



**UNIVERSIDAD FRANCISCO DE VITORIA**

**ESCUELA POLITÉCNICA SUPERIOR**

**GRADO EN INGENIERÍA MATEMÁTICA**

**PROYECTO FINAL DE GRADO**

**MODALIDAD INGENIERÍA**

**ANÁLISIS SOCIOECONÓMICO Y  
VISUALIZACIÓN DE DESIGUALDADES EN  
LOS DISTRITOS Y BARRIOS DE MADRID**

[Claudia Esnarrizaga Rodríguez]

Convocatoria de [Julio] [2025]

## CALIFICACIÓN DEL PROYECTO FINAL DE GRADO

CUALITATIVA:	
NUMÉRICA:	

Conforme Presidente:	Conforme Secretario:
Fdo.:	Fdo.:

Conforme Vocal:	Conforme Vocal:	Conforme Vocal:
Fdo.:	Fdo.:	Fdo.:

Lugar y fecha: Pozuelo de Alarcón, a \_\_\_\_ de \_\_\_\_\_ de 202\_\_\_\_



*La verdadera medida de una sociedad es cómo trata a sus miembros más vulnerables. Mahatma Gandhi.*



*A mi familia, por su apoyo incondicional.*

*A quienes me han acompañado, guiado y alentado durante este proceso, en los momentos de mayor esfuerzo y también en los de mayor aprendizaje.*

*Y, en especial, a todos aquellos que contribuyen cada día a construir ciudades más justas y equitativas. Este trabajo es también un pequeño aporte a ese propósito colectivo.*



# Agradecimientos

Este trabajo no solo refleja el esfuerzo, aprendizaje y crecimiento personal de estos años, sino que también es fruto del apoyo constante de todas las personas que han estado a mi lado en cada etapa.

En primer lugar, a mi familia, por ser siempre mi mayor pilar. A mi padre, ejemplo de esfuerzo, inteligencia, ambición y perseverancia. Su dedicación diaria, capacidad de trabajo y superación han sido siempre una inspiración para mí. A mi madre, por saber encontrar siempre el equilibrio y ayudarme a relativizar las dificultades. Por enseñarme el valor de la empatía, la sensibilidad y los principios sólidos que me han guiado tanto personal como académica y profesionalmente. A mi hermano, por ser siempre fuente de amor; y a mi hermana, por ser mi mayor confidente, mi amiga y compañera de vida.

A Iván, por estar a mi lado en cada momento de este proceso, por su apoyo incondicional y su capacidad de animarme en los momentos más exigentes. Por ser ejemplo de constancia, disciplina y exigencia sana. Por enseñarme siempre a encontrar el lado positivo incluso en los momentos más vulnerables. Gracias por ser siempre un impulso para mejorar y seguir adelante.

A mis amigas Marina, Nerea, Claudia, Ainhoa, Paula y Mayte, porque vuestra amistad ha sido siempre un refugio seguro en cada etapa de mi vida. Habéis estado presentes desde siempre, y vuestra presencia ha hecho este camino más llevadero y feliz.

A todos los docentes que han formado parte de mi aprendizaje, no solo en este proyecto sino a lo largo de la carrera, gracias por vuestra dedicación y vocación. En especial, a mi tutora Natalia, que desde el primer día ha estado presente, guiándome, enseñándome y transmitiéndome esa pasión por las matemáticas. Sin su acompañamiento y cercanía este trabajo no habría sido posible.

A mis compañeros, Celia, Julia, María y Jaime, por la ayuda mutua, el trabajo y el día a día compartidos, siempre dispuestos a aportar y a buscar juntos la mejor versión de cada uno de nosotros. Y especialmente a Inés, por ser ese apoyo incondicional durante todos estos años, por no dejarme rendirme en los momentos difíciles, y por acompañarme en cada ámbito de esta etapa.

Y finalmente, a Dios, que en silencio ha sostenido cada paso, especialmente en los momentos de incertidumbre y debilidad, recordándome que el esfuerzo y sacrificio siempre tienen sentido cuando no se camina solo.



# Resumen

El presente trabajo aborda el análisis de la vulnerabilidad territorial en los barrios y distritos de la ciudad de Madrid, partiendo de un amplio conjunto de indicadores socioeconómicos, educativos, de salud y vulnerabilidad agregada. A partir de la base de datos oficial del Ayuntamiento de Madrid [1], se ha construido una base de datos estructurada y normalizada que ha permitido aplicar técnicas de *clustering* no supervisado (*K-means*, *Agglomerative* y *DBSCAN*) con diferentes configuraciones de número de grupos. Los resultados obtenidos han sido validados mediante el *Silhouette Score* y el *Calinski-Harabasz Score*, permitiendo seleccionar los modelos más estables para cada nivel territorial. A partir de los grupos generados, se ha calculado un ranking de vulnerabilidad, representado posteriormente en mapas interactivos. El análisis muestra cómo determinados indicadores relacionados con el nivel educativo, la estructura demográfica y la renta disponible tienen un peso clave en la diferenciación territorial. Finalmente, se proponen futuras líneas de mejora para el perfeccionamiento metodológico y su posible aplicación práctica para reducir desigualdades.

## Palabras claves

Vulnerabilidad territorial, *clustering*, *K-means*, *Agglomerative*, desigualdad urbana, análisis multivariante, Madrid.

## Abstract

*This work addresses the territorial vulnerability analysis of the districts and neighborhoods of Madrid, based on a comprehensive set of socioeconomic, educational, health and aggregated vulnerability indicators. Using the official database from the Madrid City Council [1], a structured and normalized database has been built, allowing the application of multivariate analysis techniques. Several unsupervised clustering algorithms (*K-means*, *Agglomerative* and *DBSCAN*) have been applied with different cluster configurations. The results have been validated through the *Silhouette Score* and the *Calinski-Harabasz Score*, allowing the selection of the most stable models for each territorial level. Based on the generated groups, a vulnerability ranking has been calculated and subsequently represented through interactive maps. The analysis reveals that indicators related to educational attainment, demographic structure, and income level are key factors in territorial differentiation. Finally, possible future improvements are proposed for methodological refinement and potential practical application in public policy planning to reduce inequalities.*

## Keywords

*Territorial vulnerability, clustering, K-means, Agglomerative, urban inequality, multivariate analysis.*



# Índice de Contenidos

<b>1.</b>	<i>Introducción</i>	<b>1</b>
1.1.	Justificación	1
1.2.	Motivación	2
1.3.	Alcance	2
<b>2.</b>	<i>Investigación previa</i>	<b>5</b>
2.1.	Vulnerabilidad socioeconómica	5
2.2.	Métodos de evaluación de la vulnerabilidad	6
2.3.	<i>Clustering</i> en análisis socioeconómico	7
2.4.	Aplicación del <i>clustering</i> al análisis de vulnerabilidad en Madrid	8
2.5.	Metodologías alternativas de <i>clustering</i> en contexto socioeconómico	9
<b>3.</b>	<i>Objetivos</i>	<b>13</b>
3.1.	Objetivo general	13
3.2.	Lista de objetivos específicos	13
3.3.	Métodos de Validación	13
<b>4.</b>	<i>Plan de Desarrollo del Proyecto</i>	<b>15</b>
4.1.	Metodología	15
4.2.	Tecnologías	16
4.3.	Plan de desarrollo del proyecto	19
4.3.1.	PT 1 Análisis y definición del problema	19
4.3.2.	PT 2 Creación y limpieza de la base de datos	19
4.3.3.	PT 3 Análisis a priori	20
4.3.4.	PT 4 Minería de datos	20
4.4.	Plan de Trabajo	21
4.5.	Recursos	21
4.6.	Costes	22
4.7.	Condicionantes y Limitaciones	22
<b>5.</b>	<i>Desarrollo de la Solución Técnica</i>	<b>23</b>
5.1.	PT 1 – Análisis y definición del problema	23
5.1.1.	Revisión teórica del concepto de vulnerabilidad urbana	23
5.1.2.	Nociones básicas y conocimientos previos	24
5.2.	PT 2 – Limpieza y creación de la base de datos	27
5.2.1.	Selección y clasificación de variables	27
5.2.2.	Arquitectura de la base de datos	35
5.3.	Análisis a priori	37
5.3.1.	Clasificación territorial: diferenciación entre barrios y distritos	37
5.3.2.	Filtrado de variables según completitud	39

5.3.3.	Eliminación de observaciones incompletas y reconstrucción del <i>dataset</i> .....	39
5.3.4.	Normalización y estandarización de variables .....	40
5.3.5.	Variables seleccionadas para el análisis.....	40
<b>5.4.</b>	<b>Aplicación de técnicas de <i>clustering</i> .....</b>	<b>43</b>
5.4.1.	<i>K-means</i> .....	43
5.4.2.	DBSCAN .....	47
5.4.3.	<i>Agglomerative Clustering</i> .....	50
<b>5.5.</b>	<b>Generación de rankings y representación espacial .....</b>	<b>54</b>
5.5.1.	Metodología cálculo ranking.....	54
5.5.2.	Visualización de los agrupamientos mediante análisis PCA .....	55
5.5.3.	Representación cartográfica rankings.....	57
<b>6.</b>	<b>Resultados .....</b>	<b>64</b>
<b>6.1.</b>	<b>Validación de los modelos .....</b>	<b>64</b>
6.1.1.	Validación mediante <i>Silhouette Score</i> .....	64
6.1.2.	Validación mediante <i>Calinski-Harabasz Score</i> .....	65
<b>6.2.</b>	<b>Análisis de los perfiles de los clústeres .....</b>	<b>67</b>
6.2.1.	Resultados a nivel de barrios ( <i>K-means</i> k=3) .....	67
6.2.2.	Resultados a nivel de distritos ( <i>Agglomerative</i> k=4).....	71
<b>6.3.</b>	<b>Análisis comparativo entre nivel barrio y nivel distrito .....</b>	<b>75</b>
<b>6.4.</b>	<b>Ranking final de vulnerabilidad.....</b>	<b>75</b>
<b>7.</b>	<b>Implicaciones Éticas e Impacto Social.....</b>	<b>77</b>
7.1.	Introducción .....	77
7.2.	Marcos normativos y principios éticos de referencia .....	77
7.3.	Riesgos identificados y medidas adoptadas.....	79
7.4.	Matriz de riesgos del proyecto.....	80
7.5.	Impacto social esperado .....	82
7.6.	Limitaciones y posibles líneas futuras del proyecto .....	83
7.7.	Conclusión .....	83
7.7.1.	Viabilidad ética del proyecto .....	83
<b>8.</b>	<b>Conclusiones .....</b>	<b>85</b>
<b>9.</b>	<b>Otros Méritos del Proyecto.....</b>	<b>87</b>
<b>10.</b>	<b>Bibliografía .....</b>	<b>89</b>
<b>Anexo A.....</b>		<b>97</b>
<b>Anexo B.....</b>		<b>103</b>

# Índice de Tablas

<i>Tabla 1: Tabla resumen indicadores vulnerabilidad .....</i>	6
<i>Tabla 2: Tabla resumen índices vulnerabilidad.....</i>	6
<i>Tabla 3: Tabla resumen del Estado del Arte .....</i>	11
<i>Tabla 4. Tabla resumen de técnicas y herramientas utilizadas .....</i>	26
<i>Tabla 5. Tabla de indicadores iniciales de la dimensión de Bienestar Social e Igualdad.....</i>	29
<i>Tabla 6. Tabla de indicadores iniciales de la dimensión de Educación y Cultura .....</i>	31
<i>Tabla 7. Tabla de indicadores iniciales de la dimensión de Medio Ambiente Urbano y Movilidad....</i>	31
<i>Tabla 8. Tabla de indicadores iniciales de la dimensión de Economía y Empleo.....</i>	32
<i>Tabla 9. Tabla de indicadores iniciales de la dimensión de Salud .....</i>	33
<i>Tabla 10. Tabla de indicadores iniciales de la dimensión de Vulnerabilidad Territorial (IGUALA).....</i>	34
<i>Tabla 11. Tabla de indicadores finales de la dimensión de Bienestar Social e Igualdad .....</i>	41
<i>Tabla 12. Tabla de indicadores finales de la dimensión de Educación y Cultura.....</i>	41
<i>Tabla 13. Tabla de indicadores finales de la dimensión de Economía y Empleo .....</i>	42
<i>Tabla 14. Tabla de indicadores finales de la dimensión de Vulnerabilidad Territorial (IGUALA) .....</i>	42
<i>Tabla 15. Nº barrios por clúster (K-means k=4).....</i>	45
<i>Tabla 16. Nº distritos por clúster (K-means k=4) .....</i>	45
<i>Tabla 17. Nº barrios por clúster (K-means k=3).....</i>	46
<i>Tabla 18. Nº distritos por clúster (K-means k=3) .....</i>	47
<i>Tabla 19. Ajuste parámetros DBSCAN .....</i>	48
<i>Tabla 20. Nº barrios por clúster (DBSCAN) .....</i>	49
<i>Tabla 21. Nº distritos por clúster (DBSCAN).....</i>	49
<i>Tabla 22. Nº barrios por clúster (Agglomerative k=4) .....</i>	52
<i>Tabla 23. Nº distritos por clúster (Agglomerative k=4) .....</i>	52
<i>Tabla 24. Nº barrios por clúster (Agglomerative k=3) .....</i>	53
<i>Tabla 25. Nº distritos por clúster (Agglomerative k=3) .....</i>	53
<i>Tabla 26. Categoría asignada según nº ranking para k=3.....</i>	55
<i>Tabla 27. Categoría asignada según nº ranking para k=4.....</i>	55
<i>Tabla 28. Silhouette Score por modelo (barrios).....</i>	65
<i>Tabla 29. Silhouette Score por modelo (distritos).....</i>	65
<i>Tabla 30. Calinski-Harabasz Score por modelo (barrios).....</i>	66
<i>Tabla 31. Calinski-Harabasz Score por modelo (distritos) .....</i>	66
<i>Tabla 32. Variables más influyentes en clúster 0 (barrios) .....</i>	68
<i>Tabla 33. Variables más influyentes en clúster 1 (barrios) .....</i>	68

<i>Tabla 34. Variables más influyentes en clúster 2 (barrios) .....</i>	68
<i>Tabla 35. Resumen variables más influyentes en modelo K-means k=3 (barrios).....</i>	69
<i>Tabla 36. Variables más influyentes en clúster 0 (distritos).....</i>	71
<i>Tabla 37. Variables más influyentes en clúster 1 (distritos).....</i>	71
<i>Tabla 38. Variables más influyentes en clúster 2 (distritos).....</i>	72
<i>Tabla 39. Variables más influyentes en clúster 3 (distritos).....</i>	72
<i>Tabla 40. Resumen variables más influyentes en modelo Agglomerative k=4 (distritos) .....</i>	73
<i>Tabla 41. Matriz descriptiva de riesgos del proyecto.....</i>	81
<i>Tabla 42. Matriz resumen de riesgos del proyecto .....</i>	81
<i>Tabla 43. Accesos mapas interactivos barrios .....</i>	97
<i>Tabla 44. Accesos mapas interactivos distritos.....</i>	97
<i>Tabla 45. Ranking final barrios.....</i>	101
<i>Tabla 46. Ranking final distritos.....</i>	102

# Índice de Figuras

<i>Imagen 1: Tabla índices socioeconómicos y de vulnerabilidad [11]</i> .....	7
<i>Imagen 2: Diagrama de Gantt</i> .....	21
<i>Imagen 3. Reorganización a estructura tipo data warehouse</i> .....	35
<i>Imagen 4. Esquema de arquitectura de un data warehouse [70]</i> .....	36
<i>Imagen 5. Tabla variables NaNs nivel barrio</i> .....	37
<i>Imagen 6. Gráfico barras horizontal variables NaNs nivel barrio</i> .....	38
<i>Imagen 7. Método del codo - K-means barrios</i> .....	44
<i>Imagen 8. Método del codo - K-means distritos</i> .....	44
<i>Imagen 9. Distribución barrios K-means k=4</i> .....	45
<i>Imagen 10. Distribución distritos K-means k=4</i> .....	46
<i>Imagen 11. Distribución barrios K-means k=3</i> .....	46
<i>Imagen 12. Distribución distritos K-means k=3</i> .....	47
<i>Imagen 13. Gráfico distancia DBSCAN barrios</i> .....	48
<i>Imagen 14. Gráfico distancia DBSCAN distritos</i> .....	48
<i>Imagen 15. Distribución barrios DBSCAN</i> .....	49
<i>Imagen 16. Distribución distritos DBSCAN</i> .....	49
<i>Imagen 17. Dendrograma barrios</i> .....	51
<i>Imagen 18. Dendrograma distritos</i> .....	51
<i>Imagen 19. Distribución barrios Agglomerative k=4</i> .....	52
<i>Imagen 20. Distribución distritos Agglomerative k=4</i> .....	53
<i>Imagen 21. Distribución barrios Agglomerative k=3</i> .....	53
<i>Imagen 22. Distribución distritos Agglomerative k=3</i> .....	54
<i>Imagen 23. Representación PCA modelos para barrios</i> .....	56
<i>Imagen 24. Representación PCA modelos para distrito</i> .....	56
<i>Imagen 25. Mapa barrios K-means k=3</i> .....	58
<i>Imagen 26. Mapa distritos K-means k=3</i> .....	58
<i>Imagen 27. Mapa barrios K-means k=4</i> .....	59
<i>Imagen 28. Mapa distritos K-means k=4</i> .....	59
<i>Imagen 29. Mapa barrios Agglomerative k=3</i> .....	60
<i>Imagen 30. Mapa distritos Agglomerative k=3</i> .....	60
<i>Imagen 31. Mapa barrios Agglomerative k=4</i> .....	61
<i>Imagen 32. Mapa distritos Agglomerative k=4</i> .....	61
<i>Imagen 33. Mapa barrios DBSCAN</i> .....	62

<i>Imagen 34. Mapa distritos DBSCAN.....</i>	62
<i>Imagen 35. Variables más influyentes en modelo K-means k=3 (barrios).....</i>	69
<i>Imagen 36. Matriz de medias normalizadas modelo K-means k=3 (barrios) .....</i>	70
<i>Imagen 37. Variables más influyentes en modelo Agglomerative k=4 (distritos).....</i>	72
<i>Imagen 38. Matriz de medias normalizadas modelo Agglomerative k=4 (distritos) .....</i>	74
<i>Imagen 39. Mapas modelos finales .....</i>	75

# Lista de Acrónimos

Acrónimo	Significado
SEMMA	<i>Sample, Explore, Modify, Model, Assess</i>
CTP	<i>Cluster-then-Predict</i>
PCA	<i>Principal Component Analysis</i>
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
CRISP-DM	<i>Cross industry Standard Process for Data Mining</i>
ETMF	Equipo de Trabajo con Menores y Familia
CAI	Centros de Atención a la Infancia
SAF	Servicio de Ayuda a domicilio a Menores y Familias
AROPE	<i>At Risk of Poverty or Social Exclusion</i>
UE	Unión Europea
INE	Instituto Nacional de Estadística
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia.y la Cultura
IA	Inteligencia Artificial
SMI	Salario Mínimo Interprofesional
CHS	<i>Calinski-Harabasz Score</i>
GHQ-12	<i>General Health Questionnaire – 12 items</i>
NaN	<i>Not a Number</i>
IGUALA	Índice de Vulnerabilidad Territorial Agregado
FP	Formación Profesional
BUP	Bachillerato Unificado Polivalente
CEPI	Centro de Participación e Integración de Inmigrantes



# 1. INTRODUCCIÓN

---

La desigualdad socioeconómica en los barrios de Madrid ha sido objeto de creciente preocupación en los últimos años. Estudios como el *Panel de Indicadores de Distritos y Barrios de Madrid* del Ayuntamiento de Madrid [1] han puesto en evidencia disparidades significativas en aspectos clave como el acceso a servicios básicos, la renta media y la calidad de vida. Según datos recogidos en el *Atlas de Vulnerabilidad Urbana* [2], los distritos más desfavorecidos presentan no solo menores niveles de ingresos, sino también una menor esperanza de vida y un acceso limitado a infraestructuras esenciales. También se reflejan en una mayor exclusión social en determinados distritos. Estas desigualdades, que se intensificaron con la crisis económica y sanitaria de los últimos años, plantean un desafío urgente para el desarrollo sostenible y equitativo de la ciudad. Este proyecto busca analizar y visualizar las dinámicas de vulnerabilidad en la ciudad mediante la aplicación de técnicas avanzadas de análisis de datos.

Medir y analizar la vulnerabilidad socioeconómica requiere un enfoque multidimensional que permita capturar estas disparidades de una manera integral. Para abordar esto, ha sido fundamental seleccionar indicadores clave que representen adecuadamente las dimensiones de vulnerabilidad. Variables como el nivel educativo, la tasa de desempleo, la densidad poblacional y las condiciones de las viviendas se presentan como elementos críticos en la construcción de un índice que permita clasificar y entender las desigualdades dentro de la ciudad.

El proyecto se centra en identificar y clasificar los distritos de Madrid según su nivel de vulnerabilidad socioeconómica. Para ello, se utilizarán herramientas como el *clustering* y el análisis descriptivo, que permitirán no solo observar las diferencias actuales, sino también establecer relaciones clave entre variables como la educación, la salud y la economía.

## 1.1. JUSTIFICACIÓN

La desigualdad socioeconómica en los distritos de Madrid no es solo un indicador de disparidad, sino un síntoma de una problemática más profunda que afecta a la cohesión social y al desarrollo sostenible de la ciudad. Según los datos proporcionados por el *Panel de Indicadores de Distritos y Barrios de Madrid* [1], ciertas zonas presentan menores ingresos, acceso limitado a servicios básicos y una mayor vulnerabilidad en términos de vivienda y educación. Estas desigualdades perpetúan ciclos de pobreza que no solo afectan a las personas directamente involucradas, sino que también tienen un impacto en el desarrollo general de la ciudad.

Desde una perspectiva global, la medición y el análisis de la vulnerabilidad han sido reconocidos como herramientas esenciales para abordar estas desigualdades. Ejemplos como el *Atlas de Vulnerabilidad humana* [2] y otros índices utilizados en diferentes regiones han demostrado cómo la integración de múltiples variables puede proporcionar una visión más completa y efectiva de las disparidades existentes. Sin embargo, en el contexto de Madrid, resulta evidente la necesidad de desarrollar un índice específico que capture las particularidades locales y facilite la identificación de las áreas más desfavorecidas.

La justificación de este proyecto radica en la urgencia de abordar estas desigualdades con un enfoque basado en datos, que permita diagnosticar la situación actual.

## 1.2. MOTIVACIÓN

La desigualdad socioeconómica en Madrid no es solo una cuestión de cifras y estadísticas, sino una realidad que impacta directamente en la vida de muchas personas. La falta de acceso a oportunidades básicas, como una educación de calidad, refuerza los ciclos de pobreza y dificulta la integración social de numerosos colectivos en situación de vulnerabilidad. Esta problemática se manifiesta de manera especialmente evidente en algunos distritos de la ciudad, donde factores como la renta media, la tasa de desempleo y la calidad de vida condicionan el bienestar y las posibilidades de progreso de sus habitantes.

Durante mi experiencia como voluntaria en el Centro de Participación e Integración de Inmigrantes (CEPI) de San Sebastián de los Reyes, tuve la oportunidad de conocer de cerca esta realidad. A través del apoyo educativo a niños cuyas familias no podían ofrecerles ayuda con sus tareas escolares, pude observar cómo la desigualdad en el acceso a la educación no solo afecta al rendimiento académico, sino también la confianza y la integración social de estos menores. Muchos de ellos enfrentaban dificultades debido a la falta de recursos o a la necesidad de sus familias de priorizar la subsistencia sobre la educación, lo que evidenciaba una brecha significativa entre distintos sectores de la población.

A partir de esta experiencia, surgió la inquietud por entender de manera más profunda los factores que contribuyen a esta desigualdad y cómo se pueden identificar y abordar desde un enfoque basado en datos. Este proyecto busca analizar la vulnerabilidad socioeconómico en Madrid a partir del estudio de diversas variables como el desempleo, la educación, la sanidad y la vivienda, con el objetivo de identificar patrones y correlaciones que permitan comprender mejor la realidad socioeconómica de la ciudad. Se emplearán técnicas de análisis de datos, como la segmentación mediante *clustering*.

## 1.3. ALCANCE

Este proyecto busca analizar la situación socioeconómica de los distritos de Madrid mediante un estudio de indicadores clave como el desempleo, la educación, la sanidad y la vivienda. Su propósito es proporcionar información detallada que permita identificar patrones de

desigualdad y facilite la toma de decisiones informadas para la distribución equitativa de recursos en las zonas más vulnerables.

