clasificación más informativa y físicamente interpretable en el contexto de este estudio. Por este motivo, el análisis y la discusión posteriores se basan

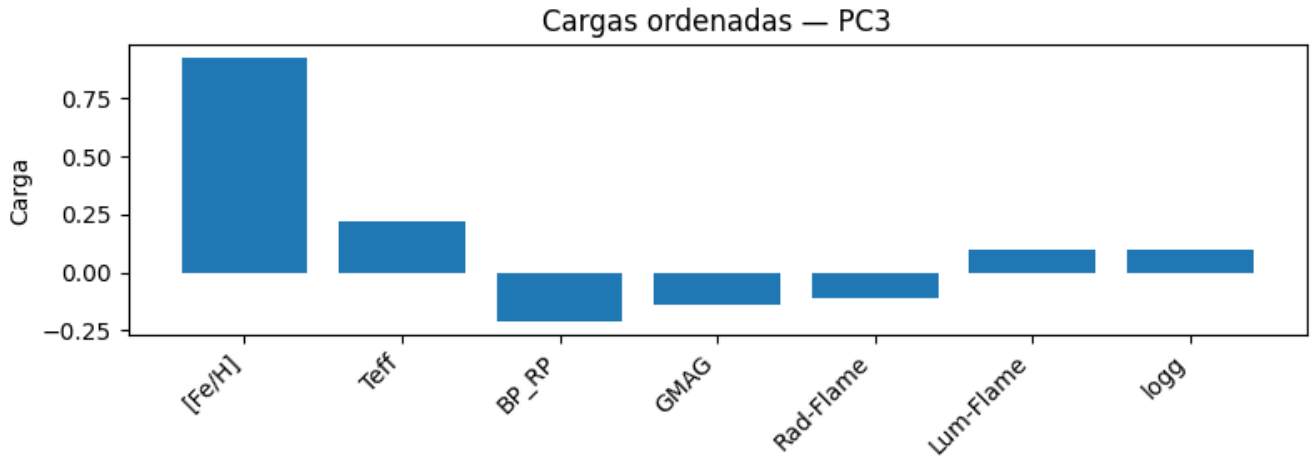

Figure 9. CargasPC3.

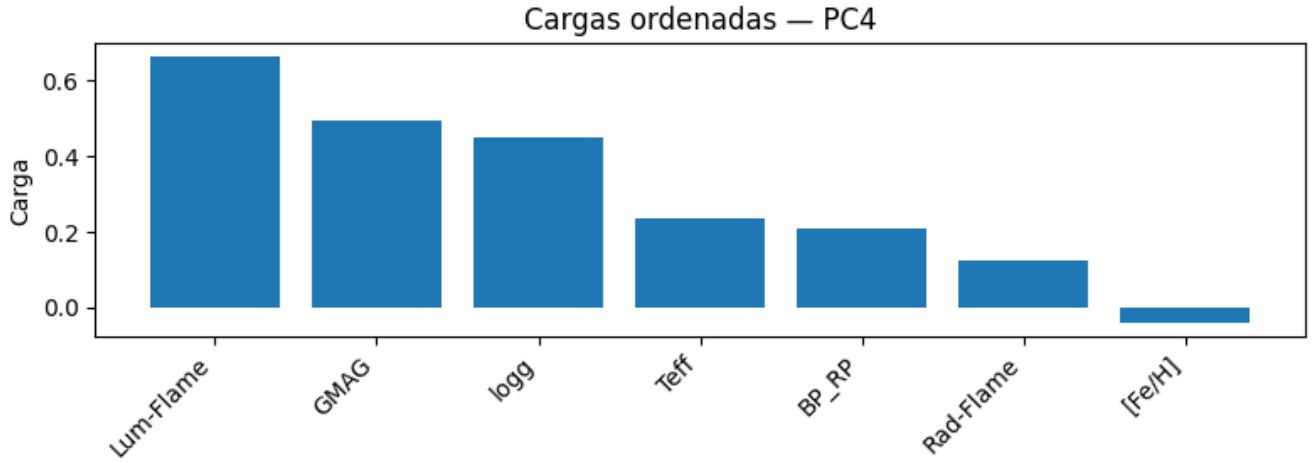

Figure 10. CargasPC4.

principalmente en los clusters obtenidos mediante *k-means*.

4.4 Separación frente a refinamiento en el espacio PCA

Los resultados muestran una diferencia clara entre componentes principales que realizan una separación global del conjunto de datos y aquellas que actúan como refinamiento interno de las poblaciones. Las primeras componentes organizan el espacio según propiedades físicas dominantes, mientras que las componentes posteriores introducen variaciones más finas relacionadas con composición química o diferencias evolutivas sutiles.

Esta distinción pone de manifiesto el papel complementario del PCA y del clustering: mientras que el PCA identifica las direcciones de máxima variación física, el clustering utiliza esta información para agrupar estrellas con propiedades similares, sin perder la continuidad inherente al sistema.

4.5 Limitaciones e implicaciones

Es importante destacar que el clustering no supervisado no pretende reproducir exactamente clasificaciones estelares discretas, sino ofrecer una segmentación basada en similitud estadística en el espacio de parámetros. La correspondencia con clases físicas conocidas debe interpretarse como una validación a posteriori, no como un objetivo impuesto.

Asimismo, la inclusión de parámetros derivados de modelos estelares, como el radio y la luminosidad, introduce una dependencia indirecta de dichos modelos, aunque sin imponer etiquetas evolutivas explícitas. Estas limitaciones no invalidan el análisis, pero deben tenerse en cuenta al interpretar los resultados.

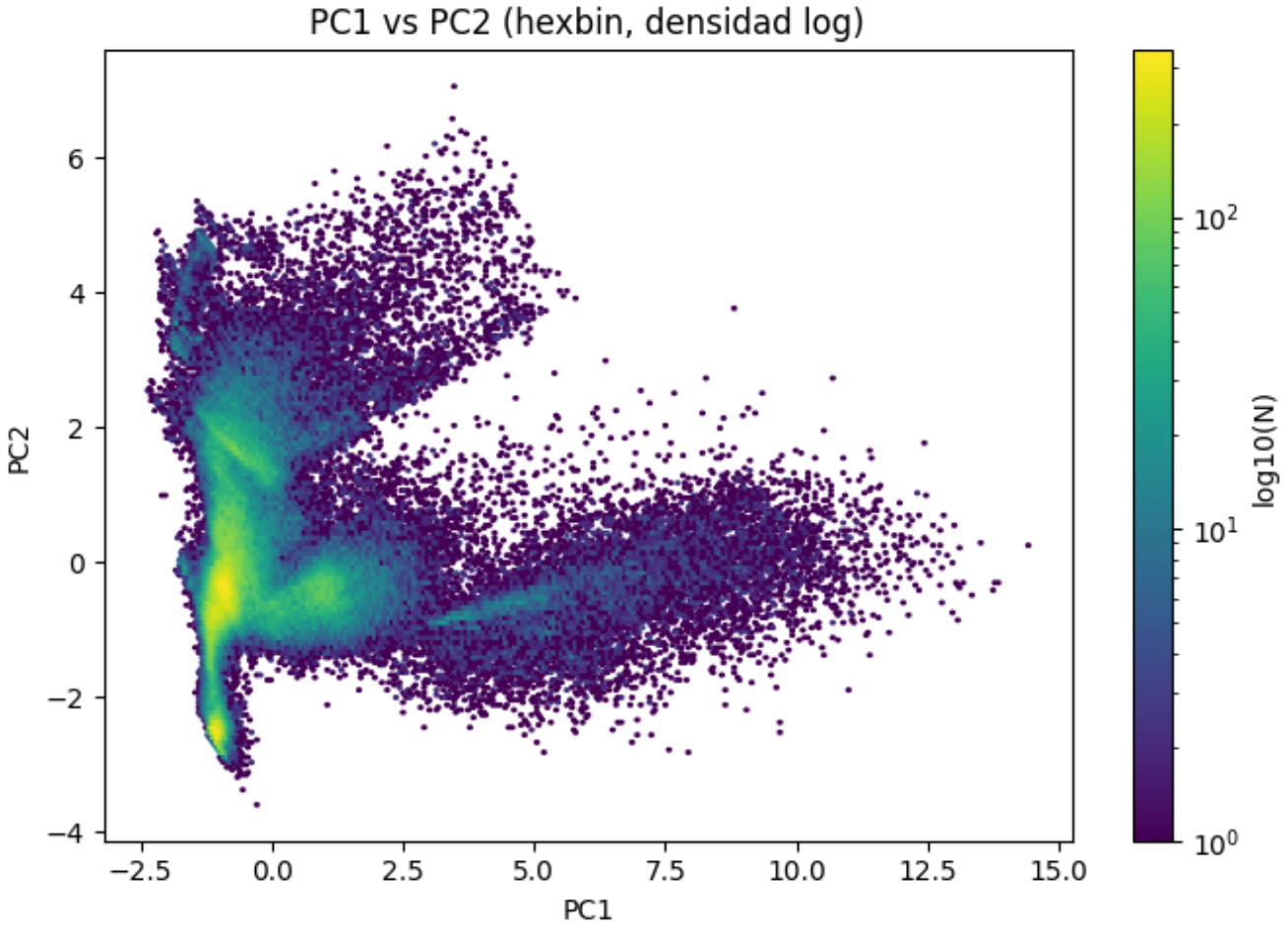

Figure 11. Proyección de las estrellas en el plano PC1–PC2. Se muestran regiones de alta y baja densidad en el espacio PCA.

4.6 Trabajo futuro

Existen varias extensiones naturales de este trabajo que permitirían profundizar en el análisis presentado.

Una posible línea de trabajo futuro consiste en ampliar el conjunto de variables consideradas, incorporando información cinemática o espectroscópica adicional disponible en Gaia DR3. Esto permitiría explorar la relación entre las propiedades físicas estelares y la estructura dinámica de la Galaxia, así como investigar posibles correlaciones entre poblaciones estelares y componentes galácticas.

Asimismo, podría evaluarse el uso de métodos de reducción de dimensionalidad no lineales, como t-SNE o UMAP, que pueden capturar estructuras más complejas en el espacio de parámetros. Aunque estos métodos presentan una interpretabilidad física más limitada que el PCA, su comparación con enfoques lineales podría aportar información complementaria sobre la geometría del espacio estelar.

En el ámbito del clustering, el uso de algoritmos jerárquicos o de modelos probabilísticos permitiría explorar segmentaciones alternativas y evaluar la estabilidad de los grupos identificados. También sería interesante analizar con mayor detalle los puntos clasificados como ruido por DBSCAN, ya que podrían corresponder a poblaciones raras o a fases evolutivas menos representadas.

Finalmente, una extensión natural del análisis consistiría en comparar los clusters obtenidos con clasificaciones estelares independientes disponibles en la literatura, utilizando estas únicamente como validación a posteriori. Este enfoque permitiría cuantificar de forma más precisa la relación entre las estructuras emergentes del aprendizaje no supervisado y las clasificaciones físicas tradicionales.

4.7 Síntesis

En conjunto, el estudio demuestra que el uso combinado de análisis de componentes principales y clustering no supervisado permite explorar de forma eficaz la estructura del espacio de parámetros estelares de Gaia DR3. Los resultados obtenidos