Para ello, se aplicarán técnicas avanzadas de análisis de datos, incluyendo la segmentación mediante *clustering* para agrupar distritos con características. Este enfoque permitirá detectar zonas con mayores dificultades socioeconómicas y evaluar el impacto de distintos factores en la calidad de vida de los ciudadanos.

Los resultados obtenidos permitirán caracterizar la vulnerabilidad en diferentes áreas de Madrid. Con ello, se espera contribuir a la reducción de desigualdades y al desarrollo de estrategias que mejoren la calidad de vida en los distritos más desfavorecidos.



## 2. INVESTIGACIÓN PREVIA

---

### 2.1. VULNERABILIDAD SOCIOECONÓMICA

La vulnerabilidad socioeconómica es un concepto clave en el estudio de las desigualdades dentro de un territorio. Se define como la propensión de ciertos grupos de población a experimentar dificultades económicas y sociales debido a factores estructurales como el desempleo, el acceso limitado a servicios básicos y la precariedad en la vivienda. Según el *Atlas de Vulnerabilidad Urbana* [2], en Madrid se observan grandes diferencias en la renta media, el acceso a la educación y la sanidad, lo que genera barreras que afectan al desarrollo social y económico de sus habitantes.

Diversos estudios han demostrado que la vulnerabilidad socioeconómica impacta en múltiples dimensiones de la vida de los ciudadanos, influyendo en su estabilidad laboral, educativa y en su bienestar general. Se han desarrollado metodologías para medir esta vulnerabilidad en diferentes ciudades del mundo, como el índice de vulnerabilidad del País Vasco [14], el estudio de Buenos Aires sobre desigualdad urbana [15] y las investigaciones sobre vulnerabilidad social en México [16]. Estos modelos permiten analizar cómo la combinación de factores como el acceso a la educación, la calidad de la vivienda y el desempleo condicionan el desarrollo de una comunidad.

Para evaluar la vulnerabilidad socioeconómica en Madrid, se han seleccionado una serie de indicadores clave basados en estudios previos, como el *Atlas de Vulnerabilidad Urbana* [2] y el *Panel de Indicadores de Distritos y Barrios de Madrid* [1]. La siguiente tabla presenta los principales factores analizados en este estudio, destacando su importancia en la identificación de zonas vulnerables:

Indicador	Fuente	Descripción	Relevancia en el análisis
Renta media	INE [6]	Ingreso anual promedio por hogar en cada distrito	Permite identificar brechas económicas entre zonas
Tasa de desempleo	Ayuntamiento de Madrid [1]	Porcentaje de la población en edad de trabajar sin empleo	Indicador clave de precariedad económica

Acceso a educación	Panel de Hogares de Madrid [4]	Número de centros educativos por cada 1000 habitantes	Refleja desigualdades en oportunidades educativas
Condiciones de vivienda	Atlas de Vulnerabilidad Urbana [2]	Porcentaje de viviendas en condiciones precarias	Relacionado con la calidad de vida y salud.

Tabla 1: Tabla resumen indicadores vulnerabilidad

## 2.2. MÉTODOS DE EVALUACIÓN DE LA VULNERABILIDAD

La evaluación de la vulnerabilidad socioeconómica ha sido abordada a través de múltiples metodologías y modelos de análisis, con el fin de cuantificar y visualizar desigualdades dentro de un territorio. En España, el Ministerio de Transportes ha desarrollado el *Atlas de Vulnerabilidad Urbana* [2], un modelo que integra datos de renta, empleo y acceso a servicios para identificar las áreas más desfavorecidas. De manera similar, en el País Vasco se ha implementado un índice basado en datos censales y encuestas de bienestar social [14], permitiendo un análisis más detallado de la vulnerabilidad en esta región.

A nivel internacional, otras ciudades han desarrollado índices específicos que permiten medir la vulnerabilidad socioeconómica. En Barcelona, por ejemplo, se ha desarrollado un *Índice Integrado de Vulnerabilidad* [11] que mide la precariedad a través de un enfoque multidimensional.

La siguiente tabla resume algunos de los índices de vulnerabilidad urbana utilizados en diferentes países y su metodología:

Índice	Año	País	Método	Indicadores
<i>Urban Vulnerability Index</i>	2008	España	Análisis factorial	Socioeconómicos, residenciales
<i>Integrates Vulnerability Index (IVI)</i>	2022	España	PCA + Clustering	Renta, empleo, acceso a servicios
<i>Deprivation Index</i>	1988	Reino Unido	Transformación logarítmica	Socioeconómicos, residenciales
<i>Comprehensive Urban Vulnerability Index</i>	2010	España	Análisis multicriterio	Sociodemográficos, ambientales

Tabla 2: Tabla resumen índices vulnerabilidad

Index	Author	Year	Area	Method	Indicators
<b>Socioeconomic</b>					
The social status index	Ley	1986	Canada	Correlation model	Socioeconomic; residential (gentrification)
Small-area index of socioeconomic deprivation	Havard	2008	France	Factorial analysis	Sociodemographic; socioeconomic
Deprivation index	Townsend	1988	UK	Logarithmic transformation	Socioeconomic; residential
Socioeconomic level index	Fernández-García	2017	Spain	Factorial analysis	Sociodemographic; socioeconomic; residential
<b>Vulnerability</b>					
Urban vulnerability index	Egea Jiménez	2008	Spain	Factorial analysis	Sociodemographic; socioeconomic; residential; subjective
Comprehensive urban vulnerability synthetic index	Ministry of Development	2010	Spain	Multicriteria analysis	Sociodemographic; socioeconomic; residential; environmental
Indices of deprivation	Ministry of Housing	2011	UK	Factorial analysis	Sociodemographic; socioeconomic; residential; environmental
Synthetic index of comprehensive urban vulnerability	Fernández Aragón	2021b	Spain	Factorial analysis	Sociodemographic; socioeconomic; sociopolitical; residential
Comprehensive vulnerability	Hanoon	2022	Iraq	Fuzzy logic function	Environmental; residential; urban
Composite vulnerability index	Gerundo	2020	Italy	Aggregation method	Sociodemographic; urban; residential

Imagen 1: Tabla índices socioeconómicos y de vulnerabilidad [11]

## 2.3. CLUSTERING EN ANÁLISIS SOCIOECONÓMICO

El *clustering* es una técnica de aprendizaje no supervisado ampliamente utilizada en el análisis socioeconómico para identificar patrones y segmentar territorios en función de sus características compartidas. En el contexto de la vulnerabilidad urbana, esta metodología permite agrupar áreas geográficas con condiciones socioeconómicas similares, facilitando la identificación de zonas de mayor riesgo social y económico.

Uno de los enfoques más utilizados en estudios previos es el *Índice Integrado de Vulnerabilidad* (IVI), desarrollado en Barcelona [11]. Este índice emplea técnicas de análisis de componentes principales (PCA) junto con métodos de *clustering*, como *K-means*, para clasificar barrios en función de factores como el nivel de ingresos, la tasa de desempleo y la accesibilidad a servicios públicos. La combinación de estas técnicas permite generar agrupaciones de barrios con condiciones similares y establecer comparaciones entre diferentes zonas urbanas.

En el contexto internacional, estudios como los desarrollados en Reino Unido y Francia han empleado métodos como la transformación logarítmica y el análisis factorial para la construcción de índices de vulnerabilidad [11]. En España, modelos como el *Atlas de Vulnerabilidad Urbana* [2] y el índice de vulnerabilidad del País Vasco [14] han aplicado metodologías multicriterio para cuantificar la precariedad urbana, destacando la utilidad de estas técnicas en la evaluación de desigualdades territoriales.

En el caso de Madrid, se aplicará *clustering* para analizar la distribución de la vulnerabilidad en los distritos. Se emplearán distintos algoritmos, entre ellos:

- **K-means clustering**: método basado en la minimización de la varianza dentro de los clústeres, ideal para segmentar distritos con características similares.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: útil para identificar agrupaciones de alta densidad en áreas urbanas con características similares, especialmente en casos donde la distribución de los datos no es uniforme.
- **Métodos jerárquicos (Agglomerative)**: permiten establecer relaciones entre zonas urbanas mediante la construcción de árboles de clasificación, facilitando el análisis de la evolución de la vulnerabilidad a lo largo del tiempo.

La combinación de estos métodos en estudios previos ha demostrado ser efectiva para realizar la distribución de la vulnerabilidad urbana y proporcionar información relevante para la toma de decisiones en materia de políticas públicas [7][11]. La adaptación de estas técnicas al caso de Madrid ha permitido evaluar la desigualdad entre distritos y clasificar en función de su nivel de precariedad.

## 2.4. APLICACIÓN DEL CLUSTERING AL ANÁLISIS DE VULNERABILIDAD EN MADRID

En el presente proyecto, el *clustering* se aplicará a los datos del *Panel de Indicadores de Distritos y Barrios de Madrid* [1] con el objetivo de segmentar los distritos en función de su nivel de vulnerabilidad socioeconómica. Para ello, se utilizarán diversas técnicas de agrupamiento con el fin de evaluar cuál proporciona una mejor representación de la realidad social de la ciudad.

El procedimiento se llevará a cabo en las siguientes etapas:

1. **Selección de variables**: se utilizarán indicadores clave como la renta media, la tasa de desempleo, el acceso a servicios de salud y educación, y las condiciones de la vivienda, en línea con estudios previos realizados en Barcelona [11] y el País Vasco [14].
2. **Preprocesamiento de datos**: se normalizarán los datos para garantizar que todas las variables sean comparables y se reducirán dimensiones mediante análisis de componentes principales (PCA), siguiendo la metodología empleada en el *Índice Integrado de Vulnerabilidad* de Barcelona [11].
3. **Aplicación de algoritmos de Clustering**: se implementarán técnicas como:
  - a. **K-means clustering**: para obtener grupos homogéneos de distritos según su nivel de vulnerabilidad.
  - b. **DBSCAN**: para detectar posibles zonas con alta concentración de precariedad.
  - c. **Métodos jerárquicos**: para observar la evolución histórica de la desigualdad en la ciudad.

4. **Evaluación de los resultados:** se aplicarán métricas como el *Silhouette Score* y el método del codo para determinar el número óptimo de clústeres y la cohesión dentro de los grupos.

Con esta metodología se espera proporcionar una clasificación robusta de la vulnerabilidad socioeconómica en Madrid.

## 2.5. METODOLOGÍAS ALTERNATIVAS DE *CLUSTERING* EN CONTEXTO SOCIOECONÓMICO

En el análisis de vulnerabilidad territorial, los métodos de *clustering* han evolucionado hacia enfoques más sofisticados que combinan distintas fases de aprendizaje. Entre ellos, destaca el modelo ***Clúster-Then-Predict*** (CTP), ampliamente empleado en contextos donde es necesario descubrir patrones no supervisados y posteriormente evaluar su relación con variables explicativas. Este enfoque ha demostrado su eficacia en estudios sobre predicción crediticia, análisis de sentimientos y evaluación territorial, como recogen investigaciones recientes [50][51][52].

El modelo CTP consta de dos fases: primero, se agrupan los datos mediante técnicas de *clustering* como *K-means* o DBSCAN, y posteriormente, se aplican modelos supervisados (*Random Forest*, regresión logística, etc.) para analizar la influencia de las variables sobre la pertenencia a cada grupo. Esto convierte un problema no supervisado en uno supervisado, ofreciendo mayor interpretabilidad y precisión.

Además del CTP otras metodologías relevantes incluyen:

- ***Model-Based Clustering*:** agrupan los datos en función de distribuciones probabilísticas subyacentes. Aplicado en contextos sociales permite asumir estructuras latentes en los datos territoriales [53].
- ***Mixture of Experts*:** combina múltiples modelos especializados asignando pesos distintos según la entrada, útil en escenarios con alta heterogeneidad [54].
- ***Two-Step Clustering*:** ampliamente utilizado en investigaciones urbanas por su capacidad de manejar datos mixtos y grandes volúmenes. Se utiliza, por ejemplo, en estudios del INE o SPSS Modeler para identificar conglomerados sociales [55].

Estos modelos aplicados en diversas investigaciones han demostrado que la combinación de fases de segmentación con modelos predictivos incrementa la capacidad del análisis urbano y la toma de decisiones basada en evidencia.

Categoría	Fuente / Título del estudio	Objetivo y relación con el estudio actual
ESTUDIO DE LA VULNERABILIDAD	<i>Atlas de Vulnerabilidad Urbana</i>	Identificar desigualdades territoriales en España mediante renta, empleo y servicios. Inspiración directa para selección de indicadores.
	<i>Índice de Vulnerabilidad del País Vasco</i>	Propuesta regional adaptada a datos censales. Referencia metodológica en análisis territorial español.
	<i>Vulnerabilidad urbana en Buenos Aires</i>	Medición de desigualdad por zonas, transferencia metodológica internacional.
	<i>Vulnerabilidad social en México</i>	Análisis multicriterio aplicado a la precariedad urbana en Latinoamérica.
MÉTODOS DE EVALUACIÓN DE LA VULNERABILIDAD	<i>IVI (Índice Integrado de Vulnerabilidad), Barcelona</i>	Aplicación de PCA y <i>clustering</i> . Influye en el uso combinado de técnicas en este proyecto.
	<i>Deprivation Index, Reino Unido</i>	Histórica metodología de análisis logarítmico para vulnerabilidad, pionera en estudios europeos.
	<i>Comprehensive Urban Vulnerability Index, España</i>	Análisis multicriterio que integra dimensiones sociales, económicas y ambientales.

CLUSTERING EN ANÁLISIS SOCIECONÓMICO	<i>Urban Clustering Methods en Reino Unido y Francia</i>	Aplicación de PCA, análisis factorial y jerárquico. Modelo base replicado en este proyecto.
	<i>Estudio sobre IVI en Barcelona con K-means y PCA</i>	Se replica el enfoque técnico completo en el análisis aplicado a Madrid.
	<i>Panel de Indicadores de Barrios y Distritos de Madrid</i>	Fuente principal de datos de este trabajo.
NUEVAS METODOLOGÍAS APLICADAS	<i>Improved Twitter Sentiment Prediction through Clúster-Then-Predict Model</i>	Demuestra cómo agrupar antes de predecir mejora la precisión de modelos. Aplicable en segmentación territorial.
	<i>A Rescaled Clúster-Then-Predict Approach for Enhanced Credit Scoring</i>	Aplica <i>clustering</i> antes del modelo predictivo para personalizar análisis. Adaptable a entornos sociales.
	<i>Benchmarking Clúster-Then-Predict Models to Challenge Prevailing Global Machine Learning Models</i>	Estudio comparativo que valida la eficacia del enfoque CTP en dominios de segmentación compleja.
	<i>Model-Based Clustering</i>	Análisis probabilístico de agrupamiento. Aplicable en contextos heterogéneos.
	<i>What is Mixture of Experts?</i>	Combinación de modelos según inputs. Útil en datos urbanos diversos.
	<i>TwoStep Clustering</i>	Técnica adaptable al análisis de datos censales mixtos.

Tabla 3: Tabla resumen del Estado del Arte



## 3. OBJETIVOS

---

### 3.1. OBJETIVO GENERAL

Identificar y analizar los factores más determinantes que influyen en la vulnerabilidad de los barrios de Madrid.

### 3.2. LISTA DE OBJETIVOS ESPECÍFICOS

- Realizar un estudio previo sobre el concepto de vulnerabilidad urbana, identificando qué dimensiones suelen considerarse más relevantes (educación, empleo, sanidad, vivienda, etc.).
- Recolectar y estructurar una base de datos que contenga indicadores socioeconómicos de los distintos distritos y barrios de Madrid.
- Aplicar técnicas de análisis y reducción de dimensionalidad con el fin de comprender las relaciones entre variables e identificar redundancias o patrones iniciales.
- Implementar técnicas de agrupamiento (*clustering*) para clasificar los barrios según su grado de vulnerabilidad, permitiendo detectar zonas homogéneas.
- Estudiar la contribución de cada variable elegida al modelo, para determinar qué factores tienen un mayor peso en la clasificación final y entender mejor las causas de la vulnerabilidad detectada.
- Visualizar los resultados facilitando su comprensión y comunicación.

### 3.3. MÉTODOS DE VALIDACIÓN

- ✓ **Validación del modelo:** Se van a utilizar métricas específicas como el *Silhouette Score* o el método del codo. Estas métricas permiten cuantificar cómo de bien se han definido los grupos, es decir, si los barrios agrupados dentro de una misma categoría presentan características similares entre sí y diferentes respecto a los de otros grupos. Además, se va a llevar a cabo una interpretación cualitativa de los clústeres,

comparándolos con las realidades socioeconómicas de la ciudad de Madrid, para comprobar si el agrupamiento es coherente con el conocimiento contextual.

- ✓ **Validación del proyecto:** Una vez construido y evaluado el modelo de análisis, es necesario verificar la veracidad de las conclusiones adquiridas para poder considerar los resultados como válidos. En un escenario ideal, se habría pensado proponer una validación por juicio de expertos, consistente en el análisis detallado del proyecto por parte de profesionales del ámbito social, urbano o de la gestión pública, quienes podrían ofrecer una visión crítica sobre la relevancia de las variables seleccionadas. Sin embargo, debido a la limitación de tiempo y a la imposibilidad de acceso a este tipo de expertos, esta validación externa ha quedado fuera del alcance del presente proyecto.

## 4. PLAN DE DESARROLLO DEL PROYECTO

---

### 4.1. METODOLOGÍA

Para el desarrollo del presente proyecto, centrado en el análisis de la vulnerabilidad socioeconómica de los barrios de Madrid mediante técnicas de minería de datos, se ha optado por la aplicación de la metodología SEMMA, cuyas siglas se corresponden con *Sample, Explore, Modify, Model, Assess*. Esta metodología está diseñada específicamente para proyectos de minería de datos y permite estructurar el proceso de análisis de forma lógica y progresiva, desde la selección y preparación de los datos hasta la construcción y evaluación de los modelos.

SEMMA fue creada por el SAS Institute en 2012 y, en ese sentido, precede a CRISP-DM, otra metodología ampliamente utilizada en el ámbito de la ciencia de datos, en cuyo desarrollo también participó SAS (actualmente parte de IBM). Aunque SEMMA fue concebida como un enfoque orientado específicamente a la minería de datos – considerado el antecedente del *machine learning* –, sus propuestas siguen siendo válidas y aplicables en el contexto actual del análisis de datos y aprendizaje automático [37].

Según la definición del propio *SAS Institute*, *data mining* es el “proceso de muestreo, exploración, modificación, modelado y evaluación de grandes volúmenes de datos para descubrir patrones previamente desconocidos que pueden ser utilizados como ventaja competitiva” [37].

Las cinco fases fundamentales que componen SEMMA son las siguientes:

1. **Muestreo (*Sample*)**: en esta primera etapa se extrae una muestra representativa del conjunto de datos que permita trabajar con una porción significativa de la población sin comprometer la calidad del análisis. En el caso de este proyecto, se ha recopilado información procedente de fuentes oficiales (como el Ayuntamiento de Madrid y el INE), estructurándola por barrios y distritos.
2. **Exploración (*Explore*)**: los datos son analizados con técnicas estadísticas y herramientas de visualización con el fin de detectar tendencias, relaciones, valores atípicos o comportamientos anómalos. Este paso es fundamental para obtener una comprensión profunda de la estructura de los datos antes de avanzar a su transformación.

3. **Modificación (Modify)**: en esta fase se lleva a cabo el tratamiento y transformación de los datos. Incluye tareas como la limpieza de datos, la imputación de valores nulos, la normalización de variables y la selección de las características más relevantes para el análisis.
4. **Modelado (Model)**: se aplican diferentes algoritmos de *clustering* (agrupamiento) para segmentar los barrios en función de su nivel de vulnerabilidad. Se han utilizado modelos como K-Means, DBSCAN y *clustering* jerárquico, evaluando el rendimiento de cada uno para determinar cuál proporciona una clasificación más coherente y útil en el contexto del análisis.
5. **Evaluación (Assess)**: finalmente, se evalúa la calidad de los modelos generados utilizando métricas específicas como el *Silhouette Score*, el método del codo o la visualización mediante técnicas de dimensionalidad (PCA). Esta fase incluye también la interpretación de los resultados y su validación desde una perspectiva ética y contextual.

La elección de SEMMA responde a su eficacia demostrada en proyectos orientados a la exploración y segmentación de datos, así como a su enfoque modular, que permite adaptar cada fase a las necesidades concretas del análisis. Además, su orientación práctica hacia la construcción de modelos lo convierte en un marco adecuado para proyectos que, como este, buscan identificar patrones en grandes volúmenes de información.

No obstante, también se ha considerado la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*), ampliamente reconocida por su estructura iterativa y orientada al negocio. Aunque no ha sido la principal guía en este proyecto, algunos de sus principios, como la comprensión del problema y la validación de resultados desde una perspectiva de aplicación real, han sido incorporados de manera transversal en el desarrollo del trabajo.

## 4.2. TECNOLOGÍAS

Para el desarrollo técnico del proyecto se han empleado principalmente dos herramientas: Python, como lenguaje de programación principal, y Microsoft Excel, como apoyo para el tratamiento preliminar de los datos. El análisis se ha llevado a cabo en el entorno *Google Colaboratory*, que permite trabajar con cuadernos interactivos en la nube, facilitando la ejecución de código, la documentación simultánea y la visualización de resultados. A continuación, se detallan las tecnologías y funciones utilizadas a lo largo del proyecto.

### PYTHON

- **Pandas**: librería fundamental para la manipulación y análisis de datos estructurados. Permite leer archivos Excel, convertir datos a formato numérico, gestionar valores faltantes, realizar transposiciones de tablas y exportar resultados, entre otras cosas.
- **Numpy**: utilizada principalmente para el manejo eficiente de matrices y vectores, soporte numérico y operaciones matemáticas [39].

- **Scikit-Learn**: biblioteca central y gratuita del aprendizaje automático en Python. Permite aplicar modelos de *clustering*, escalado, evaluación y reducción de dimensionalidad [40]. Además, se ha complementado el modelo jerárquico con funciones *scipy* para generar dendogramas explicativos de la estructura de agrupación.
- **Matplotlib.plot**: librería de visualización utilizada para la creación de graficos personalizados y visualizaciones de resultados de modelos y transformaciones.
- **Seaborn**: biblioteca construida sobre matplotlib que facilita la creación de gráficos estadísticos atractivos y con una sintaxis sencilla. Incluye funciones integradas para análisis exploratorio como mapas de calor, gráficas de dispersión y distribuciones [41].
- **Geopandas**: se trata de una extensión de pandas diseñada para facilitar el trabajo con datos geoespaciales. Permite manejar geometrías (puntos, polígonos, líneas) y trabajar con archivos como Shapefiles o GeoJSON [38].
- **Openpyxl**: utilizada indirectamente por pandas para exportar resultados a archivos Excel [42].

## FUNCIONES

- `read_excel()`: función que permite leer archivos Excel (.xlsx) y convertirlos en estructuras de datos tipo *DataFrame* para su análisis y manipulación posterior.
- `set_index()`: establece una columna como índice del *DataFrame*, facilitando la identificación y acceso a los datos por etiquetas.
- `transpose() / .T`: transpone un *DataFrame*, convirtiendo filas en columnas y viceversa, útil para reorganizar la estructura de los datos.
- `to_numeric()`: convierte valores a tipo numérico. En caso de error, puede reemplazar los datos no convertibles por valores nulos.
- `combine_first()`: combina dos Series o *DataFrames* y toma los valores no nulos del primero, usando los del segundo solo si el primero tiene valores faltantes.
- `dropna()`: elimina filas o columnas que contienen valores nulos (*NaN*), según criterios definidos por el usuario.
- `fillna()`: rellena los valores faltantes de un DataFrame con un valor especificado o calculado, como la media o la mediana.
- `mean()`: calcula la media aritmética de una columna o fila de un DataFrame.
- `isnull()`: detecta la presencia de valores nulos (*NaN*) en un DataFrame y devuelve una estructura booleana con True donde existen estos valores.
- `value_counts()`: cuenta la frecuencia de aparición de los valores únicos en una columna o serie.
- `corr()`: calcula la matriz de correlación entre las columnas numéricas de un DataFrame, evaluando relaciones lineales entre variables.
- `to_excel()`: exporta un DataFrame a un archivo Excel (.xlsx) en el disco local.

- `StandardScaler()`: escala los datos eliminando la media y ajustando la varianza a uno. Es útil para que las variables tengan una escala comparable antes de aplicar algoritmos.
- `K-means()`: algoritmo de *clustering* que divide las observaciones en k grupos o clusters, minimizando la distancia a los centroides.
- `DBSCAN()`: algoritmo de agrupamiento basado en densidad. Agrupa observaciones que están juntas en regiones densas y detecta automáticamente outliers.
- `AgglomerativeClustering()`: algoritmo jerárquico de *clustering* que agrupa progresivamente los elementos más similares en una estructura tipo árbol.
- `silhouette_score()`: mide la calidad del *clustering* calculando la cohesión dentro de los clusters y la separación entre ellos. El valor oscila entre -1 y 1.
- `PCA()`: Análisis de Componentes Principales. Técnica de reducción de dimensionalidad que transforma los datos en componentes que explican la mayor varianza posible.
- `Map()`: crea un mapa base interactivo centrado en unas coordenadas geográficas específicas.
- `GeoJson()`: permite superponer geometrías en formato GeoJSON sobre un mapa Folium, asociando propiedades a cada área visualizada.
- `GeoJsonTooltip()`: añade etiquetas emergentes (tooltips) a cada zona del mapa, mostrando propiedades personalizadas al pasar el ratón.
- `Choropleth()`: genera un mapa de coropletas (zonas coloreadas) basado en valores asociados a una variable continua.
- `save()`: guarda el mapa generado en un archivo .html interactivo que puede abrirse en cualquier navegador web.
- `linkage()`: función perteneciente al módulo `scipy.clúster.hierarchy` que construye una matriz de enlaces (linkage matrix) a partir del conjunto de datos y una medida de distancia. Esta matriz describe el orden y la distancia a la que los elementos o grupos se van fusionando en un proceso de *clustering* jerárquico. Admite varios métodos de enlace, como `ward`, `single`, `complete` o `average`, que determinan cómo se calcula la distancia entre los clusters al agruparlos.
- `dendrogram()`: función de visualización que genera un gráfico en forma de árbol (dendograma) a partir de la matriz de enlaces creada con `linkage()`. Cada rama del dendograma representa una fusión entre elementos o grupos, y la altura de la unión indica la distancia o disimilitud entre ellos. Este gráfico es útil para identificar visualmente agrupaciones naturales dentro de los datos y decidir un número adecuado de clusters cortando el árbol en un cierto nivel.

## EXCEL

Microsoft Excel se ha utilizado como herramienta de apoyo en la fase inicial del proyecto para:

- Revisar y corregir manualmente la estructura del archivo (celdas combinadas, variables prescindibles).
- Alinear visualmente los nombres de los distritos y barrios.
- Servir de punto de partida para identificar las variables de interés y clasificarlas según el sector al que mejor pertenecían (salud, educación, vulnerabilidad, socioeconómico).
- Verificar valores atípicos y facilitar la documentación previa al análisis automatizado.

## 4.3. PLAN DE DESARROLLO DEL PROYECTO

En este apartado se detalla el plan de trabajo general del proyecto, alineado con la metodología SEMMA y orientado al análisis y modelado de la vulnerabilidad territorial. El proyecto se ha dividido en distintos paquetes de trabajo (PT), cada uno de ellos con tareas concretas y objetivos específicos.

### 4.3.1. PT 1 Análisis y definición del problema

La primera fase del proyecto ha consistido en una revisión teórica y conceptual que sirviera de base para la parte práctica. Se han investigado los principales enfoques metodológicos utilizados para el análisis de desigualdades territoriales, centrándose especialmente en el contexto urbano. Asimismo, se han estudiado distintas formas de medir la vulnerabilidad, tanto a través de indicadores individuales como mediante índices compuestos. Este análisis previo ha permitido comprender qué variables podían considerarse relevantes en un contexto como el de Madrid y cómo se podrían agrupar las zonas urbanas según características comunes. También se han revisado técnicas estadísticas y de aprendizaje automático, prestando especial atención a métodos de *clustering* y reducción de dimensionalidad, como *K-means* y PCA, que posteriormente se han aplicado sobre los datos reales.

### 4.3.2. PT 2 Creación y limpieza de la base de datos

La base de datos del proyecto se ha obtenido a partir de un conjunto de indicadores publicados por el INE y el Ayuntamiento de Madrid, los cuales ofrecían información detallada a nivel de barrio y distrito sobre múltiples dimensiones sociales, demográficas y económicas. En total, la base original contenía más de 280 variables, muchas de ellas duplicadas en formato porcentual y absoluto, lo que dificultaba su manejo en herramientas de análisis como Python.

Con el fin de facilitar el trabajo y centrarse en los indicadores más relevantes, se ha realizado un proceso de selección manual de las variables más significativas para el estudio. A partir

de esta selección, se ha construido un nuevo archivo de Excel con un formato mucho más estructurado y adaptado al análisis automatizado. En este nuevo archivo se han copiado únicamente las variables previamente seleccionadas organizando la información para que cada zona correspondiera a una fila y cada variable a una columna, respetando el orden lógico de distritos y barrios.

#### 4.3.3. PT 3 Análisis a priori

Una vez creada y estructurada la base de datos definitiva, se ha llevado a cabo un primer análisis exploratorio con el objetivo de comprender mejor la naturaleza de los datos, evaluar su calidad y detectar posibles inconsistencias. Esta fase ha permitido identificar la distribución general de los valores, observar rangos atípicos y detectar correlaciones relevantes entre variables.

Durante este proceso también se han analizado los valores nulos presentes en cada variable, diferenciando entre aquellos que se debían a la falta de información en barrios concretos y los que correspondían a datos disponibles únicamente a nivel de distrito. Este análisis ha permitido tomar decisiones sobre qué variables mantener y cuáles descartar, teniendo en cuenta su completitud y relevancia para el estudio.

Asimismo, esta etapa ha sido fundamental para detectar relaciones iniciales entre los indicadores, facilitando la posterior selección de variables clave para el desarrollo del índice de vulnerabilidad. Gracias a este análisis a priori, ha sido posible establecer una base sólida para los modelos de agrupamiento, mejorando tanto la interpretación como la calidad del análisis final.

Para el desarrollo de este análisis se destinó aproximadamente una semana, en la que se generaron visualizaciones preliminares, se analizaron estadísticas descriptivas y se validó la integridad general del conjunto de datos.

#### 4.3.4. PT 4 Minería de datos

Tras realizar una primera exploración de los datos y observar ciertos patrones en algunas zonas, se inicia la fase de modelado con el objetivo de clasificar los barrios y distritos de Madrid según su nivel de vulnerabilidad. En este proyecto se han utilizado distintos modelos de agrupamiento no supervisado, con el fin de comparar resultados y seleccionar aquel que mejor represente la estructura real de los datos.

En concreto, se han desarrollado tres modelos de *clustering*:

- Algoritmo *K-means*, como modelo principal de referencia.
- Algoritmo DBSCAN, basado en la densidad de los datos.
- *Agglomerative clustering*, que permite formar clústeres de manera progresiva.

Una vez aplicados los modelos, se han comparado sus resultados mediante la métrica *Silhouette Score*, con el objetivo de evaluar la cohesión de los grupos formados y su

separación respecto al resto. Esta comparación ha servido para elegir el modelo más equilibrado y representativo.

El modelo final seleccionado se ha utilizado para interpretar los perfiles de vulnerabilidad de las distintas zonas y ha sido la base para la visualización posterior y la elaboración del índice compuesto.

## 4.4. PLAN DE TRABAJO

Se ha elaborado un diagrama de Gantt que permite visualizar de forma calendarizada el desarrollo del proyecto, estableciendo plazos e hitos clave en cada una de sus fases.

Este cronograma incluye tanto las etapas técnicas como los procesos de documentación, análisis y validación. El siguiente gráfico refleja el plan de trabajo inicial previsto para la ejecución del proyecto.

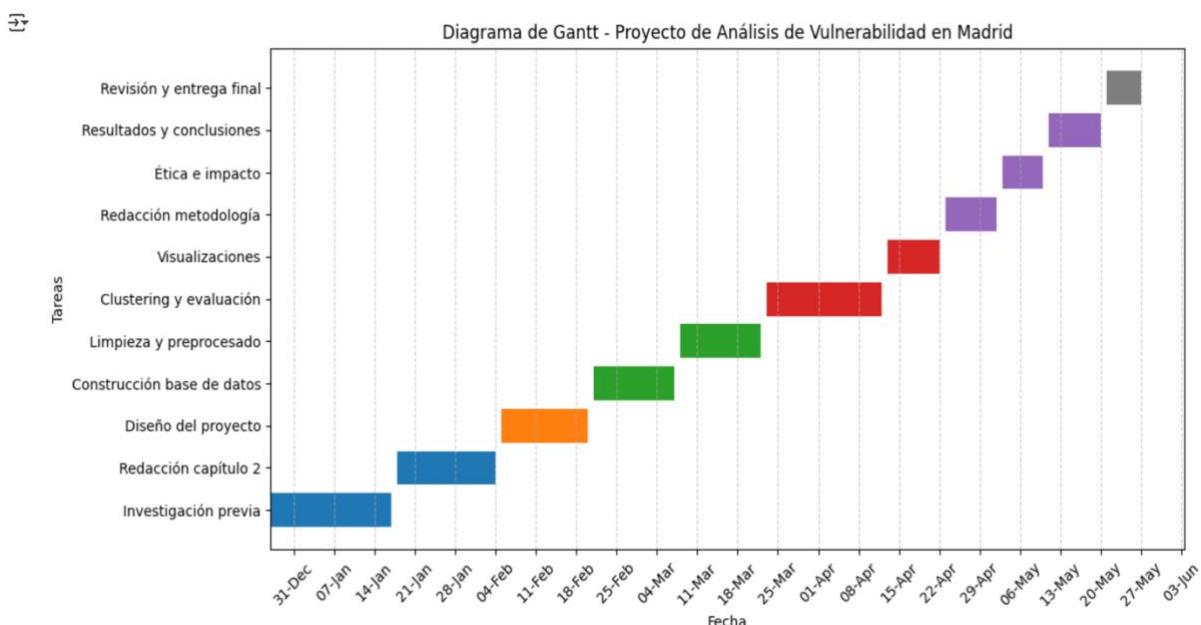


Imagen 2: Diagrama de Gantt

## 4.5. RECURSOS

El desarrollo del proyecto se ha llevado a cabo combinando recursos técnicos y humanos que han permitido la correcta implementación, análisis y validación del estudio.

Desde el punto de vista técnico, se han utilizado dos equipos personales: un portátil HP Pavilion y un iMac, alternando ambos en función de las necesidades del trabajo. El entorno principal de desarrollo ha sido *Google Colaboratory*, que ha permitido programar en Python directamente desde la nube, sin necesidad de configurar entornos locales. La preparación de los datos se ha realizado inicialmente en Microsoft Excel, herramienta clave para la selección y organización de las variables que componen la base de datos. Además, se han utilizado

diversas librerías especializadas en análisis y visualización de datos como *pandas*, *numpy*, *sklearn*, *seaborn*, *matplotlib*, *folium* y *geopandas*.

En cuanto a los recursos humanos, la tutora del Trabajo de Fin de Grado, Natalia, ha acompañado el proyecto proporcionando orientación metodológica y académica. También ha participado en la supervisión Cynthya, como coordinadora del grado, y Luis Moreno Almonacid, responsable de evaluar los aspectos éticos del proyecto. Finalmente, el desarrollo completo del análisis, la selección de variables, la construcción del modelo y la redacción de esta memoria ha sido realizado por Claudia Esnarrizaga, autora del proyecto.

## 4.6. COSTES

El desarrollo de este trabajo de fin de grado no ha supuesto ningún coste económico directo. Todas las herramientas utilizadas han sido de libre acceso, incluyendo Python, *Google Colaboratory* y las bibliotecas necesarias. Tampoco ha sido necesario adquirir licencias o equipos específicos, ya que todo el trabajo se ha realizado con recursos personales y plataformas gratuitas en la nube.

## 4.7. CONDICIONANTES Y LIMITACIONES

Durante el desarrollo del proyecto se han encontrado algunas limitaciones principalmente relacionadas con la disponibilidad y el formato de los datos. La estructura inicial de la base, con variables en filas y zonas en columnas (y muchas columnas combinadas o duplicadas), supuso una dificultad añadida para su tratamiento en Python, lo que obligó a una reestructuración manual laboriosa.

Asimismo, no todas las variables estaban disponibles para cada zona, lo que supuso un reto a la hora de mantener información relevante sin sacrificar calidad estadística. También se ha tenido en cuenta que los modelos de agrupamiento no supervisado no permiten una validación externa directa, por lo que la interpretación de los resultados requiere siempre un juicio crítico y contextualizado.

A pesar de estas limitaciones, se ha conseguido alcanzar los objetivos planteados inicialmente, obteniendo una clasificación clara y justificada de las zonas según su nivel de vulnerabilidad.

## 5. DESARROLLO DE LA SOLUCIÓN TÉCNICA

---

### 5.1. PT 1 – ANÁLISIS Y DEFINICIÓN DEL PROBLEMA

La primera fase del proyecto ha consistido en el análisis conceptual y técnico del fenómeno de la vulnerabilidad territorial, así como en la construcción de la base de datos sobre la que se sustentará todo el análisis posterior. El objetivo final de esta etapa ha sido identificar, seleccionar y estructurar los indicadores más relevantes para medir la vulnerabilidad urbana en Madrid, permitiendo aplicar técnicas de minería de datos orientadas a la creación de un ranking de zonas según su nivel de riesgo o exclusión.

#### 5.1.1. Revisión teórica del concepto de vulnerabilidad urbana

El concepto de vulnerabilidad territorial hace referencia a la exposición diferencial de determinados territorios a riesgos sociales, económicos, medioambientales y sanitarios, condicionada por factores estructurales como la renta, el acceso a servicios, la edad de la población o la calidad del entorno urbano. En el contexto urbano, esta vulnerabilidad tiende a concentrarse espacialmente, generando patrones de desigualdad que requieren intervenciones públicas específicas [58].

En este sentido, el Ayuntamiento de Madrid ha desarrollado el índice IGUALA (Índice de Vulnerabilidad Territorial Agregado), una medida que permite diagnosticar la vulnerabilidad de los diferentes barrios y distritos de la ciudad a partir de indicadores cuantitativos. Este índice se construye en torno a cinco dimensiones clave: Bienestar e Igualdad, Medio Ambiente Urbanismo y Movilidad, Educación y Cultura, Economía y Empleo y Salud, que engloban un total de 44 indicadores procedentes de 10 fuentes de datos oficiales [59].

El proyecto IGUALA ha evaluado más de 200 indicadores para monitorizar los distintos tipos de vulnerabilidad presentes en la ciudad, seleccionando aquellos que cumplen criterios de calidad técnica y operativa. Concretamente, se ha priorizado que los indicadores seleccionados fueran automatizables, actualizables periódicamente, fiables en su fuente, relevantes para la política pública y disponibles a nivel territorial. Además, estos indicadores han sido validados metodológicamente, normalizados para su comparabilidad entre zonas y agrupados en indicadores descriptivos, de acción y mixtos, lo que permite combinar análisis de diagnóstico con capacidades de intervención [59].

Para el desarrollo de este proyecto, se ha tomado como referencia la lógica estructural del índice IGUALA, tanto en la selección de fuentes como en la agrupación temática de variables. Con este objetivo, se ha utilizado el Panel de indicadores de distritos y barrios de Madrid [1], una base de datos en formato Excel (.xlsx) que contiene, en distintas pestañas, los indicadores territoriales de cada uno de los 21 distritos de la ciudad. Cada hoja corresponde a un distrito concreto y presenta los valores para dicho distrito y sus respectivos barrios. La estructura es homogénea en todas las pestañas, con las mismas variables en el mismo orden, lo que ha permitido consolidar la información en una única tabla para el análisis técnico posterior.

Una vez unificada esta base de datos original, se ha llevado a cabo una labor intensiva de limpieza, exploración y filtrado de variables. El objetivo ha sido reducir dimensionalidad y seleccionar únicamente aquellos indicadores considerados más relevantes desde un punto de vista teórico y práctico para el análisis de vulnerabilidad urbana. Este proceso ha culminado en la creación de una nueva tabla estructurada, en la que se ha mantenido un total de 74 variables, partiendo de las 285 que hay en la base de datos original. Estas 74 variables han sido clasificadas temáticamente según las dimensiones explicadas previamente.

### 5.1.2. Nociones básicas y conocimientos previos

En el presente proyecto se ha recurrido a diversas técnicas propias del análisis de datos y el aprendizaje automático con el fin de identificar patrones de vulnerabilidad en el territorio urbano de Madrid. Para comprender el enfoque metodológico adoptado, es importante introducir brevemente los conceptos clave que sustentan el análisis aplicado.

#### **Aprendizaje no supervisado y *clustering***

El análisis parte de técnicas de aprendizaje no supervisado, una rama del *Machine Learning* que permite encontrar estructuras o agrupaciones en los datos sin necesidad de etiquetas previas. En concreto, se ha aplicado la técnica de *clustering*, cuyo objetivo es agrupar observaciones (en este caso, barrios y distritos) que presenten perfiles similares según múltiples indicadores sociales, económicos, educativos, medioambientales y sanitarios.

Se han empleado tres algoritmos diferentes para realizar esta agrupación:

- **K-means**: método basado en centroides que agrupa los datos en K clústeres minimizando la varianza interna de cada grupo. Se ha utilizado el método del codo para determinar el número óptimo de clústeres.
- **DBSCAN**: técnica basada en densidad que permite identificar agrupaciones arbitrarias y detectar puntos atípicos (ruido).
- **Clustering jerárquico (agglomerative)**: técnica jerárquica que fusiona observaciones en función de su similitud, generando una estructura de árbol (dendrograma) que permite visualizar relaciones jerárquicas entre zonas.

## Normalización y escalado

Antes de aplicar estas técnicas, se ha normalizado el conjunto de datos mediante *StandardScales*, una herramienta de scikit-learn que transforma las variables para que todas tengan media cero y desviación estándar uno. Esta normalización garantiza que las variables no dominen el resultado del *clustering* por el hecho de estar en escalas distintas.

Para evitar estos sesgos entre indicadores, se ha aplicado una normalización tipo z-score, calculada como:

$$z = \frac{x - \mu}{\sigma}$$

Donde  $x$  es el valor original,  $\mu$  es la media de la variable y  $\sigma$  su desviación estándar.

## Análisis visual y validación

Una vez generados los clústeres, se ha utilizado una técnica de visualización mediante PCA con dos componentes para proyectar los datos en un plano y facilitar la representación visual de los grupos formados. Cabe destacar que el PCA no se ha utilizado como paso previo para la reducción de dimensionalidad, sino exclusivamente como herramienta de visualización posterior.

Asimismo, se ha empleado el *Silhouette Score* para evaluar la cohesión y separación de los clústeres obtenidos. Este índice toma valores entre -1 y 1, donde los valores más altos indican una mejor separación entre grupos y está definido como:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Donde  $a(i)$  es la distancia media entre  $i$  y los puntos de su mismo clúster, y  $b(i)$  la distancia media al clúster más cercano.

## Interpretación y ranking

A partir de los clústeres obtenidos, se ha generado un ranking de vulnerabilidad mediante el cálculo de la media de los valores estandarizados para cada distrito dentro de cada grupo. Esta puntuación media permite ordenar los clústeres desde los más vulnerables (mayores valores en indicadores críticos) hasta los menos vulnerables.

$$\text{Ranking}(c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \left( \frac{1}{p} \sum_{j=1}^p x_{ij} \right)$$

Además, se ha explorado la importancia relativa de cada variable mediante la varianza entre los centroides de los clústeres. Esta estrategia ha permitido identificar los indicadores que más han contribuido a la diferenciación de los perfiles de vulnerabilidad entre distritos.

Técnicas y herramientas utilizadas			
Técnica / Concepto	Descripción	Herramienta / Librería	Aplicación en el proyecto
<i>Clustering (K-means)</i>	Agrupa observaciones minimizando la varianza intra-clúster	K-means de <code>sklearn.clúster</code>	Agrupar zonas según perfil multidimensional
<i>Clustering (DBSCAN)</i>	Agrupa puntos densamente conectados y detecta ruido	DBSCAN de <code>sklearn.clúster</code>	Detectar agrupaciones naturales y outliers
<i>Clustering jerárquico</i>	Fusión progresiva de observaciones según distancia	<code>AgglomerativeClustering</code>	Generar dendograma y clusters jerárquicos
Normalización (z-score)	Transforma cada variable para tener media 0 y desviación 1	<code>StandardScaler</code> de <code>sklearn</code>	Evitar sesgos por escala entre indicadores
Reducción visual con PCA	Proyección lineal en dos dimensiones principales	PCA de <code>sklearn.decomposition</code>	Visualizar clusters en un plano
<i>Silhouette Score</i>	Mide cohesión y separación de los clusters	<code>Silhouette_score</code> de <code>sklearn.decomposition</code>	Evaluar calidad del <i>clustering</i>
Ranking por media de indicadores	Asigna un orden a los clusters según media de indicadores estandarizados	<code>pandas + media + ordenación</code>	Construir ranking de vulnerabilidad territorial
Importancia por varianza	Evalúa qué variables distinguen más a los clusters	<code>pandas + .var()</code>	Identificar indicadores más influyentes

Tabla 4. Tabla resumen de técnicas y herramientas utilizadas

## 5.2. PT 2 – LIMPIEZA Y CREACIÓN DE LA BASE DE DATOS

### 5.2.1. Selección y clasificación de variables

La identificación y selección de variables relevantes es una fase clave para garantizar la robustez analítica del modelo de vulnerabilidad territorial. A partir del panel original de indicadores del Ayuntamiento de Madrid [1], se han seleccionado un total de 74 variables que permiten medir, comparar y comprender de forma multidimensional las desigualdades sociales y urbanas existentes en los distritos y barrios de la ciudad.

Esta selección ha seguido una doble lógica: por un lado, se ha priorizado la disponibilidad, actualidad y fiabilidad de los datos (indicadores cuantificables, recientes y con cobertura a nivel de barrio); y por otro, se ha buscado una representación equilibrada de las cinco esferas que componen el Índice de Vulnerabilidad Territorial Agregado del Ayuntamiento de Madrid (IGUALA) [59]: Bienestar e Igualdad, Medio Ambiente Urbano y Movilidad, Educación y Cultura, Economía y Empleo, y Salud.

Cada variable ha sido analizada en función de su capacidad explicativa del fenómeno de la vulnerabilidad, asegurando una adecuada cobertura territorial, relevancia social, y potencial para capturar dinámicas estructurales. En este sentido, se han descartado aquellas redundantes, altamente correlacionadas o sin granularidad a nivel barrio. El resultado final es una tabla final depurada y estructurada que recoge los datos normalizados por zona y permite su posterior uso en técnicas de *clustering* y visualización geoespacial.

A continuación, se presenta una clasificación de las variables seleccionadas agrupadas por dimensión, con una tabla resumen para cada una de las cinco esferas. Cada tabla contiene el nombre del indicador, el año de actualización y la fuente original de los datos.

Bienestar Social e Igualdad		
Nombre del indicador	Año de actualización	Fuente
Personas con nacionalidad española	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Personas con nacionalidad extranjera	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Total hogares	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Hogares con una mujer sola mayor de 65 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Hogares con un hombre solo mayor de 65 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Hogares monoparentales: una mujer adulta con uno o más menores	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Tasa de crecimiento demográfico (porcentaje)	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Tasa de riesgo de pobreza o exclusión social	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Tasa de riesgo de pobreza o exclusión social Hombres	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Tasa de riesgo de pobreza o exclusión social Mujeres	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Tasa de riesgo de pobreza o exclusión social ESPAÑOLA (mayor de 16 años)	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Tasa de riesgo de pobreza o exclusión social EXTRANJERA UNIÓN EUROPEA (mayor de 16 años)	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Tasa de riesgo de pobreza o exclusión social EXTRANJERA RESTO DEL MUNDO (mayor de 16 años)	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Personas perceptoras de prestación de la Renta Mínima de Inserción	2023	Ayuntamiento de Madrid   Otras prestaciones y servicios [67]
Familias atendidas por el Equipo de Trabajo con Menores y Familia (ETMF)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Demandas de intervención en los Centros de Atención a la Infancia (CAI)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Menores atendidos por el Servicio de Ayuda a domicilio a Menores y Familias (SAF)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Intervenciones de la Policía Municipal en materia de seguridad: delitos relacionados con las personas	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Intervenciones de la Policía Municipal en materia de seguridad: relacionadas con la tenencia de armas	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Intervenciones de la Policía Municipal en materia de seguridad: relacionadas con el patrimonio	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Intervenciones de la Policía Municipal en materia de seguridad: relacionadas con la tenencia y consumo de drogas	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]

Inspecciones y actuaciones en locales de espectáculos públicos y actividades recreativas	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Expedientes instruidos por Agentes Tutores	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Atestados/partes de accidentes de tráfico confeccionados	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Personas detenidas e investigadas por la Policía Municipal en materia de seguridad: Total personas detenidas e investigadas	2023	Ayuntamiento de Madrid   Datos estadísticos actuaciones Policía Municipal [68]
Porcentaje de envejecimiento (Población mayor de 65 años/Población total)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Porcentaje de sobre-envejecimiento (Población mayor de 80 años/ Población mayor de 65 años)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Edad media de la población	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Índice de dependencia (Población de 0-15 + población 65 años y más / Pob. 16-64)	2024	Ayuntamiento de Madrid   Portal de Estadística [74]
Número de habitantes	2024	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Tabla 5. Tabla de indicadores iniciales de la dimensión de Bienestar Social e Igualdad

En la dimensión de *Bienestar Social e Igualdad*, se ha prestado especial atención a los factores que determinan la vulnerabilidad socioeconómica de diferentes grupos poblacionales. Por ello, se ha incluido un conjunto de indicadores derivados del índice AROPE (*At Risk of Poverty or Social Exclusion*), que permite detectar a la población en situación de pobreza o exclusión social. Este índice, reformulado en 2021 bajo la Estrategia Europa 2030, integra tres dimensiones clave: riesgo de pobreza, carencia material y social severa, y baja intensidad en el empleo. Para una comprensión más precisa de la distribución territorial de esta vulnerabilidad, se ha optado por desagregar este indicador en seis variantes específicas: población general, mujeres, hombres, población española, población extranjera de la Unión Europea, y población extranjera del resto del mundo.