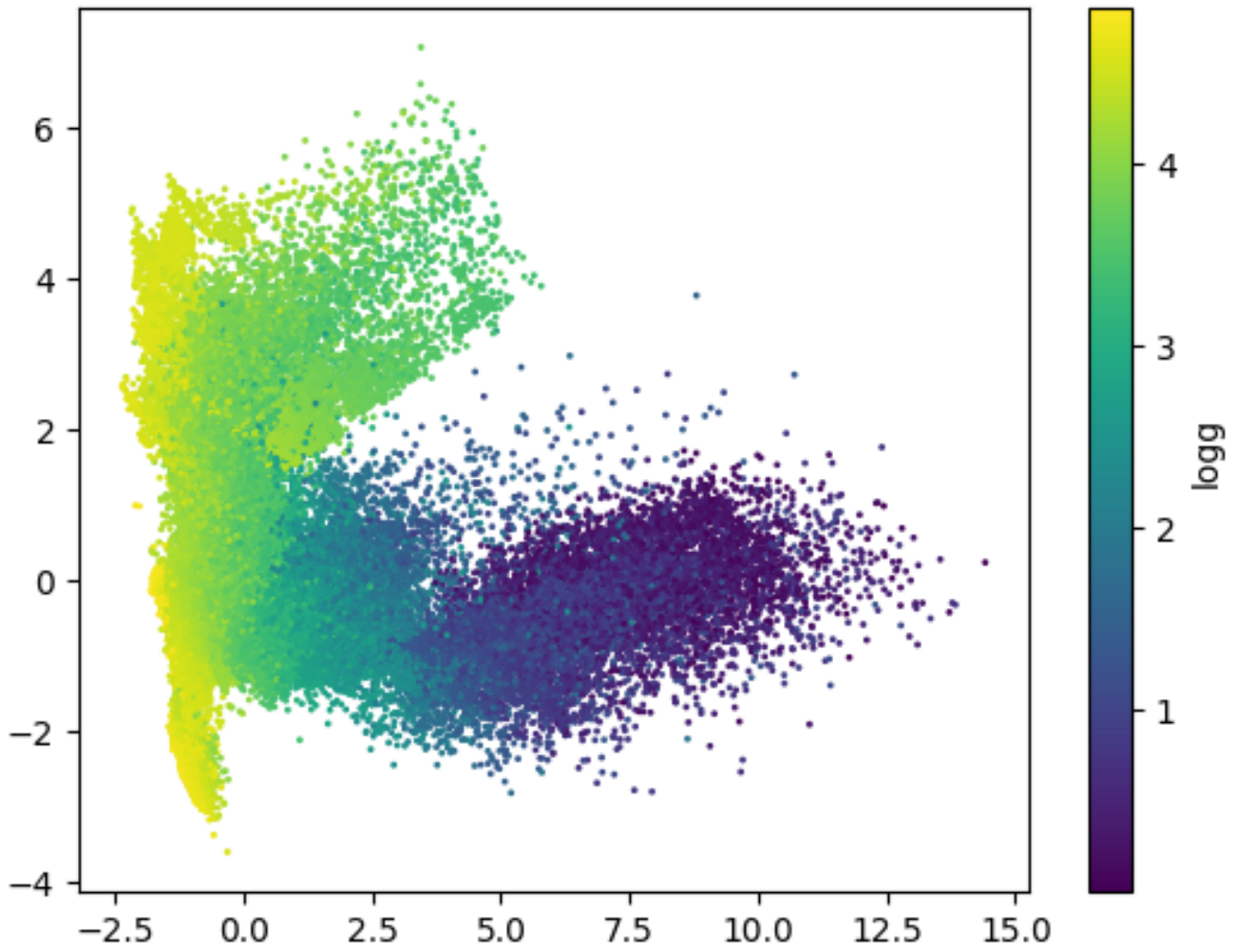

Figure 12. Proyección PC1–PC2 coloreada por $\log g$.

muestran que las principales fases evolutivas y secuencias estelares emergen de manera natural a partir de los datos, confirmando el potencial de estas técnicas para el análisis exploratorio de grandes catálogos astronómicos.

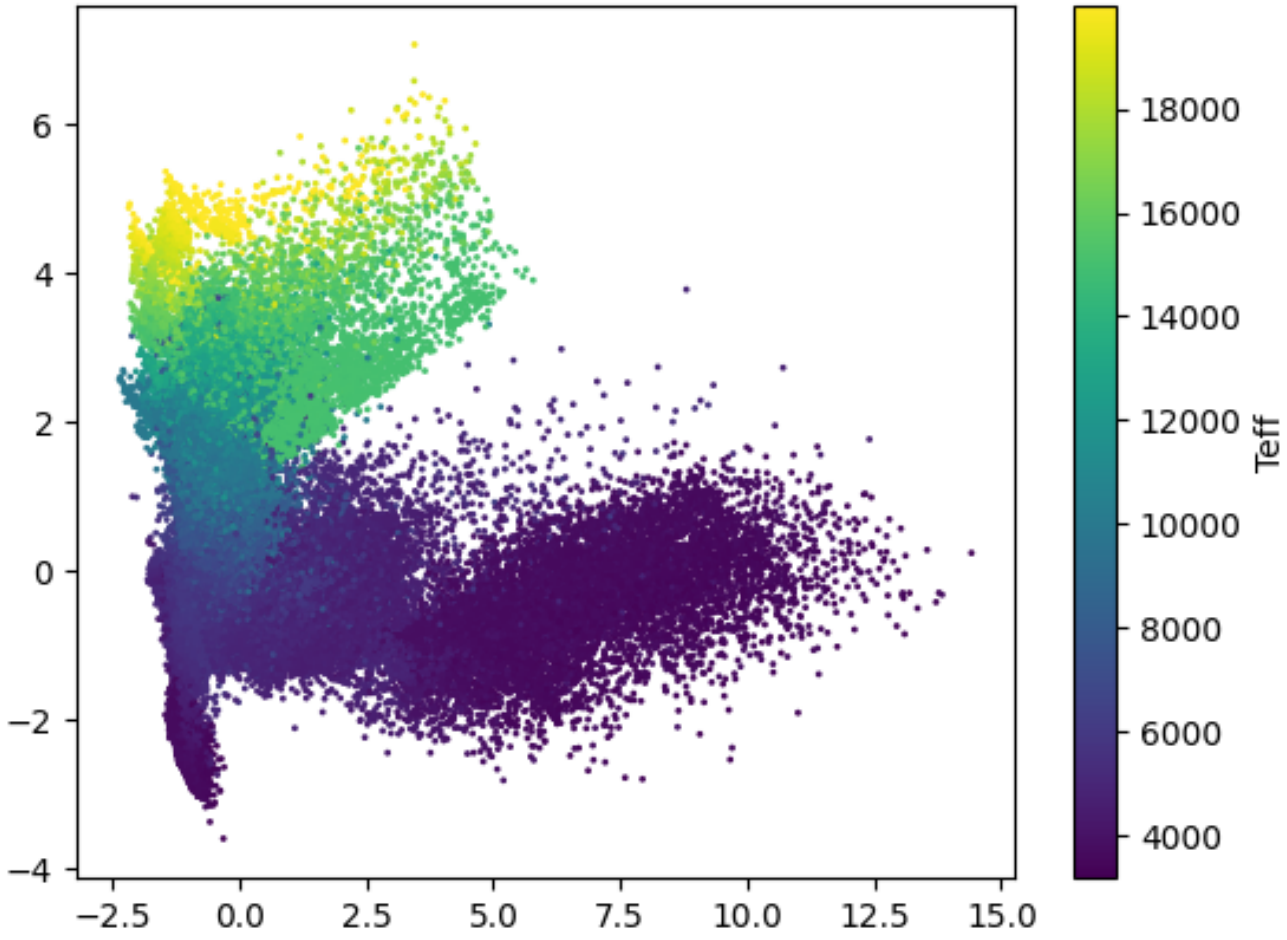

Figure 13. Proyección PC1–PC2 coloreada por $T_{\rm eff}$.