El **indicador general de tasa de riesgo de pobreza o exclusión social** refleja el porcentaje de personas que cumplen al menos una de las tres condiciones mencionadas. La **carenza material y social severa**, uno de sus componentes más recientes, se define por la imposibilidad de cubrir al menos siete de trece necesidades básicas, tanto a nivel del hogar (como irse de vacaciones, mantener una temperatura adecuada en la vivienda o pagar imprevistos) como a nivel individual (como no poder reemplazar muebles estropeados). Este

enfoque permite captar mejor las desigualdades sociales, no solo económicas sino también relacionadas con condiciones de vida dignas.

La desagregación por **sexo** permite analizar si existen diferencias estructurales en la exposición de hombres y mujeres a estas situaciones de vulnerabilidad. De la misma forma, la desagregación por **origen** (españoles, ciudadanos de la UE y personas del resto del mundo) aporta una visión crítica sobre el impacto que pueden tener la nacionalidad o el estatus migratorio en el acceso a recursos y oportunidades. Estas distinciones son fundamentales para diseñar políticas públicas con enfoque interseccional, capaces de atender la diversidad de realidades presentes en la ciudad.

Educación y Cultura		
Nombre del indicador	Año de actualización	Fuente
Población en etapa educativa (Población de 0 a 16 años -16 no incluidos)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 0 a 2 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 3 a 5 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 6 a 11 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 12 a 15 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado en Centros privados concertados	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado en Centros privados sin concierto	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado en Centros públicos	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Total alumnado extranjero	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado extranjero en Centros privados concertados	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado extranjero en Centros privados sin concierto	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado extranjero en Centros públicos	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Total alumnado con necesidades de apoyo educativo	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado con necesidades de apoyo educativo en Centros privados concertados	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Alumnado con necesidades de apoyo educativo en Centros privados sin concierto	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Alumnado con necesidades de apoyo educativo en Centros públicos	2022	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años que no sabe leer ni escribir o sin estudios	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con enseñanza primaria incompleta	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con Bachiller Elemental, Graduado Escolar, ESO, Formación profesional 1º grado	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con Formación profesional 2º grado, Bachiller Superior o BUP	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con titulación media, diplomatura, arquitectura o ingeniería técnica	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con estudios superiores, licenciatura, arquitectura, ingeniería sup., estudios sup. no universitarios, doctorado, postgrado	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Tabla 6. Tabla de indicadores iniciales de la dimensión de Educación y Cultura

Medio Ambiente Urbano y Movilidad		
Nombre del indicador	Año de actualización	Fuente
Superficie deportiva m2	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Centros deportivos municipales	2024	Ayuntamiento de Madrid   portal de datos abiertos   Instalaciones deportivas básicas municipales [66]
Instalaciones deportivas básicas	2024	Ayuntamiento de Madrid   portal de datos abiertos   Instalaciones deportivas básicas municipales [66]

Tabla 7. Tabla de indicadores iniciales de la dimensión de Medio Ambiente Urbano y Movilidad

Economía y Empleo		
Nombre del indicador	Año de actualización	Fuente
Renta disponible media por persona	2021	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Paro registrado (número de personas registradas en SEPE en febrero)	2024	Ayuntamiento de Madrid   Área de Información Estadística   Mercado de trabajo [65]
Tasa de desempleo en hombres de 25 a 44 años	2023	Ayuntamiento de Madrid   Área de Información Estadística   Mercado de trabajo [65]
Tasa de desempleo en mujeres de 25 a 44 años	2023	Ayuntamiento de Madrid   Área de Información Estadística   Mercado de trabajo [65]

Tabla 8. Tabla de indicadores iniciales de la dimensión de Economía y Empleo

Para una mejor comprensión del componente Economía y Empleo, se explican a continuación tres de las variables seleccionadas en este ámbito, ya que representan indicadores clave para evaluar el nivel de vulnerabilidad económica de la población:

- **Tasa de desempleo en mujeres de 25 a 44 años:** este indicador refleja la proporción de mujeres desempleados dentro del grupo de edad de 25 a 44 años en relación con la población activa femenina del mismo tramo.

$$\text{Tasa de desempleo mujeres 25 a 44} = \frac{\text{Mujeres desempleadas de 25 a 44 años}}{\text{Mujeres activas de 25 a 44 años}} \cdot 100$$

- **Tasa de desempleo en hombres de 25 a 44 años:** este indicador refleja la proporción de hombres desempleados dentro del grupo de edad de 25 a 44 años en relación con la población activa masculina del mismo tramo.

$$\text{Tasa de desempleo hombres 25 a 44} = \frac{\text{Hombres desempleados de 25 a 44 años}}{\text{Hombres activos de 25 a 44 años}} \cdot 100$$

- **Renta disponible media por persona:** este indicador se calcula a partir de datos del Padrón y la Agencia Tributaria. Refleja la cantidad media de ingresos netos disponibles por individuo en un área determinada, después de impuestos y otras deducciones.

$$\text{Renta disponible media} = \frac{\sum \text{Rentas netas anuales de todos los individuos}}{\text{Número total de personas}}$$

Salud		
Nombre del indicador	Año de actualización	Fuente
Sedentarismo	2023	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]

Consumo de tabaco diario	2023	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]
Consumo de medicamentos	2023	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]
Personas con sobrepeso	2023	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]
Probabilidad de padecer enfermedad mental (GHQ-12)	2023	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]
Probabilidad de padecer enfermedad mental (GHQ-12) Hombres	2020	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]
Probabilidad de padecer enfermedad (GHQ-12) Mujeres	2020	Ayuntamiento de Madrid   Resumen ejecutivo estudio salud Madrid [63]
Calidad de vida actual en su barrio	2023	Ayuntamiento de Madrid   Informe de resultados interdistrital [64]
Satisfacción de la convivencia vecinal	2023	Ayuntamiento de Madrid   Informe de resultados interdistrital [64]

Tabla 9. Tabla de indicadores iniciales de la dimensión de Salud

Dentro de la dimensión de Salud, se han seleccionado variables que permiten aproximarse a dos aspectos fundamentales del bienestar físico y psicológico de la población: los hábitos de vida relacionados con la actividad física y la salud mental. En concreto, se ha optado por explicar con mayor profundidad dos indicadores: el porcentaje de sedentarismo y la probabilidad de padecer enfermedad mental según el cuestionario GHQ-12.

El **indicador de sedentarismo** mide el porcentaje total de personas cuya actividad física es insuficiente, es decir, aquellas que realizan menos de 30 minutos de ejercicio al día y menos de tres días por semana. Este umbral está alineado con los criterios establecidos por organismos internacionales de salud pública y permite detectar zonas en las que la inactividad física podría tener un impacto relevante en la salud general de la población. La inclusión de esta variable resulta clave en contextos urbanos donde las desigualdades en el acceso a instalaciones deportivas o espacios públicos pueden agravar estas condiciones.

Por otro lado, el indicador de **probabilidad de padecer enfermedad mental** se basa en la aplicación del cuestionario GHQ-12 (*General Health Questionnaire*), una herramienta estandarizada utilizada para detectar riesgos de malestar psicológico en contextos no clínicos. Este instrumento evalúa la presencia de síntomas como ansiedad, estrés o depresión leve, y proporciona una puntuación que se traduce en una estimación porcentual del riesgo de padecer algún tipo de trastorno emocional. Se ha tomado como referencia la aplicación realizada en abril de 2020, coincidiendo con un momento especialmente crítico a nivel social y emocional tras el inicio de la pandemia de COVID-19. Su inclusión en el análisis responde a

la necesidad de visibilizar la dimensión emocional de la vulnerabilidad, a menudo menos tangible pero igualmente determinante en el bienestar global de los territorios.

Vulnerabilidad Territorial (IGUALA)		
Nombre del indicador	Año de actualización	Fuente
Índice de Vulnerabilidad Territorial Agregado	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Bienestar Social e Igualdad	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Medio Ambiente Urbano y Movilidad	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Educación y Cultura	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Economía y Empleo	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Salud	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]

Tabla 10. Tabla de indicadores iniciales de la dimensión de Vulnerabilidad Territorial (IGUALA)

Los seis indicadores anteriores constituyen variables agregadas diseñadas para sintetizar grandes volúmenes de datos en esferas clave del bienestar urbano. En particular:

- **Índice de Vulnerabilidad Territorial Agregado:** resume la vulnerabilidad global de cada unidad territorial de Madrid, integrando 40 indicadores agrupados en cinco esferas temáticas fundamentales.
- **Índice de Vulnerabilidad Bienestar Social e Igualdad:** se centra en aspectos relacionados con la cohesión social, servicios sociales, seguridad, integración y desigualdad.
- **Índice de Vulnerabilidad Medio Ambiente Urbano y Movilidad:** aborda condiciones urbanas, medioambientales y de acceso a transporte y servicios urbanos.
- **Índice de Vulnerabilidad Educación y Cultura:** incluye variables sobre nivel educativo, acceso a centros formativos, y recursos culturales, claves para el desarrollo y movilidad social.
- **Índice de Vulnerabilidad Economía y Empleo:** recoge datos sobre renta, desempleo, inserción laboral y condiciones económicas generales del territorio.
- **Índice de Vulnerabilidad Salud:** evalúa las condiciones sanitarias y factores relacionados con el acceso y calidad del sistema de salud en la zona.

### 5.2.2. Arquitectura de la base de datos

La base de datos utilizada en este análisis ha sido construida a partir de una reorganización profunda del *Panel de indicadores de distritos y barrios de Madrid 2024* [1]. En su formato original, este panel se presenta como un archivo Excel con 21 pestañas, cada una correspondiente a un distrito, y en cada hoja se incluyen los valores de más de 280 indicadores desglosados por barrios. Esta estructura, aunque adecuada para la consulta visual puntual, dificulta el análisis comparativo entre zonas y la automatización de procesos analíticos.

Por ello, se ha llevado a cabo una transformación completa hacia un modelo tabular vertical, en el que las **variables (indicadores)** ocupan las filas y las columnas corresponden a las **unidades territoriales** (los 21 distritos y sus respectivos barrios). El resultado es una base de datos unificada en la que cada fila representa una observación única por variable, y cada celda contiene el valor de dicha variable para una zona concreta.

Desde el punto de vista de arquitectura de datos, este modelo responde a una estructura lógica de tipo *data warehouse* simplificado, en el que los datos han sido integrados desde múltiples hojas de cálculo hacia una única fuente tabular estandarizada, con el objetivo de facilitar análisis multidimensionales, transformaciones estadísticas y visualizaciones avanzadas.

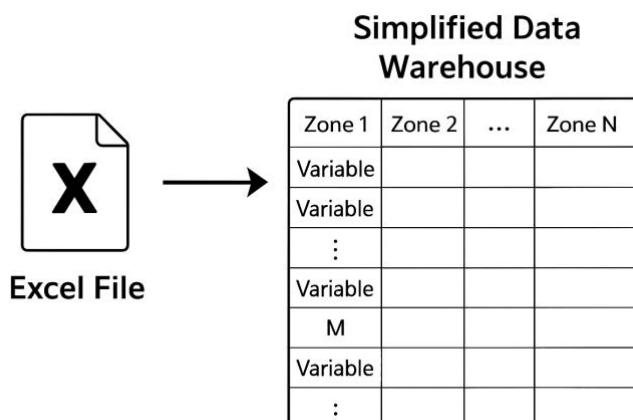


Imagen 3. Reorganización a estructura tipo data warehouse

Esta aproximación sigue los principios fundamentales definidos por IBM en relación con la arquitectura de datos [69], y más concretamente con las funciones propias de un *data warehouse*, entendido como una estructura diseñada para consolidar información desde múltiples fuentes con vistas al análisis estratégico y a la toma de decisiones [70].

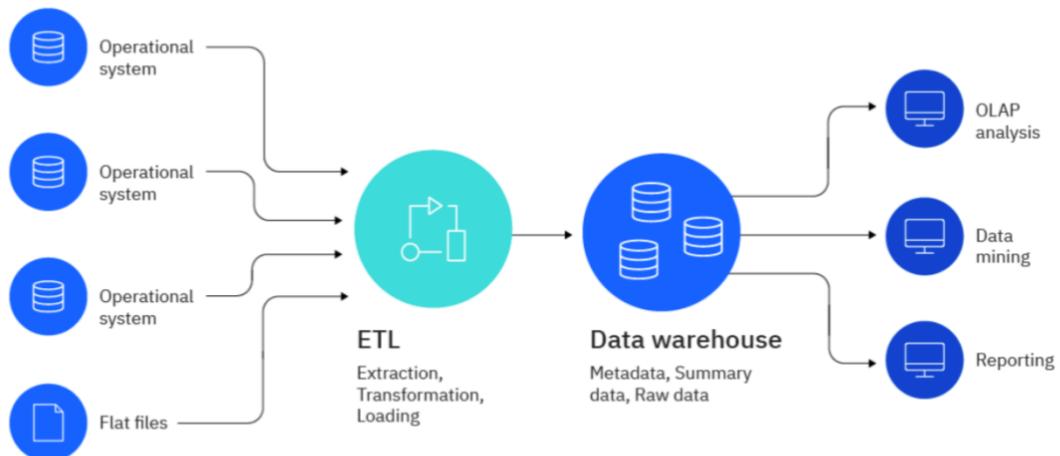


Imagen 4. Esquema de arquitectura de un data warehouse [70]

Este modelo centralizado y estructurado permite:

- **Gobierno de datos:** al consolidar todas las variables en una sola tabla, se facilita el control, trazabilidad y documentación de los datos empleados.
- **Accesibilidad y reutilización:** la estructura en formato *wide* (zonas por columnas y variables por filas) permite utilizar el *dataset* en herramientas de análisis como Python o R sin necesidad de transformaciones adicionales.
- **Escalabilidad:** la base de datos está diseñada para incorporar nuevas variables o zonas sin alterar la lógica estructural del modelo.
- **Consistencia y estandarización:** todas las variables han sido homogenizadas en cuanto a unidades, escala y formato, lo que facilita la normalización y el análisis comparativo posterior.

A nivel operativo, esta arquitectura ha permitido mantener una capa base sólida y fiable, sobre la cual se han aplicado técnicas de limpieza, normalización y minería de datos. Esta transformación inicial ha sido determinante para asegurar la calidad de los análisis posteriores de *clustering* y vulnerabilidad territorial.

## 5.3. ANÁLISIS A PRIORI

Tras la unificación de la base de datos en un único archivo tabular, el siguiente paso ha consistido en su preparación para un primer análisis. Este proceso ha incluido la distinción estructural entre distritos y barrios, el filtrado de variables según criterios de completitud, y la normalización estadística del conjunto, todo ello a garantizar la calidad y robustez de los modelos de *clustering* posteriores.

### 5.3.1. Clasificación territorial: diferenciación entre barrios y distritos

Una peculiaridad del Panel de Indicadores del Ayuntamiento de Madrid [1] es que combina en sus pestañas tanto valores agregados por distrito como desgloses por barrio. Sin embargo, al transponer la base de datos a formato tabular vertical, se pierde la referencia jerárquica directa entre ambos niveles. Por este motivo, se ha diseñado una rutina de identificación que permite distinguir automáticamente las filas correspondientes a distritos (nombres escritos completamente en mayúsculas) de las que representan barrios. A cada unidad se le ha asignado además un campo adicional (*es\_barrio*) que permite filtrar fácilmente los análisis por nivel territorial.

Esta distinción resulta especialmente relevante, ya que la disponibilidad de información no es homogénea entre ambos niveles. Mientras que los datos a nivel distrito tienden a estar más consolidados y sistematizados, la información desagregada por barrio presenta una mayor proporción de valores nulos, especialmente en indicadores sensibles como los policiales o sanitarios. En las imágenes 5 y 6 se muestran ejemplos de variables que están completas en distritos, pero no en barrios, lo que pone de manifiesto una limitación estructural en la cobertura de información a nivel micro territorial.

Variables con muchos NaNs en barrios pero no en distritos:		
VARIABLES ESCOGIDAS/ZONA	NanNs en barrios (%)	NanNs en distritos (%)
Consumo de medicamentos	100.0	0.0
Personas con sobrepeso	100.0	0.0
Consumo de tabaco diario	100.0	0.0
Alumnado en Centros privados concertados	100.0	0.0
Alumnado en Centros públicos	100.0	0.0
Alumnado en Centros privados sin concierto	100.0	0.0
Alumnado extranjero en Centros privados concertados	100.0	0.0
Total alumnado extranjero	100.0	0.0
Renta disponible media por persona	100.0	0.0
Intervenciones de la Policía Municipal en materia de seguridad: relacionadas con la tenencia y consumo de drogas	100.0	0.0
Inspecciones y actuaciones en locales de espectáculos públicos y actividades recreativas	100.0	0.0
Expedientes instruidos por Agentes Tutores	100.0	0.0
Atestados/partes de accidentes de tráfico confeccionados	100.0	0.0
Personas detenidas e investigadas por la Policía Municipal en materia de seguridad: Total personas detenidas e investigadas	100.0	0.0
Intervenciones de la Policía Municipal en materia de seguridad: delitos relacionados con las personas	100.0	0.0
Intervenciones de la Policía Municipal en materia de seguridad: relacionadas con la tenencia de armas	100.0	0.0
Intervenciones de la Policía Municipal en materia de seguridad: relacionadas con el patrimonio	100.0	0.0
Centros deportivos municipales	100.0	0.0
Instalaciones deportivas básicas	100.0	0.0
Alumnado con necesidades de apoyo educativo en Centros públicos	100.0	0.0
Alumnado con necesidades de apoyo educativo en Centros privados sin concierto	100.0	0.0
Alumnado con necesidades de apoyo educativo en Centros privados concertados	100.0	0.0
Total alumnado con necesidades de apoyo educativo	100.0	0.0
Alumnado extranjero en Centros públicos	100.0	0.0
Alumnado extranjero en Centros privados sin concierto	100.0	0.0
Sedentarismo	100.0	0.0
Personas perceptoras de prestación de la Renta Mínima de Inserción	100.0	0.0
Familias atendidas por el Equipo de Trabajo con Menores y Familia (ETMF)	100.0	0.0
Demandas de Intervención en los Centros de Atención a la Infancia (CAI)	100.0	0.0
Menores atendidos por el Servicio de Ayuda a domicilio a Menores y Familias (SAF)	100.0	0.0
Índice de Vulnerabilidad Territorial Agregado	100.0	0.0
Probabilidad de padecer enfermedad (GHQ-12) Mujeres	100.0	0.0
Probabilidad de padecer enfermedad mental (GHQ-12) Hombres	100.0	0.0
Probabilidad de padecer enfermedad mental (GHQ-12)	100.0	0.0
Superficie deportiva m2	100.0	0.0

Imagen 5. Tabla variables NaNs nivel barrio

Indicadores

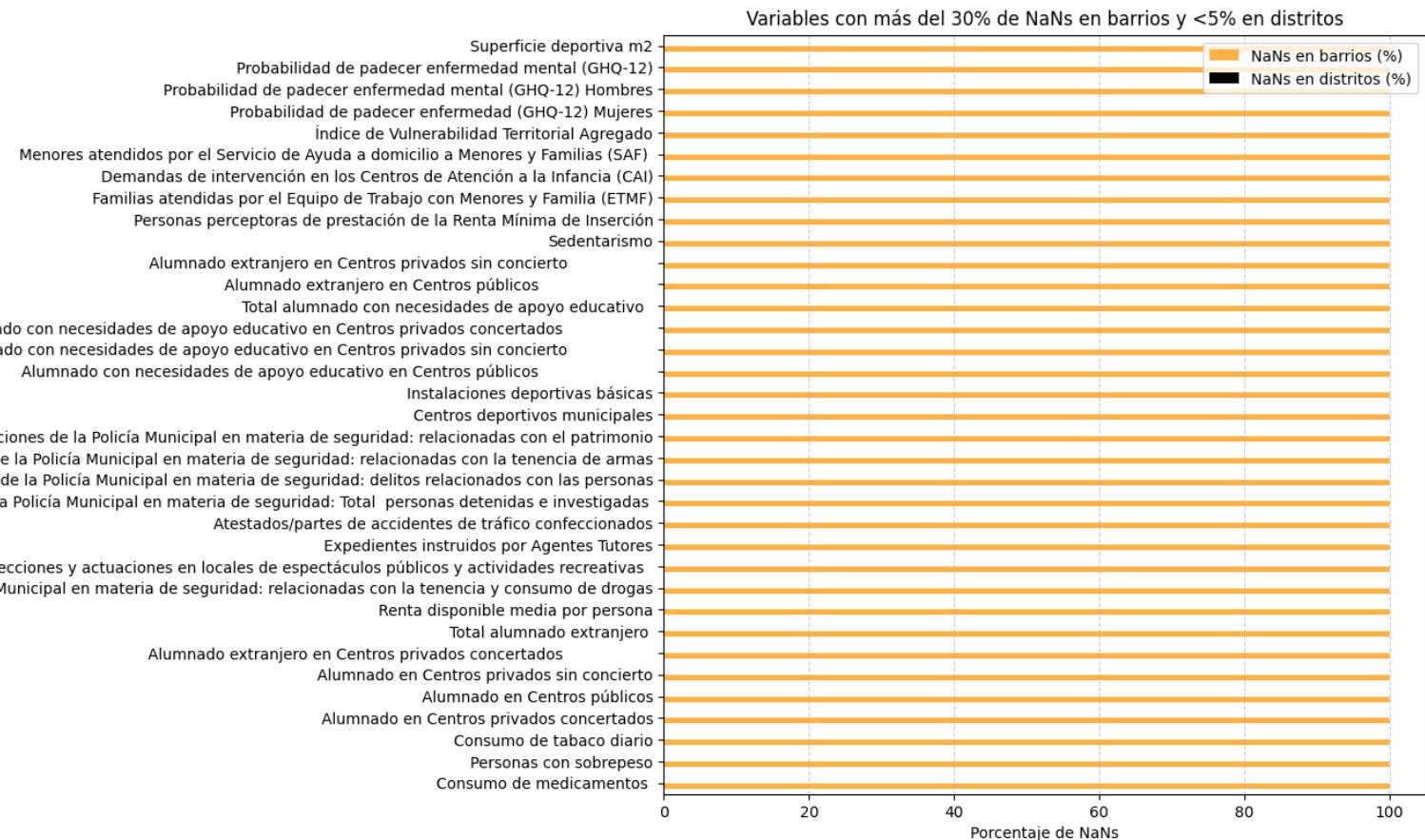


Imagen 6. Gráfico barras horizontales variables NaNs nivel barrio

### 5.3.2. Filtrado de variables según completitud

Dado que la calidad de los modelos de *clustering* puede verse fuertemente afectada por la presencia de valores nulos, se ha establecido un umbral mínimo de completitud para mantener una variable en el análisis. Concretamente, se han conservado aquellas columnas con al menos un 70% de los datos no nulos, criterio habitual en la literatura para evitar distorsiones en el escalado o en la distancia euclídea utilizada por muchos algoritmos de agrupamiento [71].

De forma complementaria, se ha aplicado un criterio de conservación forzosa sobre determinados indicadores clave. En particular, se han mantenido las seis variables compuestas que conforman el Índice de Vulnerabilidad Territorial Agregado (IGUALA), además de la edad media de la población, por considerarse relevantes a efectos comparativos e interpretativos, a pesar de que algunas de ellas no superaban el umbral de completitud en barrios.

Esta decisión metodológica implica que, aunque ciertas variables sí presentaban valores válidos a nivel distritos, no han sido incluidas en el análisis por barrios al no disponer de información suficiente. Como muestran en las imágenes 5 y 6, un conjunto significativo de indicadores – relacionados principalmente con salud, educación o intervenciones policiales – muestra un 100% de valores nulos en barrios, pero no así en distritos, lo que evidencia una brecha de cobertura entre niveles. No obstante, con el fin de mantener la coherencia metodológica entre ambos análisis y facilitar la comparación de resultados, se ha optado por emplear un conjunto común de variables tanto para barrios como para distritos, descartando aquellas que no alcanzaban el umbral de completitud en alguno de los dos niveles.

### 5.3.3. Eliminación de observaciones incompletas y reconstrucción del *dataset*

Una vez filtradas las variables, se ha procedido a eliminar también aquellas observaciones (filas) que no alcanzaban un umbral mínimo de completitud, también fijado en el 70%. Esta operación ha reducido el número total de zonas a analizar en el caso de los barrios, pero ha mejorado de forma significativa la estabilidad de los modelos posteriores, evitando distorsiones derivadas de la imputación arbitraria o del sesgo de distribución.

No obstante, al tratarse de una transformación potencialmente excluyente, se han conservado copias paralelas del conjunto de datos original para realizar análisis complementarios en niveles más agregados, en caso de ser necesario. Esta decisión ha permitido preservar la flexibilidad analítica del modelo, facilitando futuras comparaciones o extensiones del estudio.

A continuación, se ha reintegrado la información territorial original al nuevo *DataFrame* depurado, de modo que cada observación conserva tanto su valor numérico normalizado como su identificación territorial. Esto asegura la trazabilidad de cada fila y permite representar los resultados posteriores mediante mapas o gráficos categorizados.

### 5.3.4. Normalización y estandarización de variables

El conjunto de datos resultante ha sido normalizado mediante la técnica z-score, que transforma cada variable para que presente media cero y desviación típica uno. Esta transformación permite eliminar el sesgo derivado de las diferentes escalas de medición entre indicadores (por ejemplo, población absoluta frente a porcentajes) y es especialmente necesaria en algoritmos de *clustering* basados en distancia, como *K-means* o *Agglomerative Clustering* [72][73].

La fórmula aplicada para cada valor  $x$  ha sido:

$$z = \frac{x - \mu}{\sigma}$$

Donde  $\mu$  representa la media de la variable y  $\sigma$  su desviación estándar. Este procedimiento ha sido implementado mediante la clase *StandardScaler* de la librería scikit-learn [74].

### 5.3.5. Variables seleccionadas para el análisis

Tras el proceso de filtrado y normalización, se ha mantenido un total de 30 variables, que representan un conjunto robusto y equilibrado de dimensiones demográficas, educativas, socioeconómicas y de vulnerabilidad. Estas son:

Bienestar Social e Igualdad		
Nombre del indicador	Año de actualización	Fuente
Personas con nacionalidad española	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Personas con nacionalidad extranjera	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Total hogares	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Hogares con una mujer sola mayor de 65 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Hogares con un hombre solo mayor de 65 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Hogares monoparentales: una mujer adulta con uno o más menores	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Porcentaje de envejecimiento (Población mayor de 65 años/Población total)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Porcentaje de sobre-envejecimiento (Población mayor de 80 años/ Población mayor de 65 años)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Edad media de la población	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Índice de dependencia (Población de 0-15 + población 65 años y más / Pob. 16-64)	2024	Ayuntamiento de Madrid   Portal de Estadística [74]
Número de habitantes	2024	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Tabla 11. Tabla de indicadores finales de la dimensión de Bienestar Social e Igualdad

Educación y Cultura		
Nombre del indicador	Año de actualización	Fuente
Población en etapa educativa (Población de 0 a 16 años -16 no incluidos)	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 0 a 2 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 3 a 5 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 6 a 11 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población en etapa educativa de 12 a 15 años	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con enseñanza primaria incompleta	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con Bachiller Elemental, Graduado Escolar, ESO, Formación profesional 1º grado	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con Formación profesional 2º grado, Bachiller Superior o BUP	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con titulación media, diplomatura, arquitectura o ingeniería técnica	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]
Población mayor/igual de 25 años con estudios superiores, licenciatura, arquitectura, ingeniería sup., estudios sup. no universitarios, doctorado, postgrado	2023	Ayuntamiento de Madrid   Banco de datos de Madrid [61]

Tabla 12. Tabla de indicadores finales de la dimensión de Educación y Cultura

Economía y Empleo		
Nombre del indicador	Año de actualización	Fuente
Paro registrado (número de personas registradas en SEPE en febrero)	2024	Ayuntamiento de Madrid   Área de Información Estadística   Mercado de trabajo [65]
Tasa de desempleo en hombres de 25 a 44 años	2023	Ayuntamiento de Madrid   Área de Información Estadística   Mercado de trabajo [65]
Tasa de desempleo en mujeres de 25 a 44 años	2023	Ayuntamiento de Madrid   Área de Información Estadística   Mercado de trabajo [65]

Tabla 13. Tabla de indicadores finales de la dimensión de Economía y Empleo

Vulnerabilidad Territorial (IGUALA)		
Nombre del indicador	Año de actualización	Fuente
Índice de Vulnerabilidad Territorial Agregado	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Bienestar Social e Igualdad	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Medio Ambiente Urbano y Movilidad	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Educación y Cultura	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Economía y Empleo	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]
Índice de Vulnerabilidad Salud	2022	Ayuntamiento de Madrid   Distritos y barrios   Índice IGUALA [62]

Tabla 14. Tabla de indicadores finales de la dimensión de Vulnerabilidad Territorial (IGUALA)

Estas variables conforman el conjunto final con el que se llevarán a cabo los análisis de agrupamiento, permitiendo una caracterización multivariable de los barrios y distritos de Madrid en términos de vulnerabilidad.

## 5.4. APLICACIÓN DE TÉCNICAS DE CLUSTERING

Una vez estructurado, filtrado y normalizado el conjunto de datos, se ha procedido a aplicar distintas técnicas de agrupamiento no supervisado (*clustering*) con el objetivo de identificar patrones territoriales de vulnerabilidad en la ciudad de Madrid. Este análisis se ha realizado de forma diferenciada para barrios y distritos, manteniendo constante el conjunto de variables en ambos casos para asegurar la comparabilidad de los resultados.

Para ambos niveles territoriales se han aplicado tres algoritmos clásicos de *clustering*: *K-means*, DBSCAN y *Agglomerative Clustering*. Cada uno de estos métodos presenta una lógica distinta de agrupación, tal y como se explica en el punto 5.1.2. *Nociones básicas y conocimientos previos*.

Estos algoritmos se han ejecutado por separado para distritos y barrios, ya que las diferencias en número de observaciones, calidad de datos y granularidad territorial justifican la ejecución de análisis paralelos.

### 5.4.1. *K-means*

El primer algoritmo de *clustering* aplicado ha sido *K-means*, ampliamente utilizado en análisis multivariantes por su capacidad para generar particiones compactas y fácilmente interpretables [75]. Esta técnica busca particionar el conjunto de observaciones en  $k$  grupos o clústeres, de forma que se minimice la variabilidad interna de cada grupo. El criterio de optimización se basa en reducir la suma de distancias cuadradas entre los puntos y el centroide de su correspondiente clúster.

Matemáticamente, el objetivo de *K-means* consiste en minimizar la función de inercia  $W_k$ , definida como:

$$W_k = \sum_{r=1}^k \sum_{x_i \in C_r} \|x_i - \mu_r\|^2$$

Donde  $x_i$  representa cada observación,  $\mu_r$  es el centro del clúster  $r$ , y la suma total agrega las distancias cuadradas de todos los puntos a sus respectivos centroides. Esta es precisamente la función de inercia que calcula el atributo `.inertia_` de la implementación de scikit-learn, utilizada en el presente análisis [76].

#### Determinación del número óptimo de clústeres (k)

Para la elección del parámetro  $k$ , se ha aplicado el denominado método del codo, que consiste en representar gráficamente la evolución de la inercia  $W_k$  en función del número de clústeres, observando el punto a partir del cual las reducciones sucesivas de la inercia son marginales. Este punto de inflexión sugiere el número óptimo de agrupamientos, ya que añadir nuevos clústeres no mejor sustancialmente la cohesión interna [77].

Aunque existen variantes alternativas de cálculo del índice  $W_k$ , como la expresada por López-Solís et al. (2023) [78], que normaliza la dispersión intra-clúster dividiendo la suma de distancias por el tamaño de cada clúster:

$$W_k = \sum_{r=1}^k \frac{1}{n_r} D_r$$

Donde  $n_r$  es el número de observaciones en el clúster  $r$  y  $D_r$  corresponde a la suma de distancias dentro del clúster, en este proyecto se ha seguido el criterio clásico de inercia total de scikit-learn (suma de cuadrados no normalizada), ampliamente aceptado en problemas de segmentación multivariable [76],

Se ha probado con valores de  $k$  entre 2 y 10, observando la gráfica de la suma de errores cuadráticos.

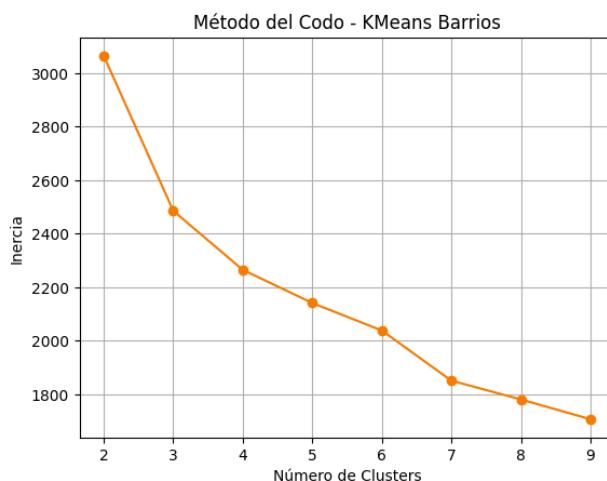


Imagen 7. Método del codo - K-means barrios

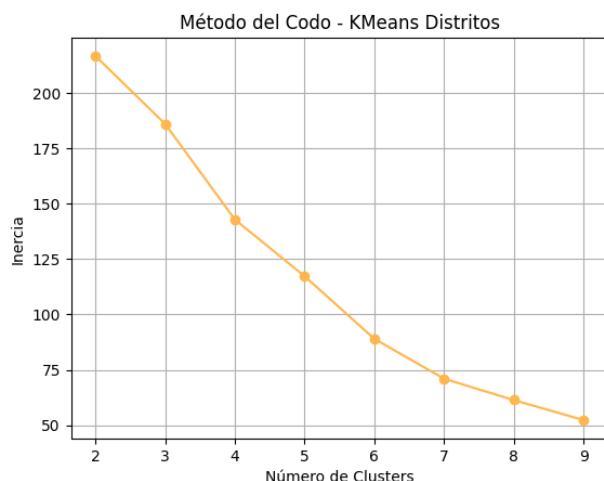


Imagen 8. Método del codo - K-means distritos

Tanto en barrios como distritos, el codo de la curva sugiere dos posibles valores adecuados:  $k = 3$  y  $k = 4$ . Por este motivo, se han generado modelos con ambos valores, comparando

su distribución y estructura para evaluar su adecuación según la cohesión interna de los grupos y su capacidad explicativa.

### Resultados para k = 4

En primer lugar, se ha ejecutado el modelo *K-means* para  $k = 4$  tanto en el análisis de distritos como en el de barrios. Los resultados muestran una distribución relativamente equilibrada de observaciones entre los diferentes grupos, aunque con ciertas asimetrías.

En el caso de los barrios, la partición ha quedado distribuida de la siguiente forma:

Nº clúster	Nº barrios
0	39
1	55
2	20
3	10

Tabla 15. Nº barrios por clúster (*K-means* k=4)

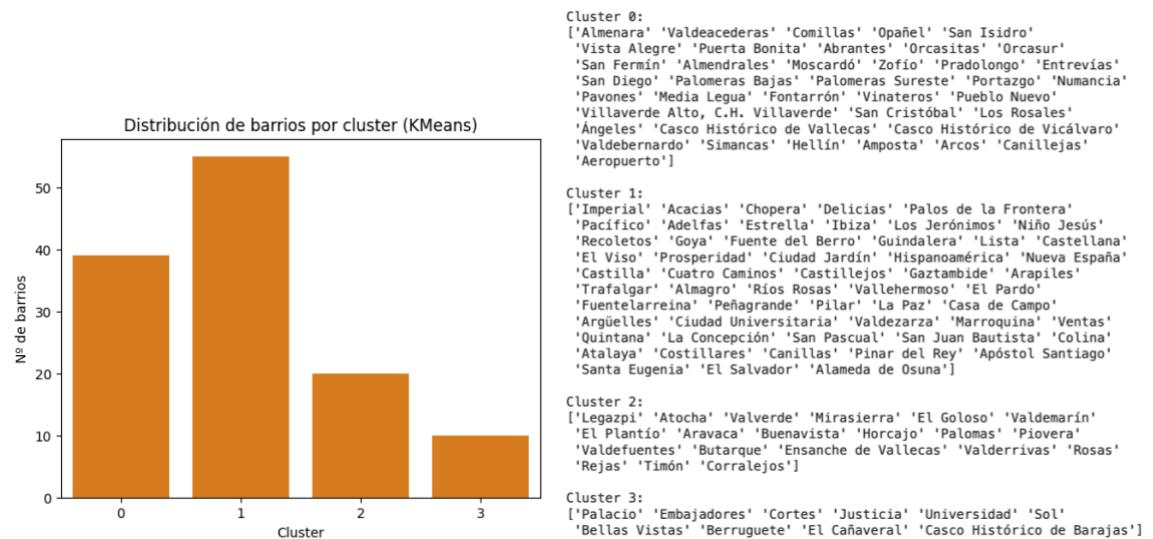


Imagen 9. Distribución barrios *K-means* k=4

En el caso de los distritos, la partición ha quedado distribuida de la siguiente forma:

Nº clúster	Nº distritos
0	4
1	7
2	8
3	2

Tabla 16. Nº distritos por clúster (*K-means* k=4)

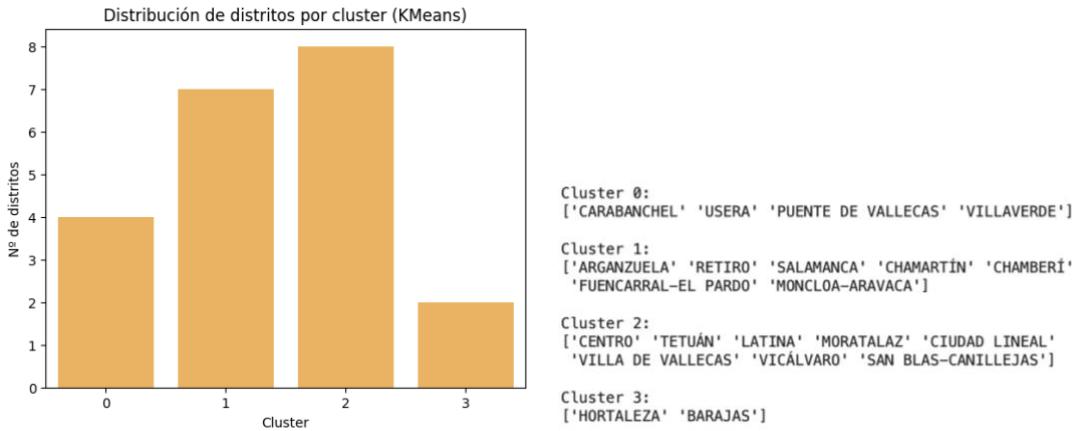


Imagen 10. Distribución distritos K-means k=4

### Resultados para k = 3

Como alternativa, se ha evaluado igualmente el modelo con  $k = 3$  para contrastar la robustez de las particiones y analizar si la segmentación adicional permite capturar matices relevantes en los patrones de vulnerabilidad.

En el caso de los barrios, la partición ha generado:

Nº clúster	Nº barrios
0	42
1	62
2	20

Tabla 17. Nº barrios por clúster (K-means k=3)



Imagen 11. Distribución barrios K-means k=3

En el caso de los distritos, el modelo ha agrupado las 21 unidades territoriales en:

Nº clúster	Nº distritos
0	4
1	10
2	7

Tabla 18. Nº distritos por clúster (K-means k=3)

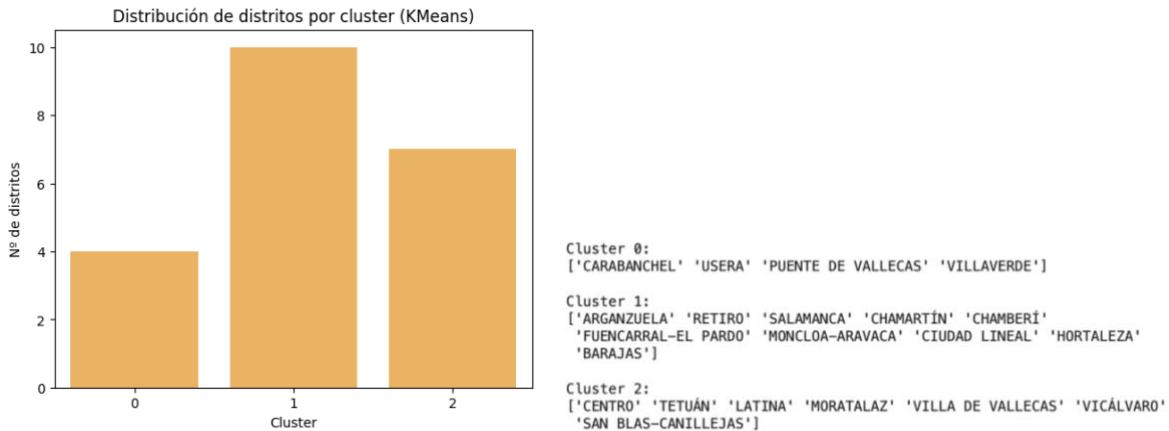


Imagen 12. Distribución distritos K-means k=3

#### 5.4.2. DBSCAN

El segundo algoritmo implementado ha sido DBSCAN, una técnica basada en densidad especialmente adecuada para identificar agrupaciones de forma arbitraria y detectar observaciones atípicas o ruido [79]. A diferencia de *K-means*, DBSCAN no requiere especificar a priori el número de clústeres, sino que forma agrupamientos a partir de la concentración espacial de los puntos según parámetros de densidad.

##### Funcionamiento

DBSCAN agrupa observaciones que cumplen dos condiciones simultáneas:

- Que se encuentren dentro de un radio máximo de vecindad (`eps`).
- Que haya al menos un número mínimo de puntos (parámetro `min_samples`) en dicha vecindad.

Los puntos que no cumplen estos criterios son etiquetados como ruido (valor -1).

##### Ajuste de parámetros

La selección del valor óptimo de `eps` es determinante para el comportamiento del algoritmo. Para su ajuste, se ha utilizado el denominado *k-distance graph*, que representa la distancia al *k*-ésimo vecino más próximo para cada punto ordenado, permitiendo identificar visualmente un punto de inflexión (codo) que sirve de referencia para establecer `eps` [80].

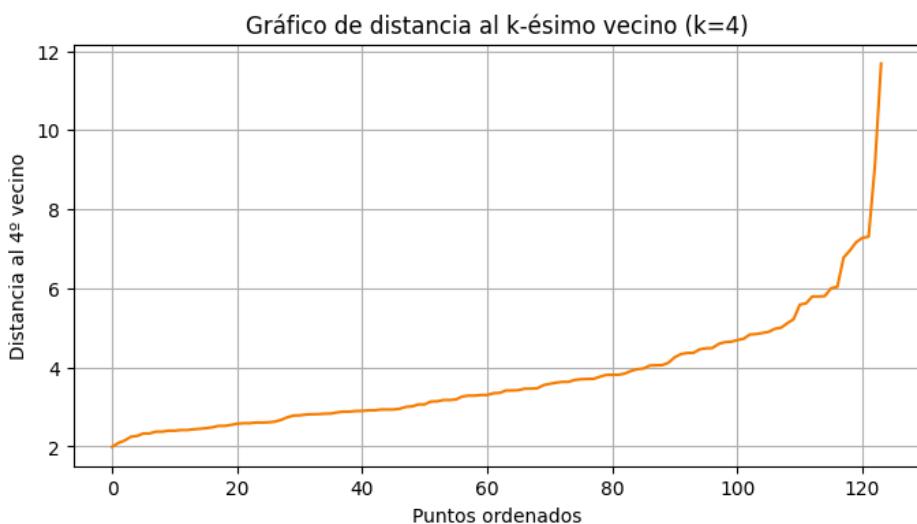


Imagen 13. Gráfico distancia DBSCAN barrios

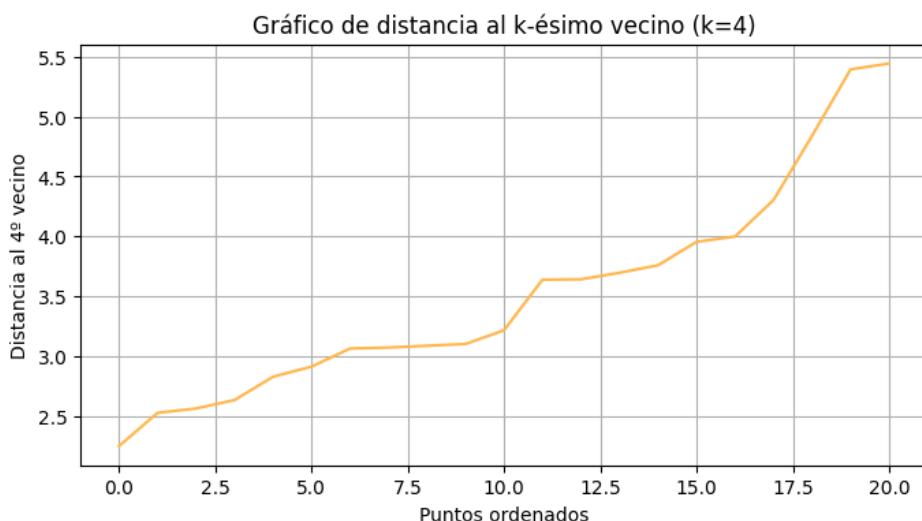


Imagen 14. Gráfico distancia DBSCAN distritos

En este proyecto se han fijado los siguientes parámetros como punto de partida para ambos niveles territoriales:

<i>min_samples</i>	Distancia al kº vecino	<i>eps</i>
4	4	2.5

Tabla 19. Ajuste parámetros DBSCAN

## Resultados obtenidos

Pese a los ajustes realizados, DBSCAN ha mostrado dificultades para generar agrupaciones estables en este conjunto de datos, principalmente debido a:

- La dispersión relativa de los valores normalizados en varias dimensiones.
- La escasa densidad de algunos subconjuntos territoriales.