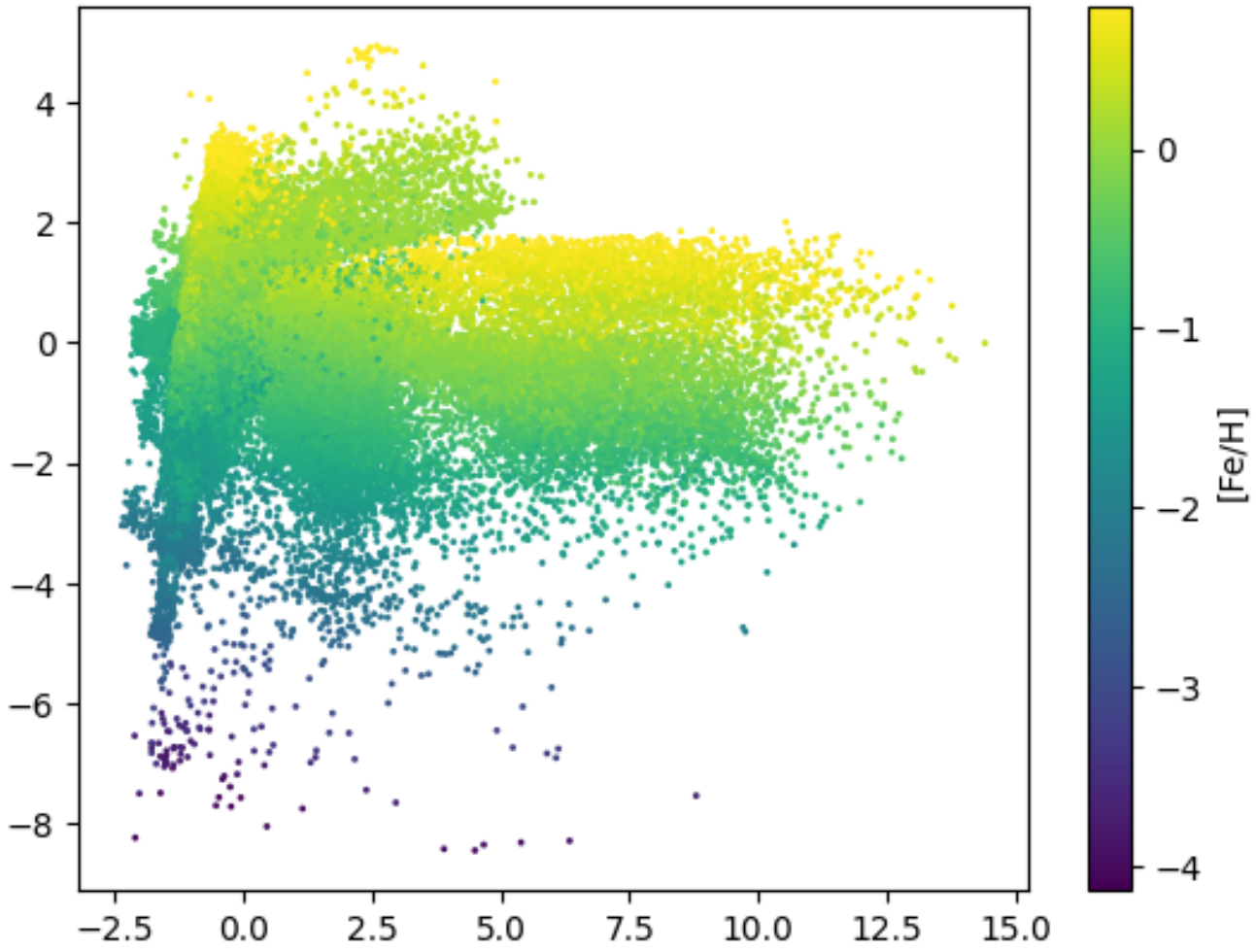

Figure 14. Proyección PC1–PC2 coloreada por [Fe/H].