En el caso de los barrios el resultado fue:

Nº clúster	Nº barrios
-1 (ruido)	90
0	17
1	4
2	9
3	4

Tabla 20. Nº barrios por clúster (DBSCAN)

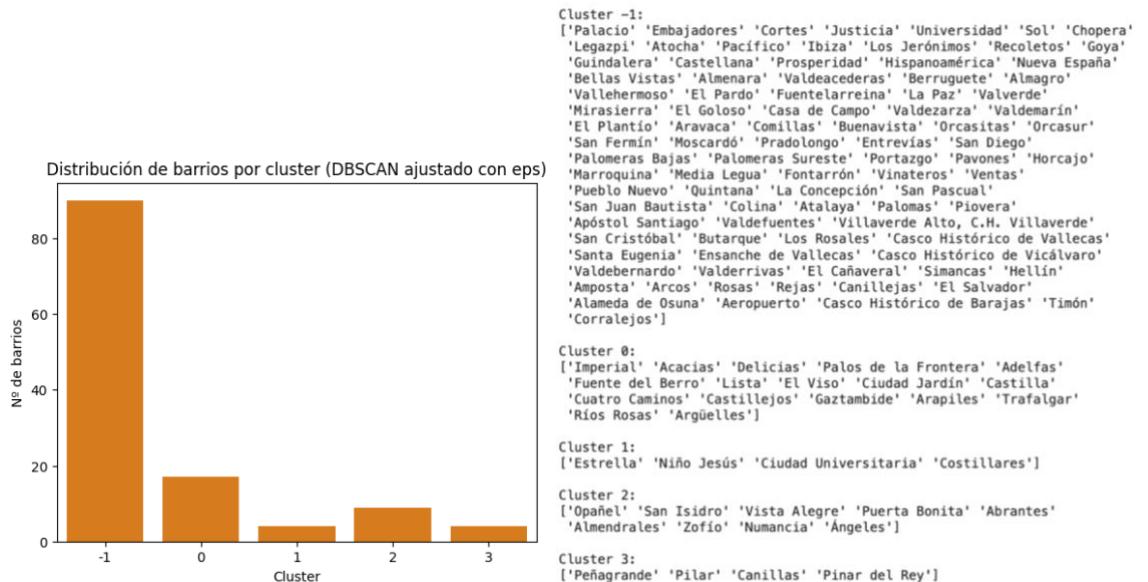


Imagen 15. Distribución barrios DBSCAN

En el caso de los distritos, DBSCAN clasificó de la siguiente manera:

Nº clúster	Nº distritos
-1 (ruido)	4
0	17

Tabla 21. Nº distritos por clúster (DBSCAN)

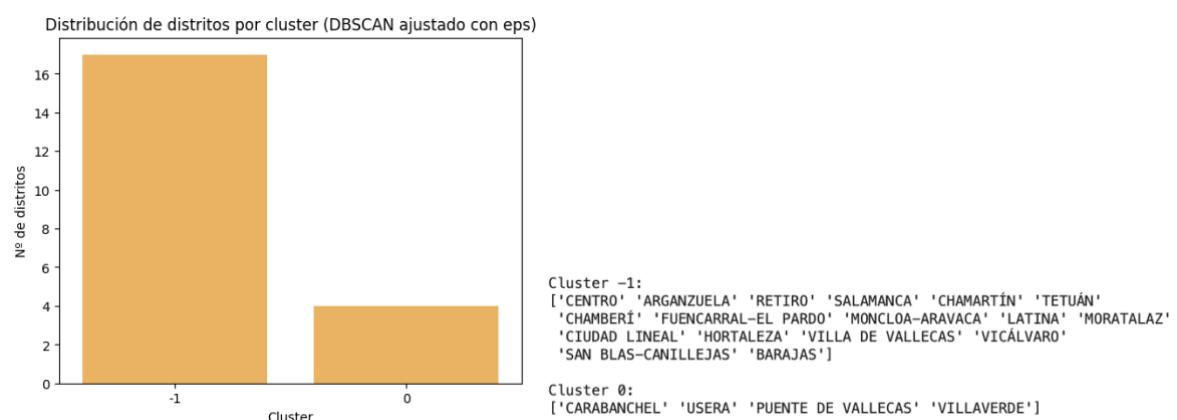


Imagen 16. Distribución distritos DBSCAN

La aplicación de DBSCAN ha demostrado las limitaciones de este algoritmo en escenarios de altas dimensiones con distribuciones dispersas como es el caso. Aunque es muy eficaz a la hora de detectar *outliers* o clústeres densos bien definidos, los patrones de vulnerabilidad territorial estudiados en Madrid presentan transiciones graduales más adecuadas para algoritmos como *K-means* o modelos jerárquicos.

Por este motivo, DBSCAN ha sido descartado como método principal de agrupamiento.

#### 5.4.3. *Agglomerative Clustering*

Como tercer enfoque de agrupamiento, se ha aplicado la técnica de *clustering* jerárquico aglomerativo, ampliamente empleada en contextos donde se busca explorar estructuras de similitud progresiva entre observaciones [81]. A diferencia de *K-means* o DBSCAN, este algoritmo no requiere especificar una distribución inicial de grupos a priori, sino que construye de forma iterativa una jerarquía de fusiones sucesivas.

##### **Funcionamiento**

*Agglomerative Clustering* parte de cada observación como un clúster independiente, fusionando en cada paso los dos grupos más próximos según un criterio de distancia. Este proceso continúa hasta formar un único clúster que agrupa todos los datos. El resultado puede representarse gráficamente mediante un dendrograma, que permite visualizar las fusiones y las distancias relativas entre agrupamientos.

El criterio de fusión ha sido el método de Ward, el cual minimiza el incremento total de varianza intra-clúster tras cada fusión. Matemáticamente, la distancia  $d(C_i, C_j)$  entre dos clústeres  $C_i$  y  $C_j$  según Ward se define como:

$$d(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|\bar{x}_i - \bar{x}_j\|^2$$

Donde  $n_i$  y  $n_j$  son los tamaños de cada clúster, y  $\bar{x}_i$ ,  $\bar{x}_j$  sus centroides respectivos [83]. Esta estrategia favorece la formación de grupos compactos y es especialmente apropiada en contextos de análisis de vulnerabilidad multivariante.

##### **Visualización previa: dendrogramas jerárquicos**

Antes de seleccionar el número final de los clústeres, se han generado los dendrogramas correspondientes para los dos niveles territoriales:

En el caso de los barrios se observan claramente tres ramas principales de separación, lo que sugiere de forma visual un número óptimo de  $k = 3$  clústeres.

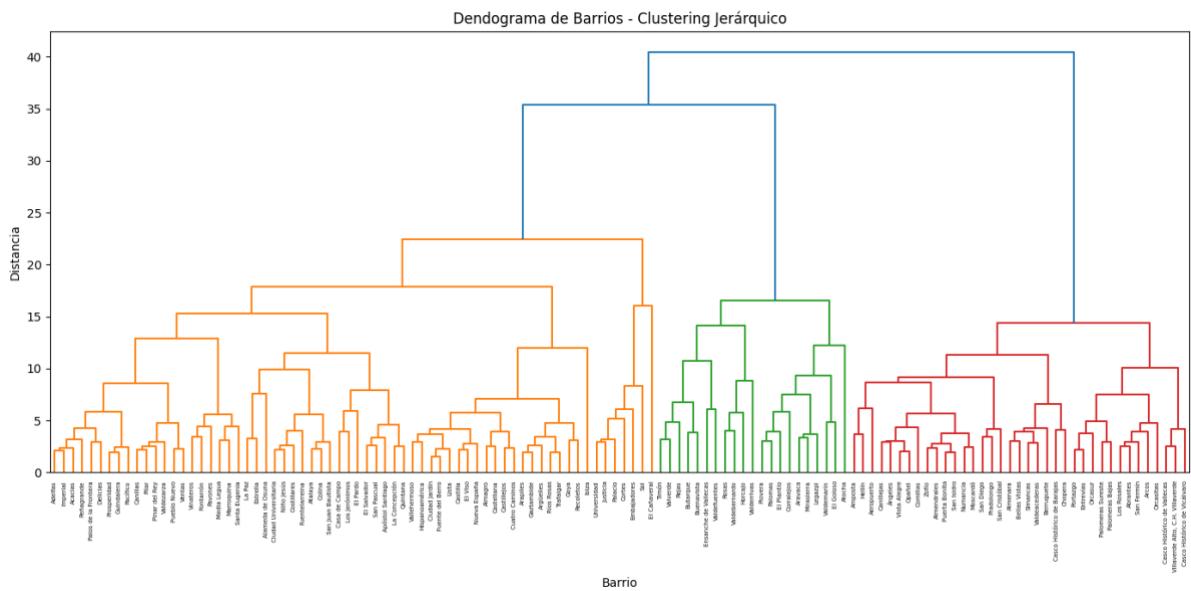


Imagen 17. Dendrograma barrios

En cuanto a los distritos, la estructura muestra una separación inicial en dos grandes grupos, con ramificaciones que permiten identificar claramente cuatro agrupamientos diferenciados, lo que sugiere  $k = 4$  como número óptimo.

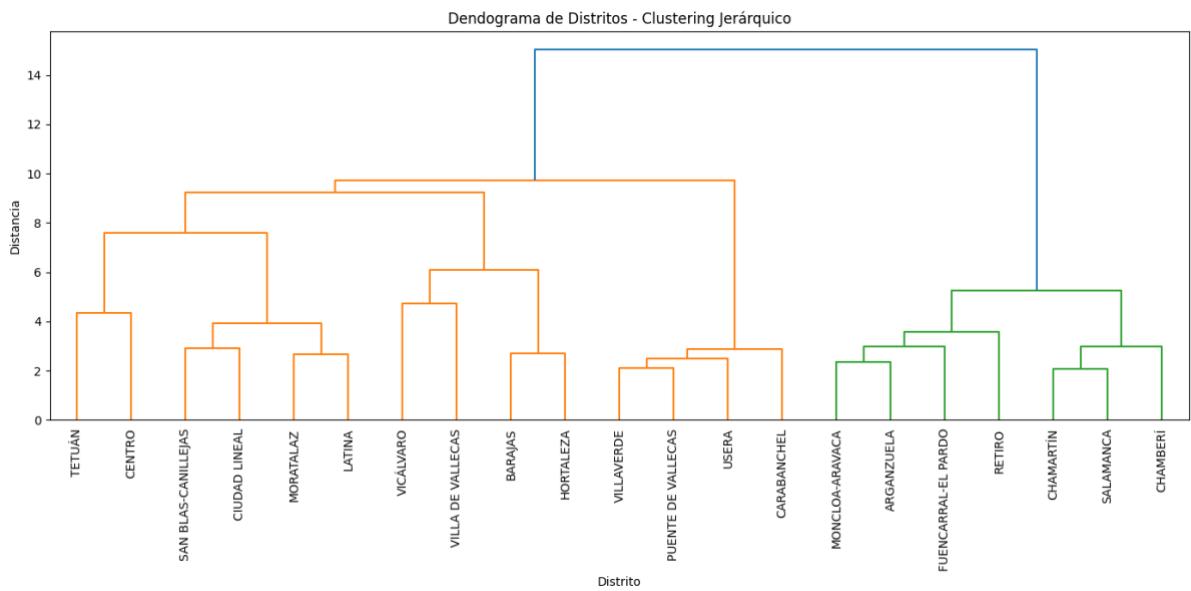


Imagen 18. Dendrograma distritos

### Determinación del número de clústeres ( $k$ )

Tomando como referencia la estructura visual de los dendrogramas, se han definido dos escenarios analíticos:

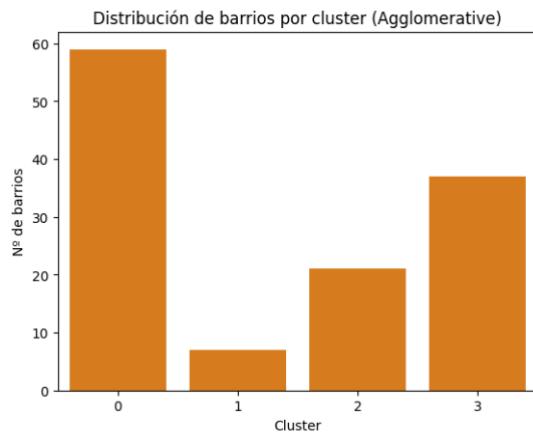
- Para barrios se han probado  $k = 3$  y  $k = 4$ , aunque el dendrograma sugiere que la primera opción es más natural.
- Para distritos se han probado  $k = 3$  y  $k = 4$  también, si bien el dendrograma indica que  $k = 4$  refleja mejor las divisiones estructurales observadas.

## Resultados para k=4

En primer lugar, se ha ejecutado el modelo *Agglomerative Clustering* para  $k = 4$  tanto en el análisis de distritos como de barrios. Los resultados muestran:

Nº clúster	Nº barrios
0	59
1	7
2	21
3	37

Tabla 22. Nº barrios por clúster (Agglomerative k=4)



Cluster 0:  
['Imperial' 'Acacias' 'Delicias' 'Palos de la Frontera' 'Pacífico'  
'Adefas' 'Estrella' 'Ibiza' 'Los Jerónimos' 'Niño Jesús' 'Recoletos'  
'Goya' 'Fuente del Berro' 'Guindalera' 'Lista' 'Castellana' 'El Viso'  
'Prosperidad' 'Ciudad Jardín' 'Hispanoamérica' 'Nueva España' 'Castilla'  
'Cuatro Caminos' 'Castillejos' 'Gaztambide' 'Arapiles' 'Trafalgar'  
'Almagro' 'Ríos Rosas' 'Vallehermoso' 'El Pardo' 'Fuentelarreina'  
'Peñagrande' 'Pilar' 'La Paz' 'Casa de Campo' 'Argüelles'  
'Ciudad Universitaria' 'Valdezarza' 'Pavones' 'Marroquina' 'Media Legua'  
'Fontarrón' 'Vinateros' 'Ventas' 'Pueblo Nuevo' 'Quintana'  
'La Concepción' 'San Pascual' 'San Juan Bautista' 'Colina' 'Atalaya'  
'Costillares' 'Canillas' 'Pinar del Rey' 'Apóstol Santiago'  
'Santa Eugenia' 'El Salvador' 'Alameda de Osuna']

Cluster 1:  
['Palacio' 'Embajadores' 'Cortes' 'Justicia' 'Universidad' 'Sol'  
'El Cañaveral']

Cluster 2:  
['Legazpi' 'Atocha' 'Valverde' 'Mirasierra' 'El Goloso' 'Valdemarín'  
'El Plantío' 'Aravaca' 'Buenavista' 'Horcajo' 'Palomas' 'Piovera'  
'Valdefuentes' 'Butarque' 'Ensanche de Vallecas' 'Valdebernardo'  
'Valderrivas' 'Rosas' 'Rejas' 'Timón' 'Corralejos']

Cluster 3:  
['Chopera' 'Bellas Vistas' 'Almenara' 'Valdeacederas' 'Berruguete'  
'Comillas' 'Opañel' 'San Isidro' 'Vista Alegre' 'Puerta Bonita'  
'Abantes' 'Orcasitas' 'Orcasur' 'San Fermín' 'Almendrales' 'Moscardó'  
'Zofío' 'Pradolongo' 'Entrevías' 'San Diego' 'Palomeras Bajas'  
'Palomeras Sureste' 'Portazgo' 'Numancia'  
'Villaverde Alto, C.H. Villaverde' 'San Cristóbal' 'Los Rosales'  
'Ángeles' 'Casco Histórico de Vallecas' 'Casco Histórico de Vicálvaro'  
'Simanca' 'Hellín' 'Amposta' 'Arcos' 'Callejas' 'Aeropuerto'  
'Casco Histórico de Barajas']

Imagen 19. Distribución barrios Agglomerative k=4

Nº clúster	Nº distritos
0	6
1	4
2	4
3	7

Tabla 23. Nº distritos por clúster (Agglomerative k=4)

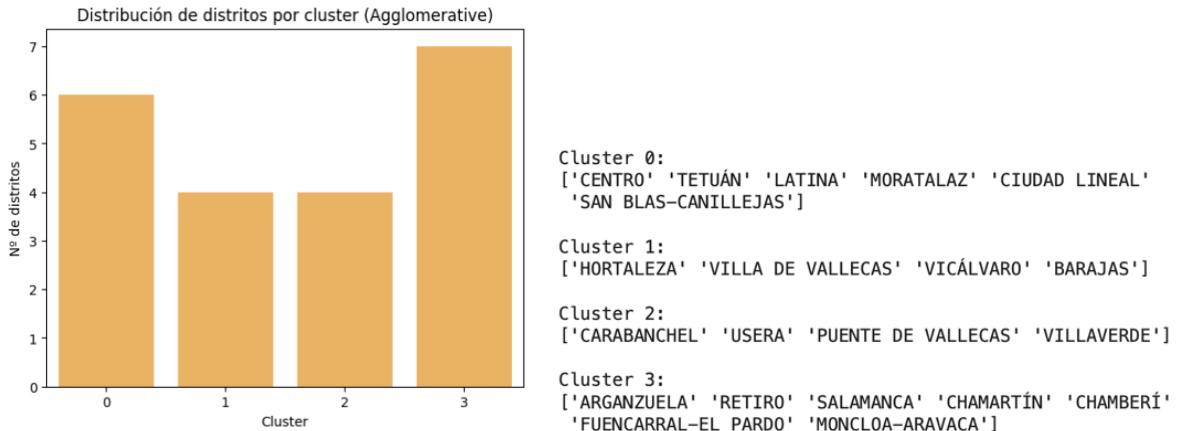


Imagen 20. Distribución distritos Agglomerative k=4

### Resultados para k=3

Como alternativa, se ha ejecutado el modelo con  $k = 3$ , con los siguientes resultados:

Nº clúster	Nº barrios
0	66
1	37
2	21

Tabla 24. Nº barrios por clúster (Agglomerative k=3)

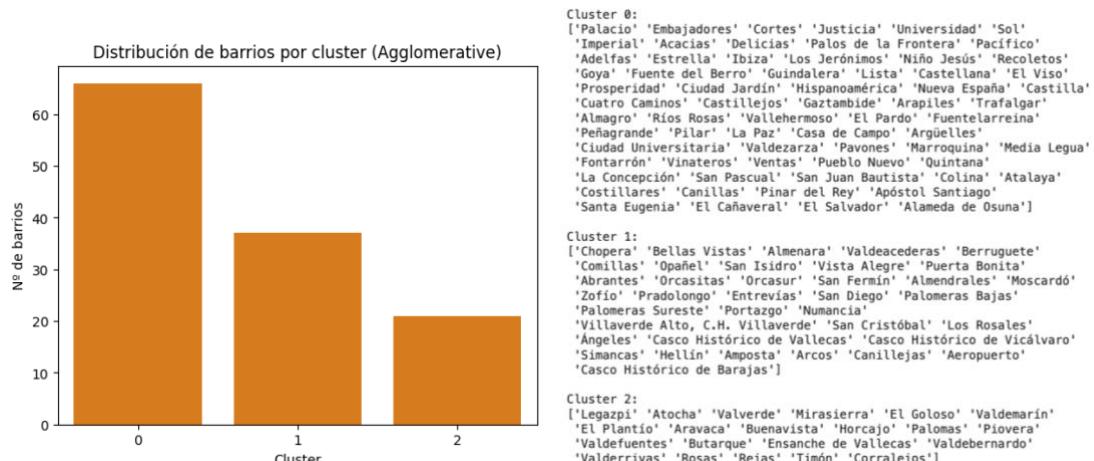


Imagen 21. Distribución barrios Agglomerative k=3

Nº clúster	Nº distritos
0	10
1	7
2	4

Tabla 25. Nº distritos por clúster (Agglomerative k=3)

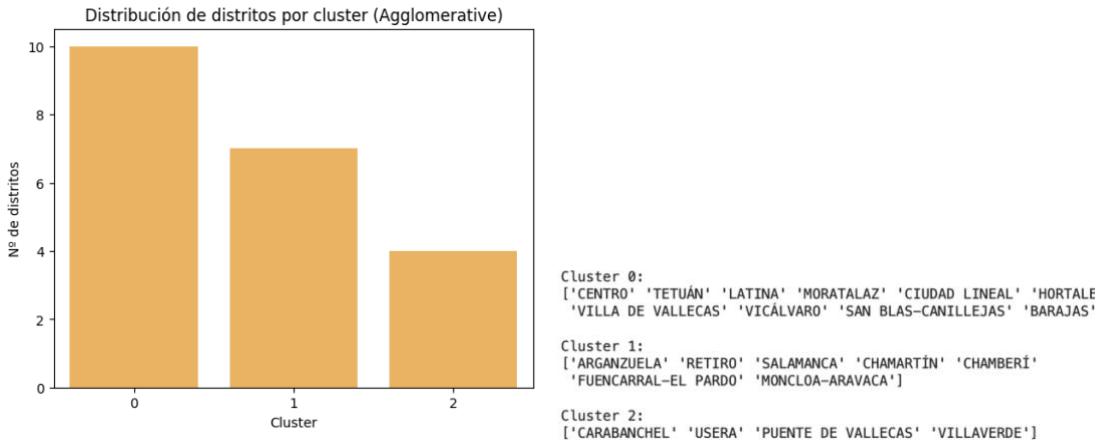


Imagen 22. Distribución distritos Agglomerative k=3

## 5.5. GENERACIÓN DE RANKINGS Y REPRESENTACIÓN ESPACIAL

### 5.5.1. Metodología cálculo ranking

Una vez obtenidas las particiones de los diferentes modelos de *clustering* aplicados, se ha procedido a establecer un ranking de vulnerabilidad territorial que permita comparar y visualizar de forma ordenada los niveles de riesgo socioeconómico identificados en cada zona.

El procedimiento aplicado parte de los datos normalizados (*z-score*) utilizados en el *clustering*. Para cada clúster generado, se calcula la media de los valores estandarizados de todos los indicadores considerados, siguiendo la fórmula:

$$\bar{z}_c = \frac{1}{p} \sum_{j=1}^p \frac{1}{n_c} \sum_{i=1}^{n_c} z_{ij}$$

Donde  $p$  es el número total de variables,  $n_c$  es el número de observaciones dentro del clúster  $c$  y  $z_{ij}$  es el valor estandarizado de la variable  $j$  para la observación  $i$ .

Posteriormente, los clústeres son ordenados de menor a mayor media global ( $\bar{z}_c$ ), bajo la premisa de que una media más baja representa un perfil de mayor vulnerabilidad relativa en el conjunto de dimensiones analizadas. Este criterio es coherente con el escalado previo, donde valores positivos representan mejores condiciones socioeconómicas, educativas o de salud

Como resultado, se asigna a cada clúster una posición de ranking, donde el grupo con la media más alta recibe el rango 1 (máxima vulnerabilidad), descendiendo sucesivamente hasta el de menor vulnerabilidad. Este proceso ha sido aplicado de forma independiente para cada algoritmo y configuración (*K-means* con  $k = 3$  y  $k = 4$ ; *Agglomerative* con  $k = 3$  y  $k = 4$ ; y *DBSCAN*), diferenciando también entre barrios y distritos. De esta forma:

Para k=3	
Nº Ranking	Categoría asignada
1	Muy vulnerable
2	Vulnerable
3	Poco vulnerable

Tabla 26. Categoría asignada según nº ranking para k=3

Para k=4	
Nº Ranking	Categoría asignada
1	Muy vulnerable
2	Vulnerable
3	Poco vulnerable
4	Muy poco vulnerable

Tabla 27. Categoría asignada según nº ranking para k=4

### 5.5.2. Visualización de los agrupamientos mediante análisis PCA

Como complemento al cálculo del ranking, se ha implementado un análisis exploratorio visual basado en el Análisis de Componentes Principales (PCA), proyectando las observaciones multivariantes en un espacio bidimensional.

El PCA permite reducir dimensionalidad extrayendo las combinaciones lineales de variables originales que explican el máximo de varianza, facilitando la representación gráfica de los clústeres formados. Aunque esta proyección no interviene en la clasificación ni en el ranking, resulta útil para validar de forma preliminar la separación estructural de los grupos generados.

En los siguientes gráficos se representan los barrios/distritos según sus dos primeros componentes principales, coloreados por clúster, lo que permite observar visualmente la compactación interna de cada grupo, el grado de separación o solapamiento entre clústeres y la estructura relativa de los modelos para  $k = 3$  y  $k = 4$ .

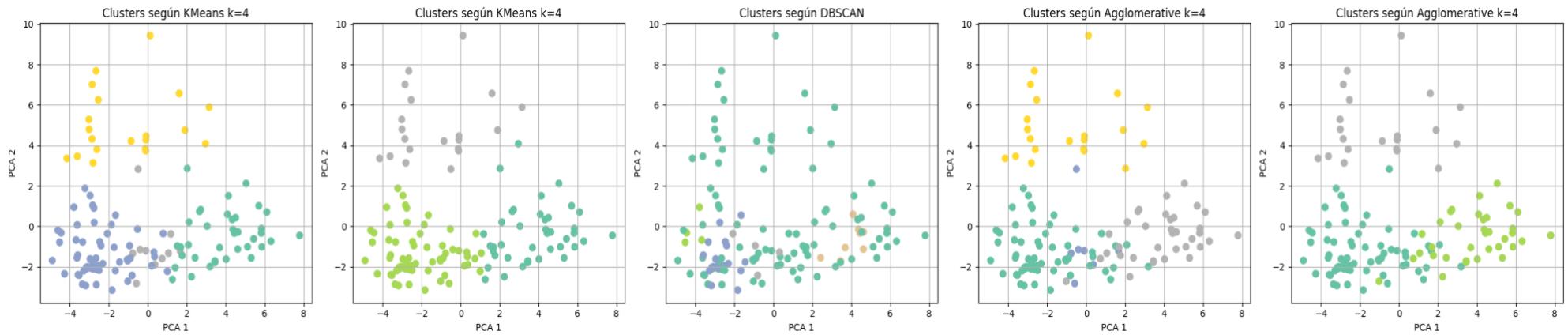


Imagen 23. Representación PCA modelos para barrios

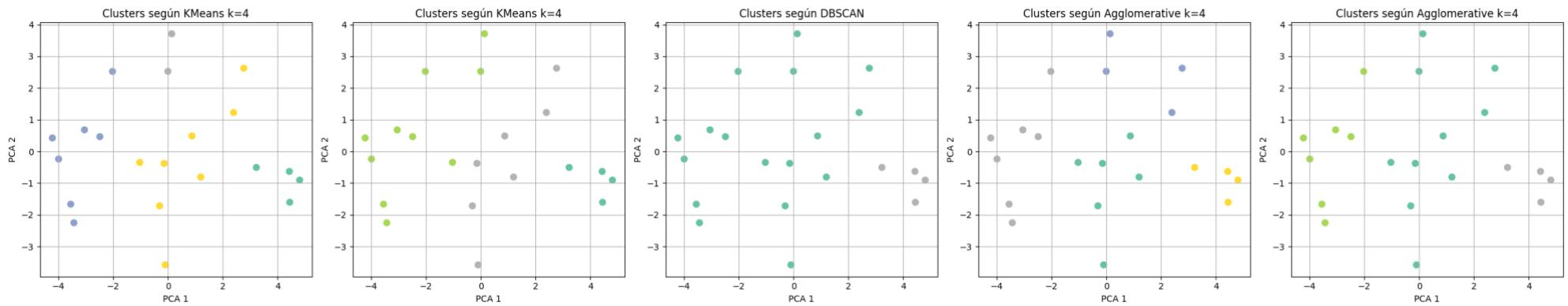


Imagen 24. Representación PCA modelos para distrito

Como se puede observar en la Imagen 23, la proyección bidimensional de los distintos modelos permite visualizar de forma intuitiva la estructura interna de los agrupamientos generados. En el caso de *K-means*, tanto para  $k = 4$  como para  $k = 3$ , los clústeres parecen relativamente bien definidos y separados en el espacio PCA, evidenciando la capacidad del algoritmo para identificar perfiles diferenciados de vulnerabilidad. No obstante, aunque la opción  $k = 4$  muestra una segmentación más granular (permite identificar subgrupos adicionales dentro de los perfiles vulnerables), visualmente parece existir un mayor solapamiento entre algunos clústeres. Por el contrario, en el modelo con  $k = 3$  se aprecia una separación algo más limpia entre grupos, lo que sugiere una estructura más coherente en términos de distancia entre observaciones. Este comportamiento será evaluado posteriormente de forma cuantitativamente por el índice de Silhouette Score.

En el caso de DBSCAN, como era de esperar por su dificultad en escenarios multivariantes con poca homogeneidad, los agrupamientos se muestran claramente menos definidos, con gran dispersión de puntos y presencia destacada de observaciones clasificadas como ruido (grupo -1), especialmente en barrios. El modelo *Agglomerative* muestra, en general, una distribución visual muy próxima a la de *K-means*, confirmando su buena adaptación a los patrones de datos multivariantes heterogéneos. Nuevamente, se aprecia que  $k = 4$  genera divisiones más finas, pero no necesariamente mejor definidas que en  $k = 3$ .

Por su parte, en la Imagen 24 correspondiente al nivel de distritos, la proyección PCA refleja un comportamiento parecido. En *K-means* y *Agglomerative*, los clústeres con  $k = 3$  tienden a agrupar los distritos de forma más compacta, mientras que  $k = 4$  produce una mayor división interna. Dada la menor cantidad de observaciones en este caso (solo 21 distritos), las diferencias en la separación de grupos resultan especialmente visibles. Aunque  $k = 4$  permite identificar subdivisiones adicionales, a simple vista  $k = 3$  muestra un reparto algo más estable y con menor solapamiento entre grupos, algo que también se validará de forma objetiva más adelante en el análisis de los *Silhouette Scores*.

### 5.5.3. Representación cartográfica rankings

Para completar la interpretación espacial de los resultados de agrupamiento, se han generado distintos mapas interactivos que permiten visualizar de forma geográfica los niveles de vulnerabilidad territorial calculados para cada modelos y nivel de análisis (barrios y distritos). Dichos mapas interactivos han sido anexados en el Anexo A.

En todos los casos, los mapas muestran el ranking de vulnerabilidad asignado a cada zona, siguiendo la metodología descrita previamente, donde los grupos han sido clasificados desde muy vulnerables hasta muy poco vulnerables, en función de la media estandarizada de sus indicadores.

#### Escala de colores aplicada

Con el objetivo de facilitar la lectura e interpretación visual, se ha utilizado una escala de colores en tonos naranjas, donde:

- Naranja oscuro representa las zonas con mayor nivel de vulnerabilidad.

- Naranja claro representa las zonas con menos vulnerabilidad.

## Agrupación de mapas por algoritmo

Para una visión comparativa más clara, los resultados se presentan agrupados por algoritmo, mostrando tanto los resultados a nivel de barrios como de distritos de forma paralela, lo que permite comprobar si los patrones de vulnerabilidad detectados son coherentes entre los distintos niveles territoriales:

### K-means k=3

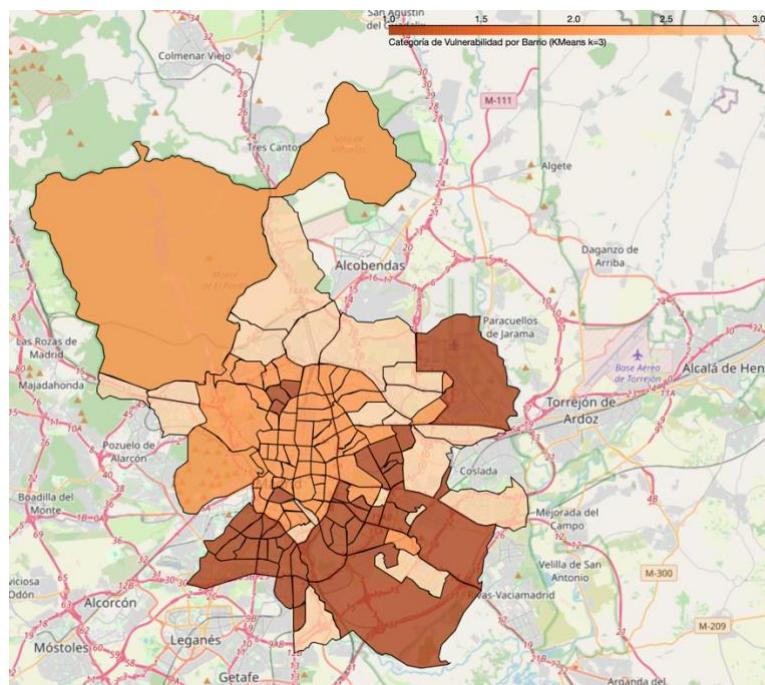


Imagen 25. Mapa barrios K-means k=3

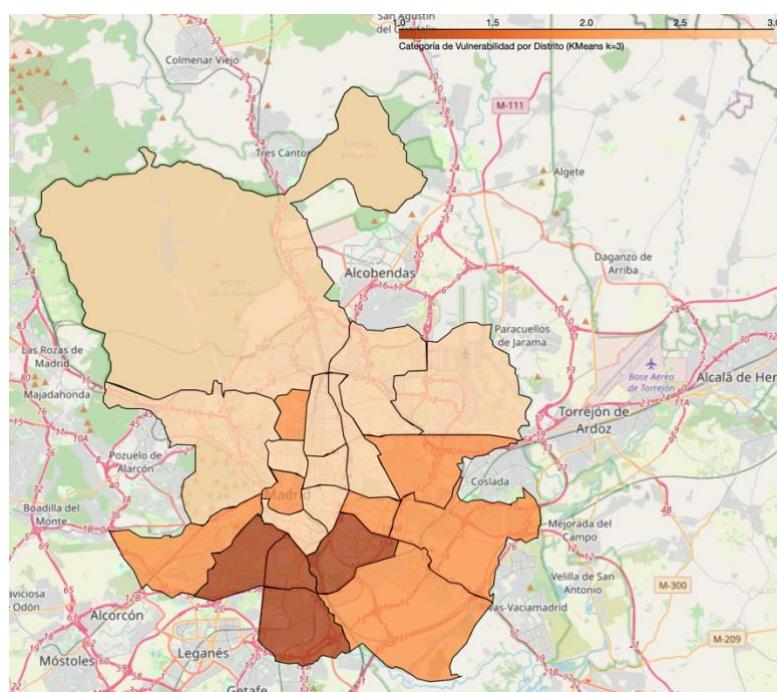


Imagen 26. Mapa distritos K-means k=3

### K-means k=4

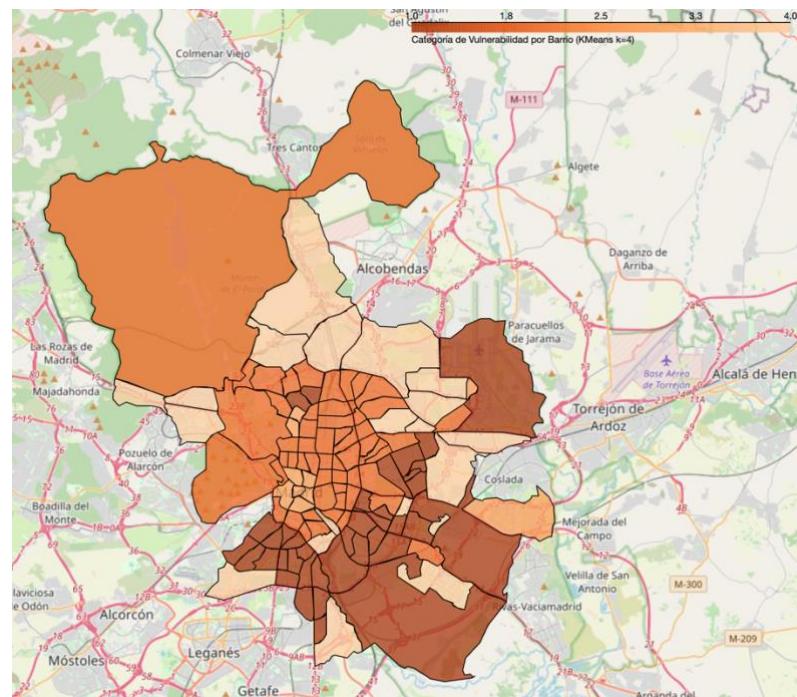


Imagen 27. Mapa barrios K-means k=4

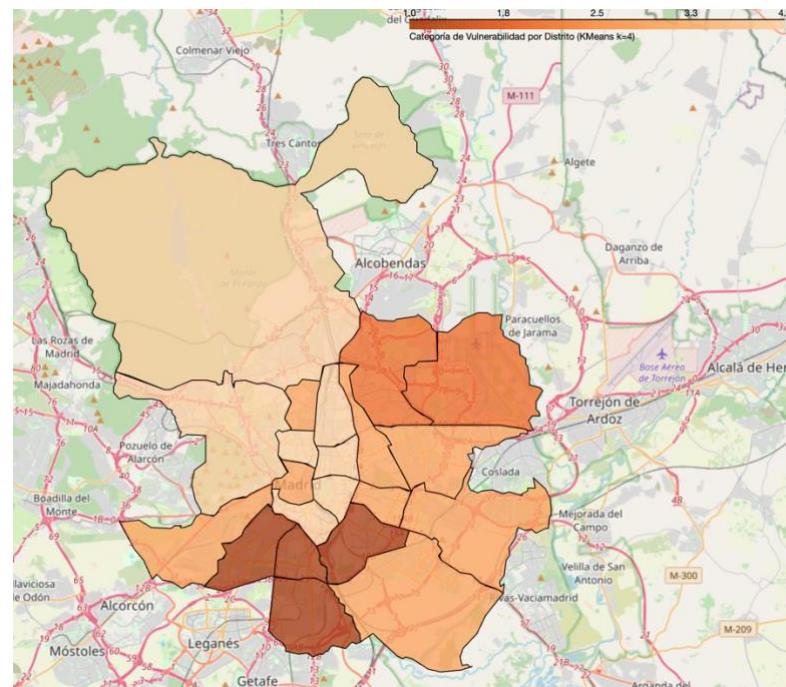


Imagen 28. Mapa distritos K-means k=4

## Agglomerative k=3

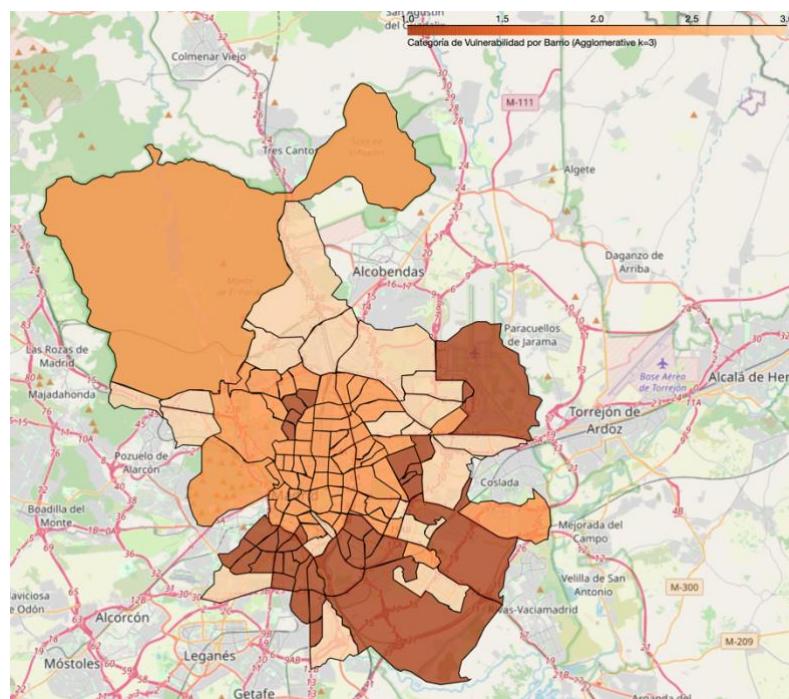


Imagen 29. Mapa barrios Agglomerative k=3

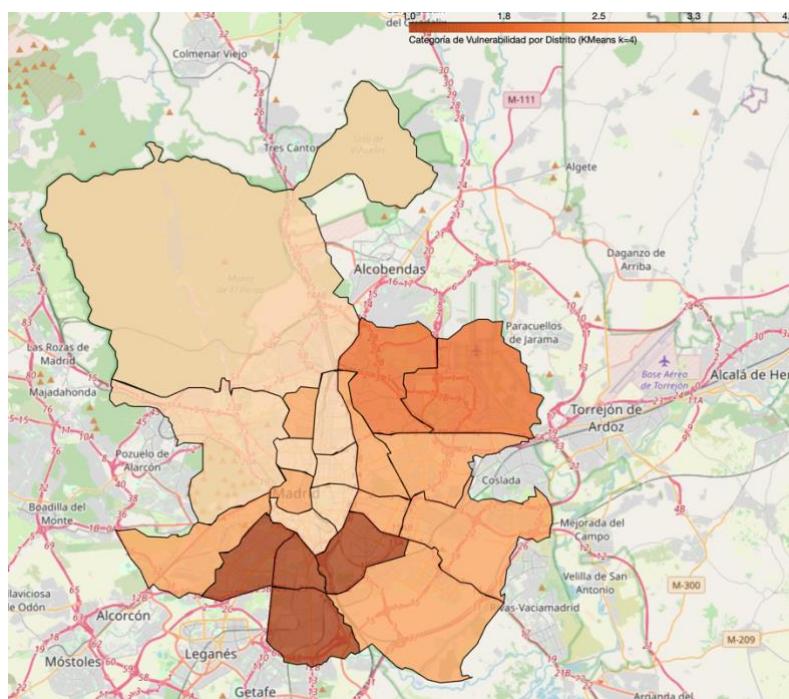
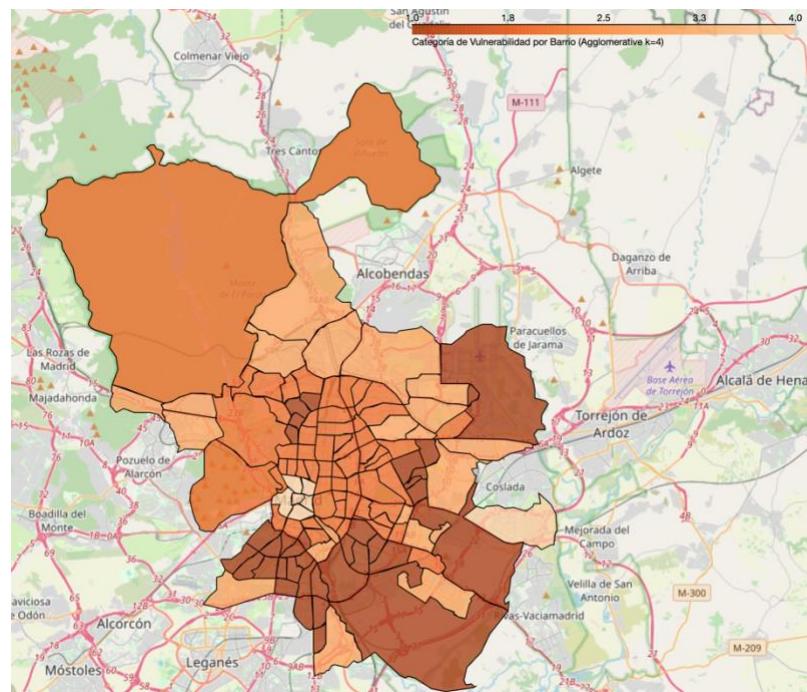
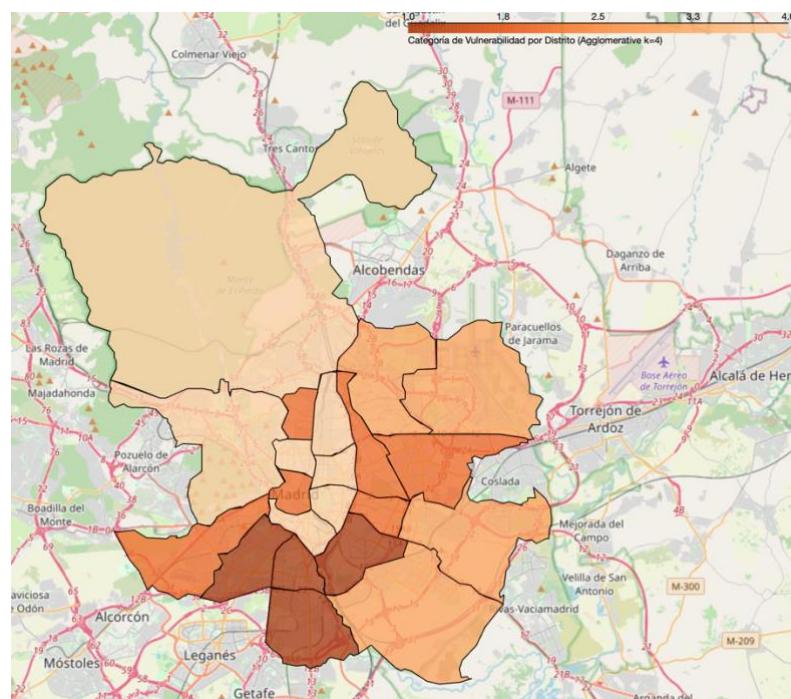


Imagen 30. Mapa distritos Agglomerative k=3

### **Agglomerative k=4**



*Imagen 31. Mapa barrios Agglomerative k=4*



*Imagen 32. Mapa distritos Agglomerative k=4*

## DBSCAN

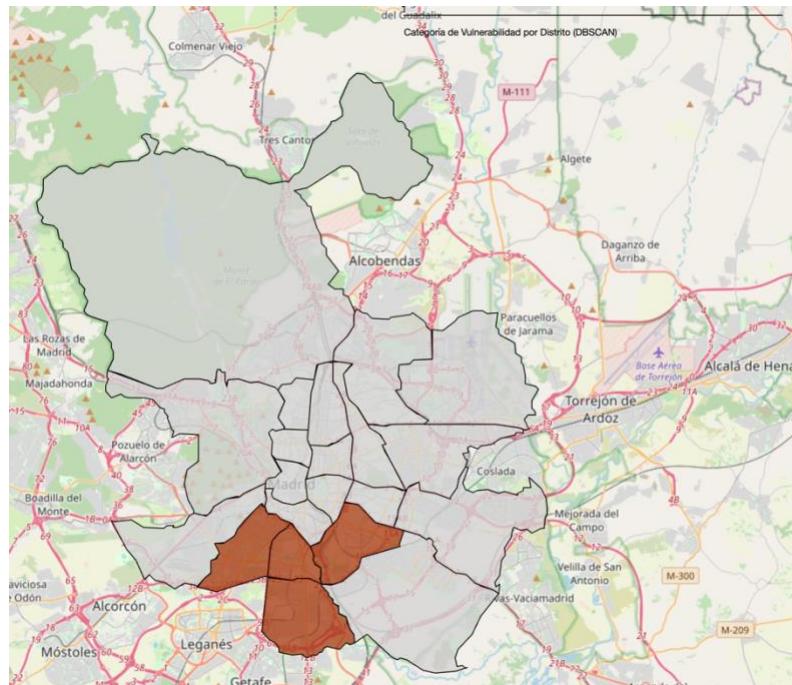


Imagen 33. Mapa barrios DBSCAN

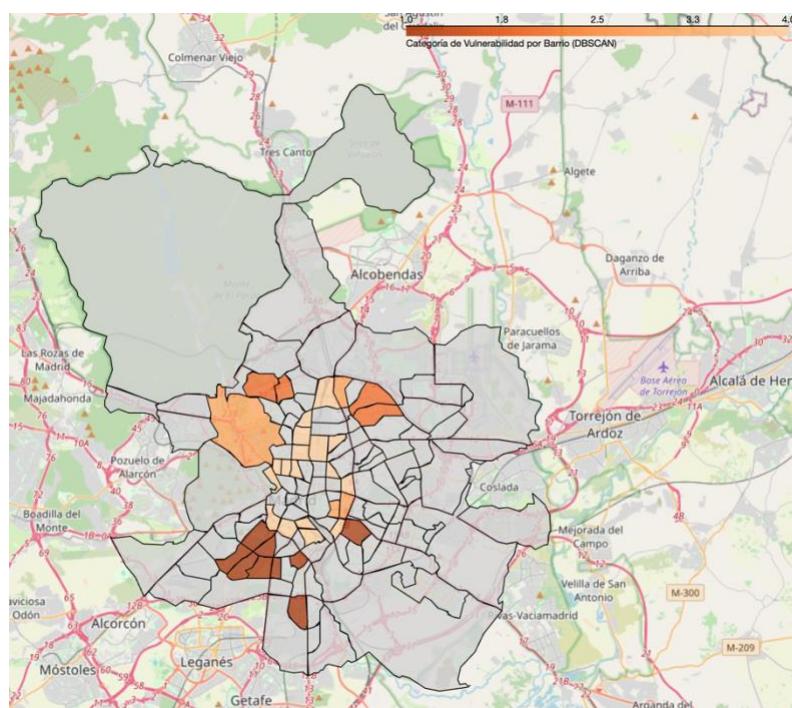


Imagen 34. Mapa distritos DBSCAN

## Observaciones preliminares

Esta representación permite observar cómo, en general, existe una elevada coherencia territorial: los distritos identificados como más vulnerables tienden a englobar también a barrios pertenecientes a esos mismos territorios que presentan igualmente mayores niveles de riesgo, lo que valida la consistencia de los patrones detectados.

Por otra parte, como ya se venía anticipando en los análisis previos, el algoritmo DBSCAN muestra nuevamente sus limitaciones en este contexto: visualmente, es evidente cómo deja sin clasificar (etiqueta como ruido, coloreado de gris en los mapas) un número considerable de barrios y distritos, lo que dificulta su interpretación práctica.

En cuanto a la comparación entre los valores de  $k$ , puede observarse que la diferencia entre las segmentaciones obtenidas para  $k = 3$  y  $k = 4$  no resulta excesivamente marcada en la mayoría de los casos. La opción  $k = 4$  permite descomponer algo más los perfiles intermedios, mientras que  $k = 3$  ofrece una clasificación más simplificada pero igualmente coherente.

Además, al comparar los resultados entre los algoritmos *K-means* y *Agglomerative*, se aprecia que, aunque presentan ligeras diferencias en los agrupamientos concretos de algunas zonas, los patrones generales identificados son muy similares, reforzando la robustez de los perfiles de vulnerabilidad detectados en la ciudad de Madrid con ambos enfoques.