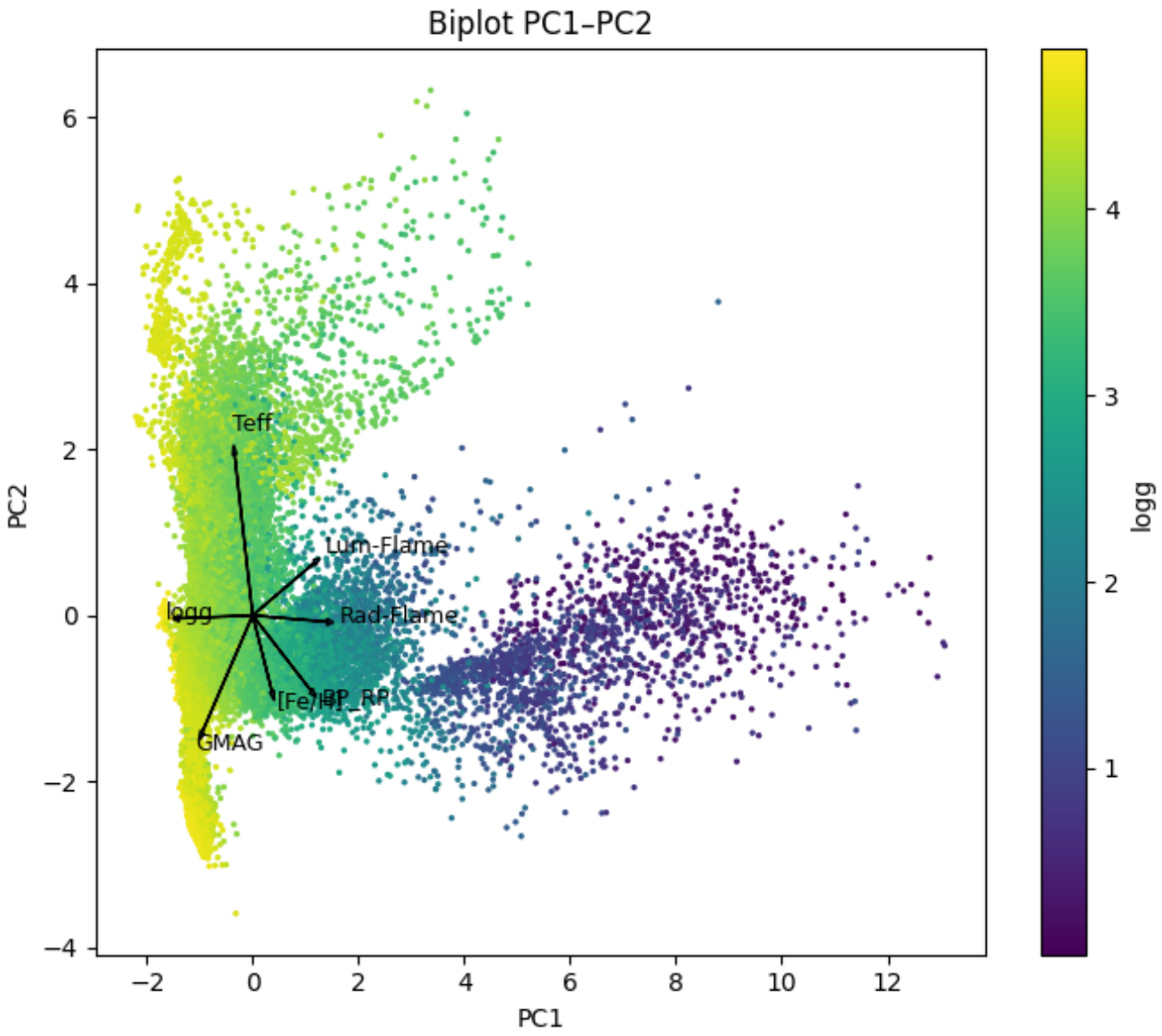

Figure 15. Biplot en el plano PC1–PC2 mostrando la orientación de las variables originales en el espacio de componentes principales.

Figure 16. Métricas internas de validación del clustering (*silhouette* y Davies–Bouldin) en función del número de clusters *k*.

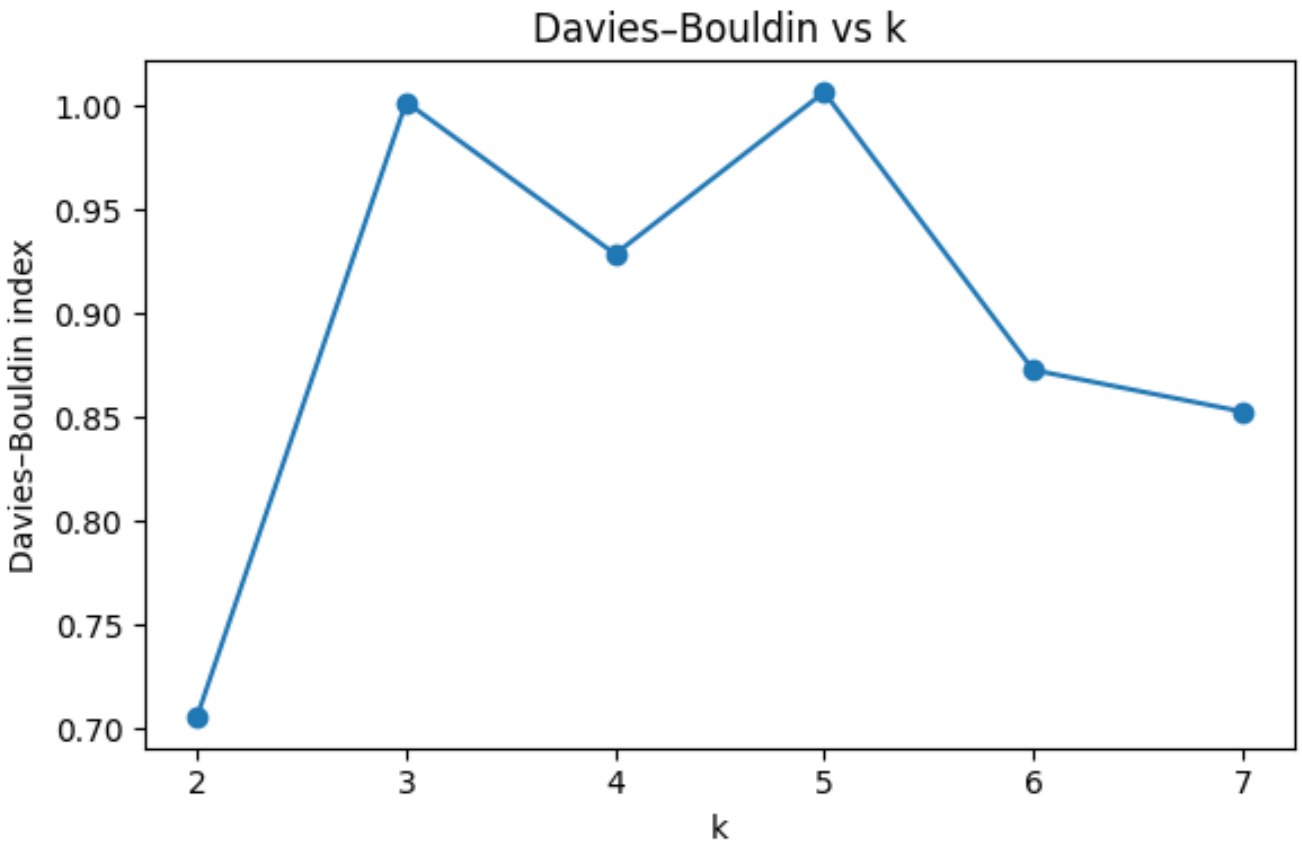

Figure 17. Métricas internas de validación del clustering (*silhouette* y Davies–Bouldin) en función del número de clusters *k*.

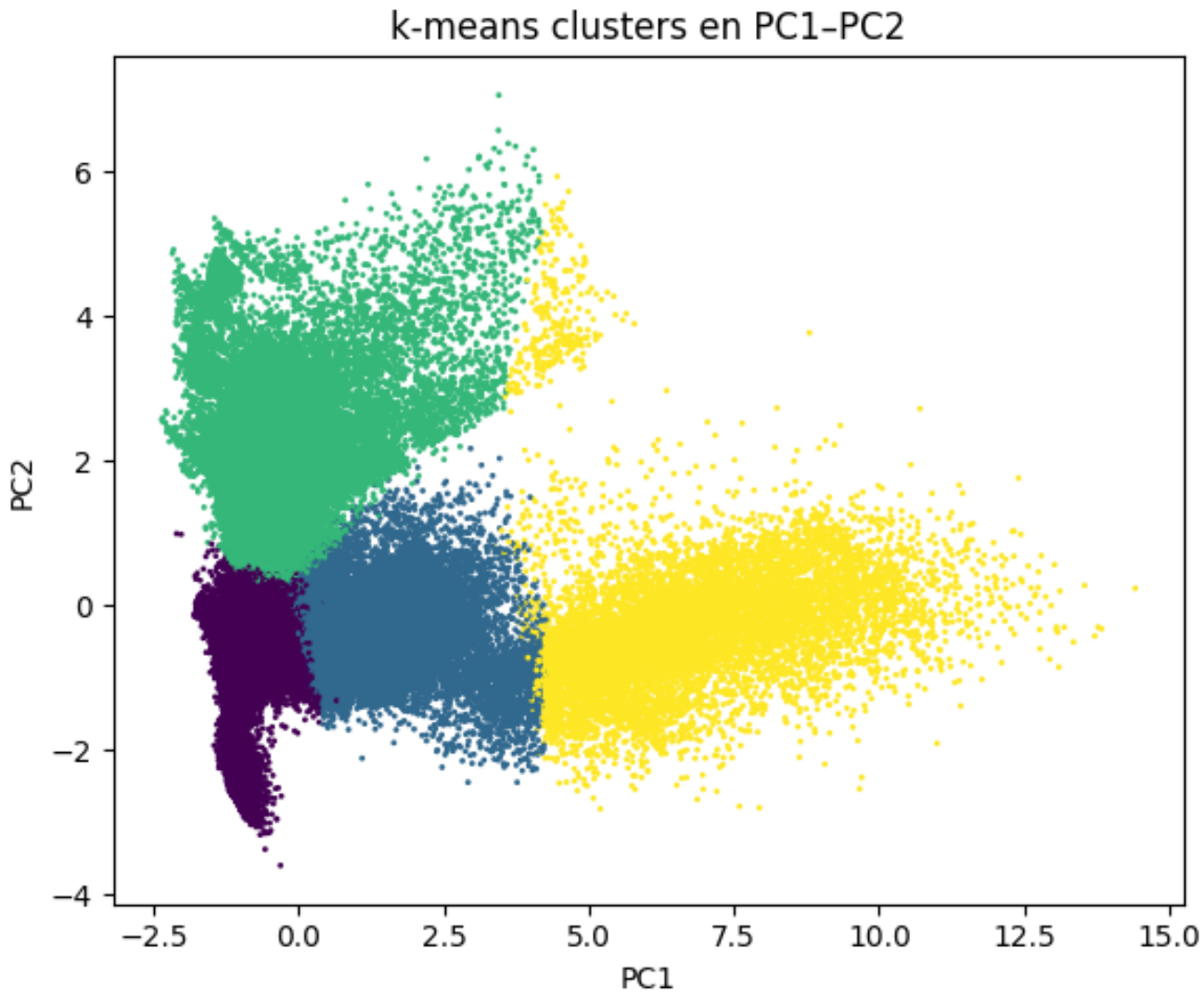

Figure 18. Proyección de los clusters obtenidos mediante *k-means* en el plano PC1–PC2.

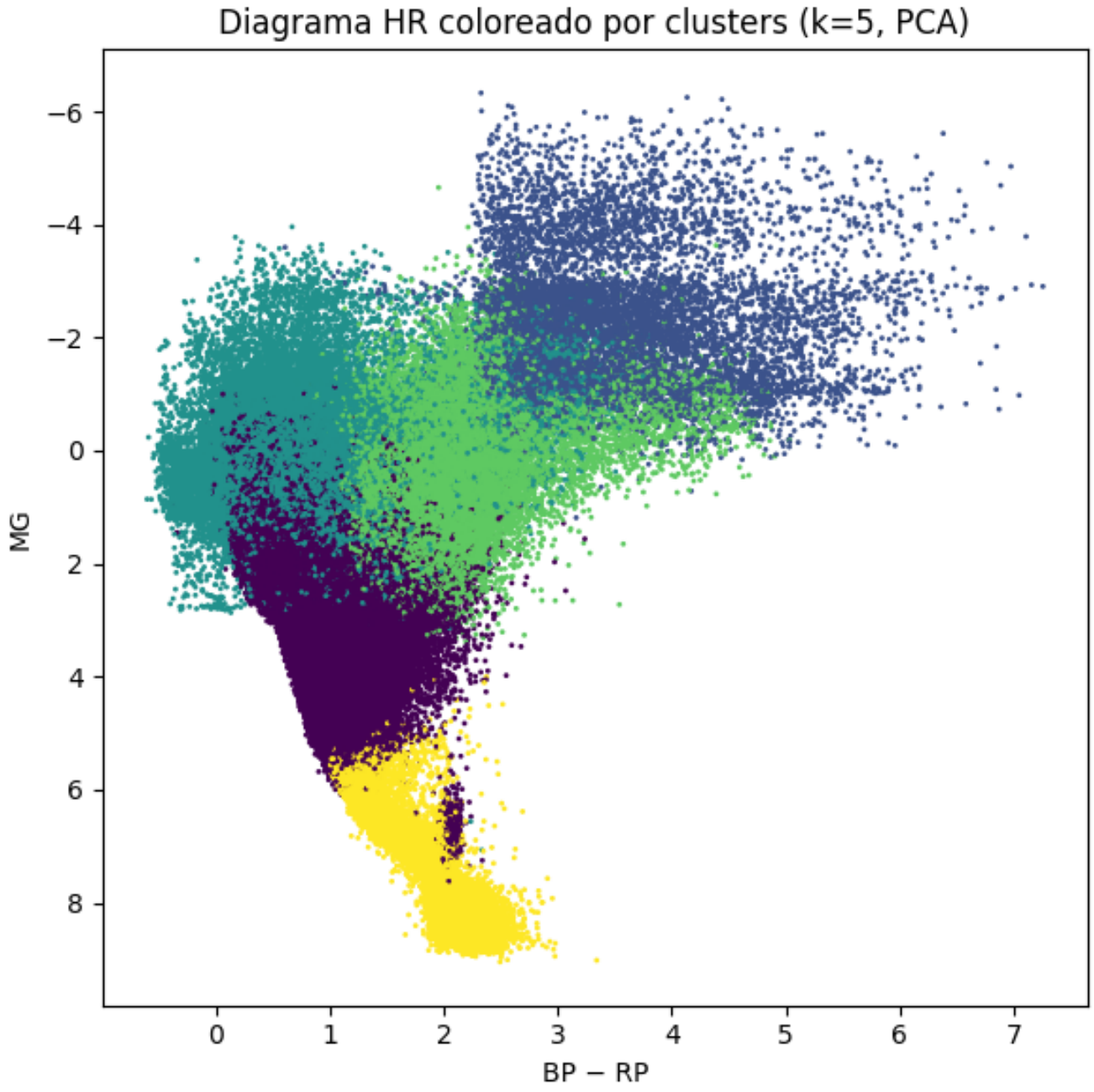

Figure 19. Diagrama de Hertzsprung–Russell (M_G frente a $BP - RP$) coloreado según los clusters obtenidos con *k-means*.

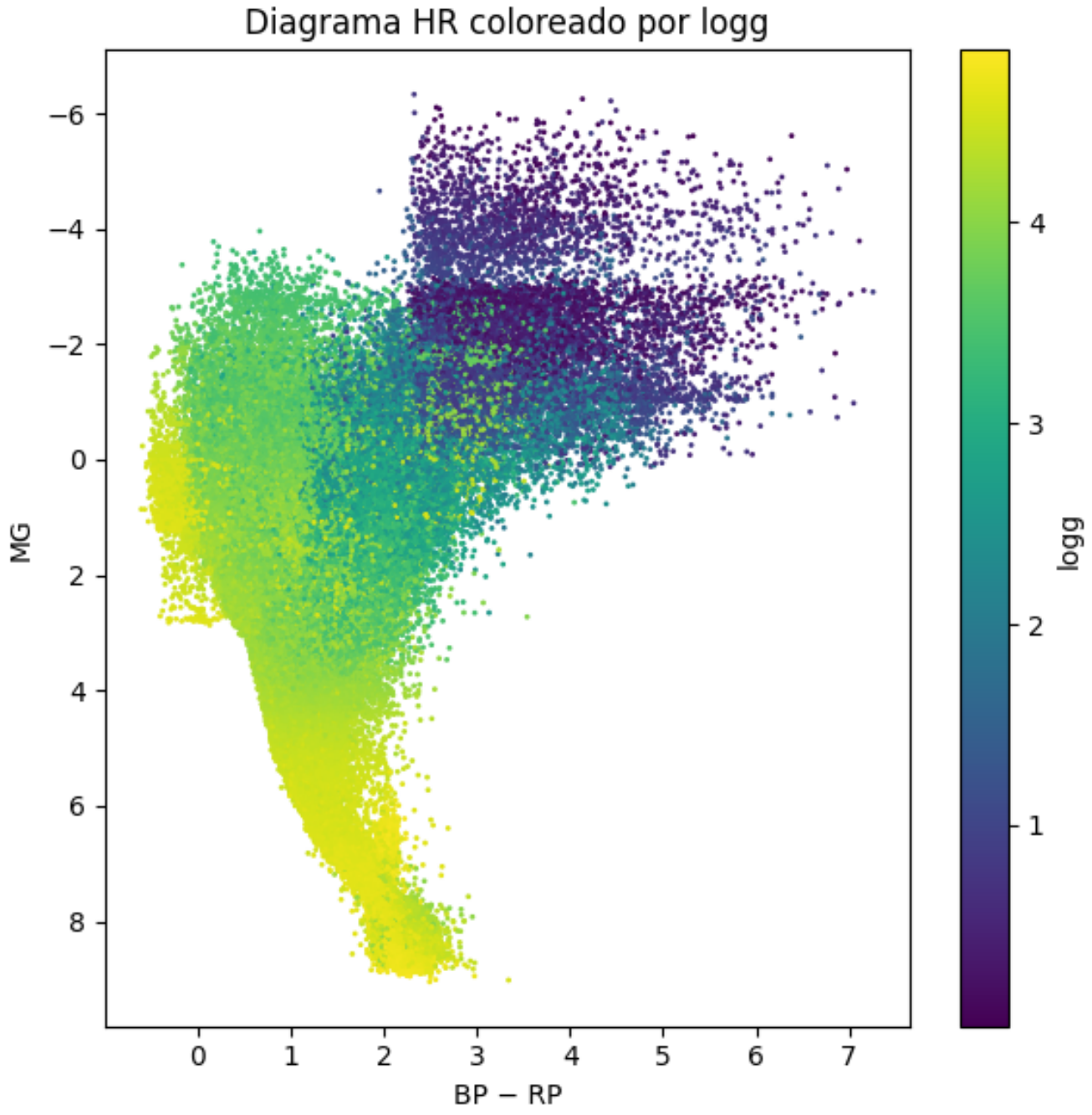

Figure 20. Diagrama de Hertzsprung–Russell (M_G frente a $BP - RP$) coloreado según logg.

	Teff		logg		[Fe/H]		\
	mean	std	mean	std	mean	std	
cluster_km							
0	6456.035395	1137.677046	4.079502	0.261683	-0.355716	0.330986	
1	3992.805546	1661.298446	0.831310	0.561690	-0.089288	0.500459	
2	11719.389029	2624.789543	3.970679	0.306438	-0.754171	0.525428	
3	5193.727021	753.229980	2.706735	0.543863	-0.198009	0.380797	
4	4129.105177	536.136204	4.610337	0.122950	-0.139291	0.352221	

	GMAG		Rad-Flame		Lum-Flame	
	mean	std	mean	std	mean	std
cluster_km						
0	3.171411	1.307660	1.841215	0.804320	9.019045	14.584645
1	-2.181461	1.161844	77.461358	27.194580	1127.422611	665.131485
2	0.188517	1.056877	2.743137	1.250147	226.926739	396.452001
3	0.377468	1.008927	12.240699	8.429575	97.213758	95.761459
4	7.619904	1.034174	0.636073	0.108054	0.139733	0.142544

Figure 21. Datos de los clusters.

```
Silhouette (sin ruido): 0.43440744669214926
Davies-Bouldin (sin ruido): 0.47264707732511707
```

Figure 22. metricas dbscan.

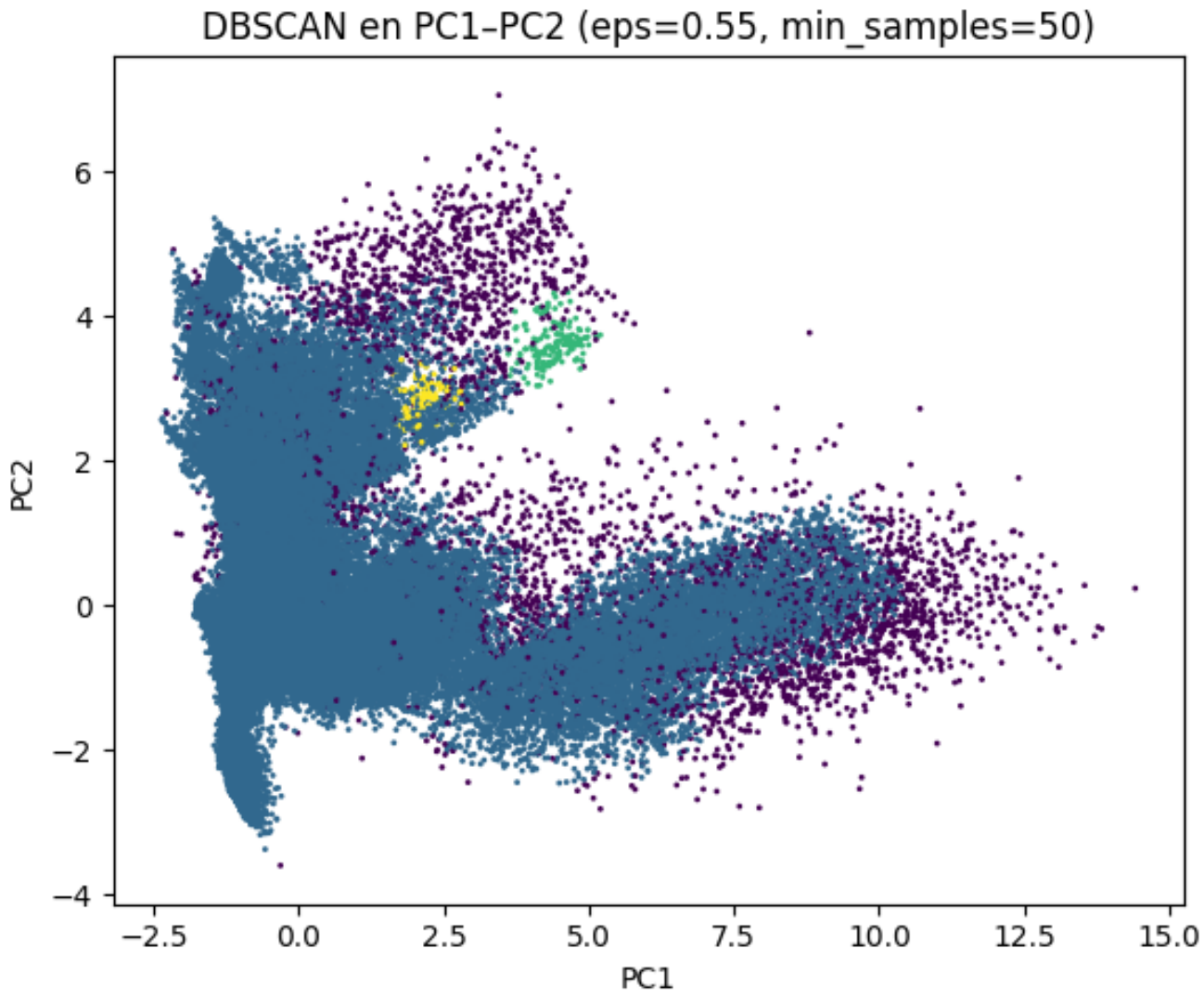

Figure 23. Clusters dbscan en PC1-PC2.

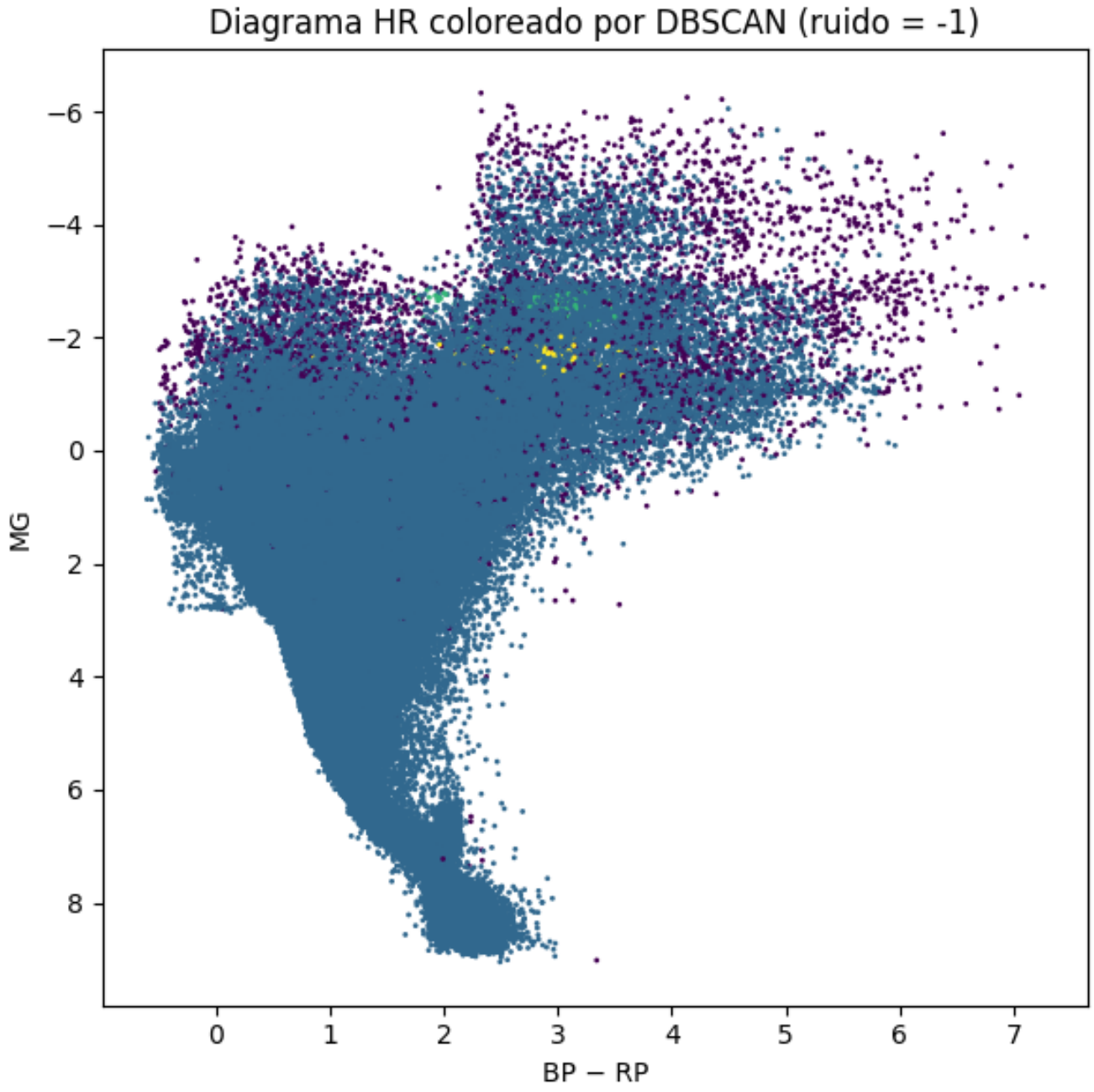

Figure 24. Clusters dbscan en HR.