En los siguientes apartados se procederá a evaluar cuantitativamente la calidad de estas segmentaciones mediante el cálculo de los *Silhouette Scores*, así como analizar en ranking los rankings finales.

## 6. RESULTADOS

---

En este capítulo se describen e interpretan los resultados obtenidos tras aplicar los algoritmos de *clustering* al conjunto de datos de vulnerabilidad territorial. Se presentan las métricas de validación empleadas para evaluar la calidad de los modelos y se analiza en detalle la configuración final seleccionada, incorporando las variables más influyentes y los rankings de vulnerabilidad generados.

### 6.1. VALIDACIÓN DE LOS MODELOS

Para evaluar la calidad de las particiones generadas por los diferentes algoritmos y configuraciones probadas, se han aplicado dos métricas de validación interna ampliamente aceptadas en estos contextos: el *Silhouette Score* y el *Calinski-Harabasz Score*.

#### 6.1.1. Validación mediante *Silhouette Score*

El primer método aplicado ha sido el *Silhouette Score* [84], el cual mide simultáneamente la cohesión intra-clúster y la separación inter-clúster. Su fórmula se expresa como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Donde  $a(i)$  es la distancia media de la observación  $i$  respecto al resto de puntos de su mismo clúster y  $b(i)$  es la distancia media de  $i$  al clúster más próximo. El índice toma valores entre -1 y 1, donde los valores cercanos a 1 indican una mejor estructura de *clustering*. Los resultados obtenidos se resumen en las siguientes tablas:

Barrios	
Modelo	<i>Silhouette Score</i>
<i>K-means</i> k=4	0.26
<i>K-means</i> k=3	0.26
DBSCAN	-1
<i>Agglomerative</i> k=4	0.25

<i>Agglomerative k=3</i>	0.25
--------------------------	------

Tabla 28. Silhouette Score por modelo (barrios)

Distritos	
Modelo	Silhouette Score
<i>K-means k=4</i>	0.26
<i>K-means k=3</i>	0.26
DBSCAN	-1
<b><i>Agglomerative k=4</i></b>	<b>0.27</b>
<b><i>Agglomerative k=3</i></b>	<b>0.27</b>

Tabla 29. Silhouette Score por modelo (distritos)

Como se observa, los valores son relativamente bajos, pero muestran coherencia entre modelos y tamaños de  $k$ . Además, las diferencias entre  $k = 3$  y  $k = 4$  son mínimas, lo que sugiere que ambos proporcionan segmentaciones relativamente estables. El caso de DBSCAN ha resultado claramente inadecuado en este contexto, obteniendo un *Silhouette Score* de -1 al dejar numerosos puntos sin agrupar (ruido), lo que ya se había detectado visualmente en las representaciones PCA.

### 6.1.2. Validación mediante *Calinski-Harabasz Score*

Dado que el *Silhouette Score* ofrecía resultados poco discriminativos entre configuraciones, se ha recurrido a un segundo método de validación: el *Calinski-Harabasz Score* (CHS) [85]. Este índice se calcula según:

$$s = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n_E - k}{k - 1}$$

Donde  $tr(B_k)$  es la traza de la matriz de dispersión entre clústeres,  $tr(W_k)$  es la traza de la matriz de dispersión intra-clúster,  $n_E$  el número total de observaciones y  $k$  el número de clústeres. Estas matrices de dispersión se calculan según:

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^k n_q (c_q - c_E)(c_q - c_E)^T$$

Siendo  $C_q$  el conjunto de observaciones asignadas al clúster  $q$ ,  $c_q$  el centroide del clúster  $q$ ,  $c_E$  el centroide global del conjunto de datos y  $n_q$  el número de observaciones en el clúster  $q$ .

En este caso, cuanto mayor sea el valor de  $s$ , mejor es la calidad del agrupamiento, ya que indica mayor separación inter-clúster en relación con la variación interna. Los resultados fueron:

Barrios	
Modelo	Calinski-Harabasz Score
<i>K-means</i> k=4	31.09
<b><i>K-means</i> k=3</b>	<b>37.5</b>
<i>Agglomerative</i> k=4	29.09
<i>Agglomerative</i> k=3	33.79

Tabla 30. Calinski-Harabasz Score por modelo (barrios)

Distritos	
Modelo	Calinski-Harabasz Score
<i>K-means</i> k=4	7.58
<i>K-means</i> k=3	7.16
<b><i>Agglomerative</i> k=4</b>	<b>8.78</b>
<i>Agglomerative</i> k=3	8.3

Tabla 31. Calinski-Harabasz Score por modelo (distritos)

De acuerdo con los resultados combinados de ambas métricas:

- Los modelos con  $k = 3$  en barrios (*K-means*) y  $k = 4$  en distritos (*Agglomerative*) presentan los mejores indicadores de separación.
- Las diferencias entre *K-means* y *Agglomerative* son reducidas, lo que sugiere que ambos métodos modelan patrones muy similares en el espacio multivariante.
- El algoritmo DBSCAN vuelve a confirmar su baja aplicabilidad a estos datos, dejando múltiples observaciones sin asignación.

## 6.2. ANÁLISIS DE LOS PERFILES DE LOS CLÚSTERES

Una vez validados los modelos de *clustering*, se ha procedido al análisis detallado de los perfiles resultantes, tanto para los distritos como para los barrios. Para cada clúster se calculan las medias de cada variable normalizada (*z-score*) dentro de cada clúster y se comparan con las medias globales del conjunto de datos. Este enfoque permite identificar qué indicadores presentan valores significativamente por encima o por debajo de la media global del conjunto de datos, reflejando las características distintivas de cada grupo.

Concretamente, para cada clúster  $c$  y cada variable  $j$ , se calcula:

$$\Delta z_{c,j} = \bar{z}_{c,j} - \bar{z}_j$$

Donde  $\bar{z}_{c,j}$  es la media de la variable  $j$  dentro del clúster  $c$  y  $\bar{z}_j$  es la media global de la variable  $j$  en todo el conjunto de datos.

Este desplazamiento  $\Delta z_{c,j}$  indica cómo de alejado está el clúster de la media global para esa variable. Los valores positivos reflejan que, en ese clúster, la variable está por encima de la media global; los valores negativos indican que está por debajo.

Aunque el algoritmo de *clustering* (*K-means* o *Agglomerative*) no asigna pesos a las variables de forma explícita, al calcular estas medias dentro de cada clúster observamos en qué variables las zonas agrupadas presentan valores significativamente diferentes respecto al conjunto. Cuanto mayor es la diferencia, mayor ha sido el papel de esa variable en provocar que esas zonas hayan quedado agrupadas. Es decir, si varias zonas coinciden en tener valores extremos en una misma variable, es muy probable que esa variable haya contribuido al agrupamiento.

Además, para interpretar y comparar estos resultados de forma más completa, se han generado para cada modelo matrices de correlación, que muestran el perfil medio de cada variable por clúster, y se han representado mediante mapas de calor (*heatmaps*). Esta información permite no solo describir los grupos de forma individual, sino también comparar los perfiles entre sí y evaluar la coherencia de los patrones que se han detectado.

En los siguientes cuadros se presentan, para cada clúster, las cinco variables cuyo desplazamiento respecto a la media global es mayor, es decir, aquellas que definen con más fuerza el perfil diferencial de cada grupo. En el anexo B se puede consultar el listado completo de las variables ordenadas de mayor a menor influencia dentro de cada clúster, según el valor del desplazamiento calculado.

### 6.2.1. Resultados a nivel de barrios (*K-means* k=3)

A nivel de barrios, el modelo de *K-means* con  $k = 3$  ha mostrado patrones muy consistentes, alineados con los perfiles obtenidos a nivel distrital que se observarán más adelante, lo que evidencia coherencia territorial en la distribución de la vulnerabilidad:

- **Clúster 0:** caracteriza barrios con fuerte componente demográfico envejecido (hogares unipersonales mayores de 65 años), alta edad media y familias monoparentales.

Variable	$\Delta z_{c,j}$
Índice de Vulnerabilidad Economía y Empleo	1.16
Población mayor/igual de 25 años con enseñanza primaria incompleta	1.14
Población mayora/igual de 25 años con Bachiller Elemental, Graduado Escolar, ESO, FP 1º grado	1.13
Población mayor/igual de 25 años que no sabe leer ni escribir o sin estudios	1.13
Índice de Vulnerabilidad Educación y Cultura	1.08

Tabla 32. Variables más influyentes en clúster 0 (barrios)

- **Clúster 1:** agrupa barrios con mayor proporción de población infantil, alta concentración de hogares con menores y todavía elevados niveles de vulnerabilidad económica y educativa.

Variable	$\Delta z_{c,j}$
Porcentaje de envejecimiento (Población mayor de 65 años/Población total)	0.67
Edad media de la población	0.65
Población mayor/igual de 25 años con estudios superiores, licenciatura, arquitectura, ingeniería sup., estudios sup. no universitarios, doctorado, postgrado	0.64
Hogares con un hombre solo mayor de 65 años	0.61
Hogares con una mujer sola mayor de 65 años	0.48

Tabla 33. Variables más influyentes en clúster 1 (barrios)

- **Clúster 2:** recoger los barrios con mejores condiciones socioeconómicas y educativas, población más joven y mejor nivel de estudios.

Variable	$\Delta z_{c,j}$
Población en etapa educativa (Población de 0 a 16 años -16 no incluidos)	1.8
Hogares monoparentales: una mujer adulta con uno o más menores	1.63
Población mayora/igual de 25 años con titulación media, diplomatura, arquitectura o ingeniería técnica	1.16
Población en etapa educativa de 12 a 15 años	0.75
Población en etapa educativa de 6 a 11 años	0.73

Tabla 34. Variables más influyentes en clúster 2 (barrios)

En términos de variables más influyentes en la segmentación de los barrios, destacan:

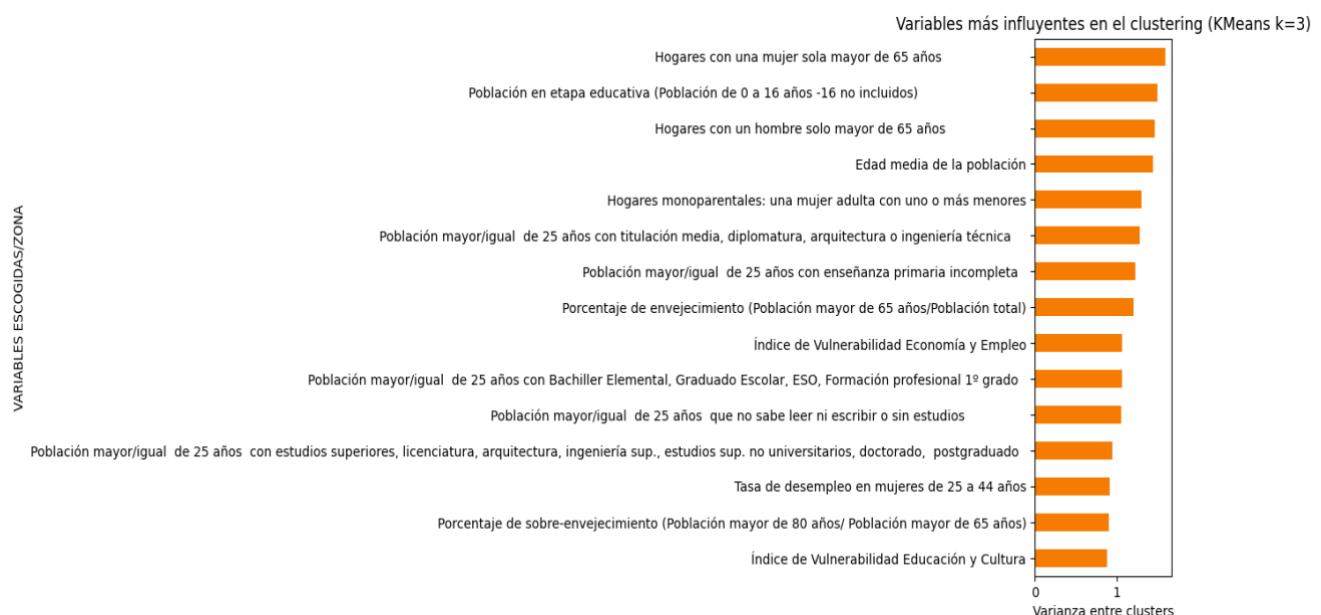


Imagen 35. Variables más influyentes en modelo K-means k=3 (barrios)

Variable	Varianza entre clústeres
Hogares monoparentales y unipersonales	1.4-1.5
Población en etapa educativa (0-16 años)	1.4
Edad media de la población	1.3
Nivel educativo bajo (ESO, Primaria Incompleta...)	1.2
Vulnerabilidad económica, educación y empleo	1-1.1

Tabla 35. Resumen variables más influyentes en modelo K-means k=3 (barrios)

Se observa nuevamente el alto peso que tienen las variables demográficas (estructura familiar, envejecimiento), educativas (nivel formativo alcanzado) y económicas (empleo, renta) en la diferenciación de perfiles.

Por otro lado, si observamos la matriz de medias normalizadas, se confirma el papel clave de variables como el Índice de Vulnerabilidad Económica y Empleo (1.2) o la tasa de desempleo (0.99-1.1) en la definición del clúster 0, junto a indicadores educativos como la población con estudios básicos o sin estudios (1.1). También se aprecian valores negativos en variables de envejecimiento o dependencia en los clústeres menos vulnerables. Un valor más alto y positivo refleja mayor contribución de esa variable a diferenciar el grupo, mientras que uno bajo y negativo indica menos influencia o condiciones más favorables, en este caso. De esta manera queda respaldado lo comentado anteriormente.

A continuación, se muestra la matriz de medias normalizadas para este modelo:

Indicadores

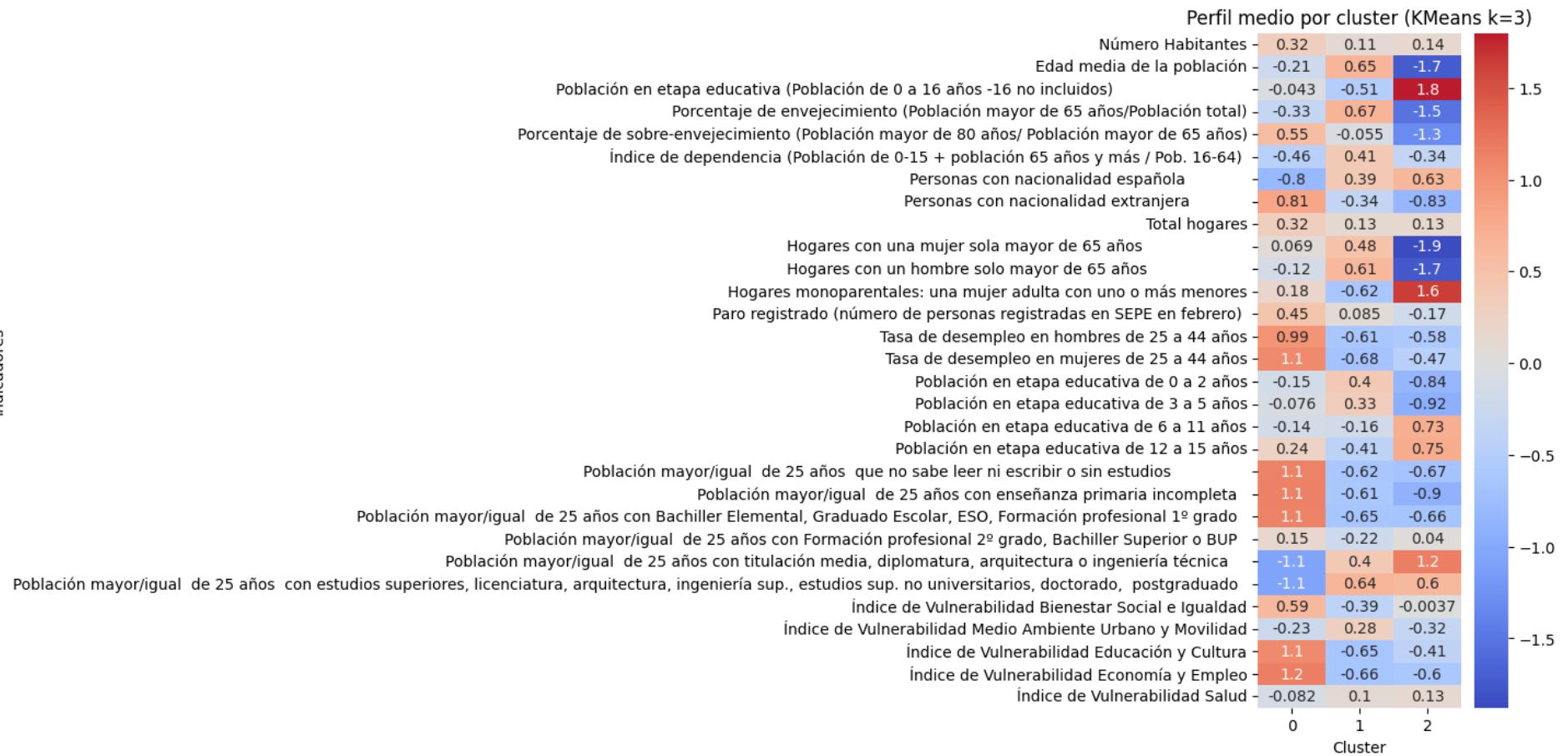


Imagen 36. Matriz de medias normalizadas modelo K-means k=3 (barrios)

### 6.2.2. Resultados a nivel de distritos (*Agglomerative k=4*)

El modelo de *Agglomerative clustering* con  $k = 4$  para los distritos ha permitido segmentar el territorio madrileño en cuatro perfiles diferenciados. El análisis de los centroides medios de cada clúster muestra los siguientes patrones:

- **Clúster 0:** asociado a distritos con mayor porcentaje de sobre-envejecimiento, alta edad media de la población y presencia relevante de hogares unipersonales de personas mayores (hombres solos >65 años). También destacan valores elevados en personas con nacionalidad extranjera.

Variable	$\Delta z_{c,j}$
Porcentaje de sobre-envejecimiento (Población mayor de 80 años/ Población mayor de 65 años)	0.37
Personas con nacionalidad extranjera	0.36
Edad media de la población	0.36
Hogares con un hombre mayor solo mayor de 65 años	0.33
Población mayor/igual de 25 años con FP 2º grado, Bachiller Superior o BUP	0.32

Tabla 36. Variables más influyentes en clúster 0 (distritos)

- **Clúster 1:** caracterizado por una mayor proporción de hogares monoparentales, elevada población en etapa educativa (0-16 años), así como altos niveles en los índices de vulnerabilidad económica y educativa. Este perfil refleja territorios más expuestos a problemáticas sociales vinculadas a la infancia, precariedad educativa y vulnerabilidad económica.

Variable	$\Delta z_{c,j}$
Hogares monoparentales: una mujer adulta con uno o más menores	1
Población en etapa educativa (Población de 0 a 16 años -16 no incluidos)	0.99
Índice de Vulnerabilidad Economía y Empleo	0.91
Índice de Vulnerabilidad Educación y Cultura	0.84
Población en etapa educativa de 6 a 11 años	0.73

Tabla 37. Variables más influyentes en clúster 1 (distritos)

- **Clúster 2:** agrupa los distritos con mayores niveles globales de vulnerabilidad, mostrando los valores más altos en los indicadores del índice IGUALA (vulnerabilidad educativa, territorial agregada, desempleo femenino, bajo nivel educativo, etc.). este clúster concentra los distritos estructuralmente más desfavorecidos.

Variable	$\Delta z_{c,j}$

Índice de Vulnerabilidad Educación y Cultura	1.34
Índice de Vulnerabilidad Territorial Agregado	1.33
Población mayor/igual de 25 años con Bachiller Elemental, Graduado Escolar, ESO, FP 1º grado	1.29
Población mayor/igual de 25 años con enseñanza primaria completa	1.22
Tasa de desempleo en mujeres de 25 a 44 años	1.22

Tabla 38. Variables más influyentes en clúster 2 (distritos)

- **Clúster 3:** recoge distritos más favorecidos, con población altamente cualificada (estudios superiores, diplomaturas e ingenierías), alta proporción de población española y mejores indicadores demográficos (baja dependencia y envejecimiento).

Variable	$\Delta z_{c,j}$
Población mayor/igual de 25 años con estudios superiores, licenciatura, arquitectura, ingeniería sup., estudios sup. no universitarios, doctorado, postgrado	0.95
Personas con nacionalidad española	0.66
Índice de dependencia (Población de 0-15 + población 65 años y más / Pob. 16-64)	0.63
Porcentaje de envejecimiento (Población mayor de 65 años/Población total)	0.58
Población mayor/igual de 25 años con titulación media, diplomatura, arquitectura o ingeniería técnica	0.53

Tabla 39. Variables más influyentes en clúster 3 (distritos)

En términos de influencia global de las variables en la formación de los clústeres, destacan como más determinantes:

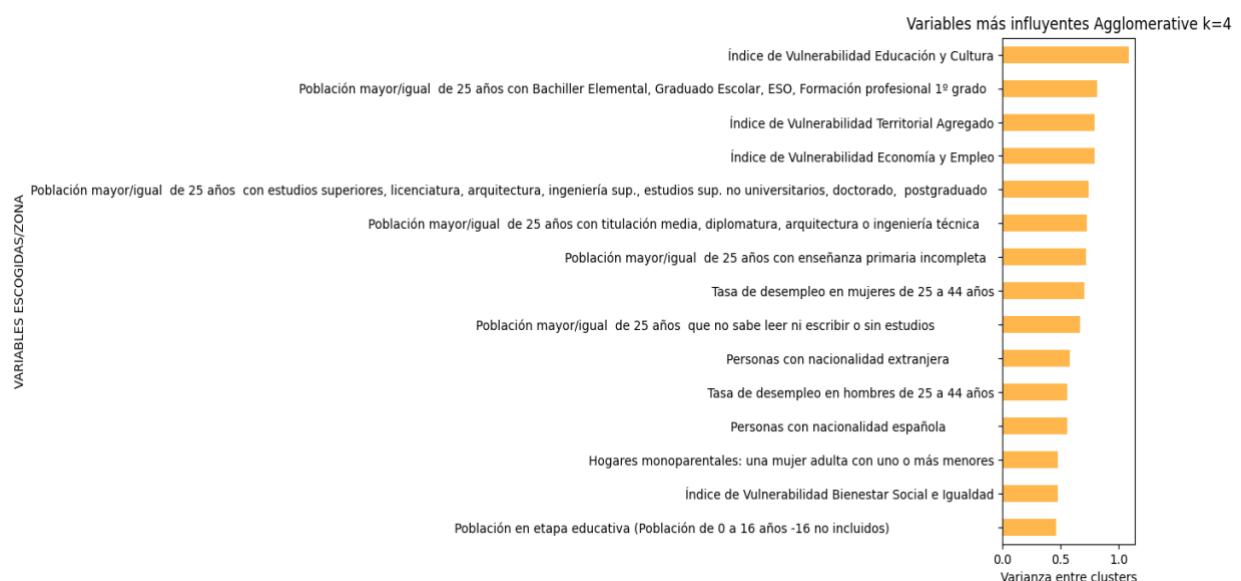


Imagen 37. Variables más influyentes en modelo Agglomerative k=4 (distritos)

Variable	Varianza entre clústeres
Índice de Vulnerabilidad Educación y Cultura	1.08
Índice de Vulnerabilidad Educación y Cultura	0.79
Vulnerabilidad Economía y Empleo	0.79
Nivel educativo (Bachiller, ESO, Primaria Incompleta...)	0.72-0.81
Nacionalidad (extranjera/española) y desempleo	0.55-0.7

Tabla 40. Resumen variables más influyentes en modelo Agglomerative k=4 (distritos)

Se confirma así el importante peso que tiene el componente educativo y económico en la configuración de los perfiles de vulnerabilidad a nivel distritos.

Por otro lado, la matriz de medias normalizadas muestra cómo algunos clústeres presentan valores claramente altos en indicadores clave como el Índice de Vulnerabilidad Territorial Agregado (1.3) o el de Educación y Cultura en el clúster 2 (1.4), confirmando que esas variables tienen un peso importante. Por el contrario, otros clústeres destacan por valores negativos en envejecimiento o dependencia, indicando perfiles más favorables. Como ya se ha explicado a nivel barrio, un valor más alto y positivo implica una mayor contribución de esa variable al perfil del clúster, mientras que uno negativo refleja menor influencia. De igual forma y como se puede apreciar, las diferencias entre clústeres respaldan la clasificación realizada previamente.

A continuación, se muestra la matriz de medias normalizadas para este modelo:

Indicadores

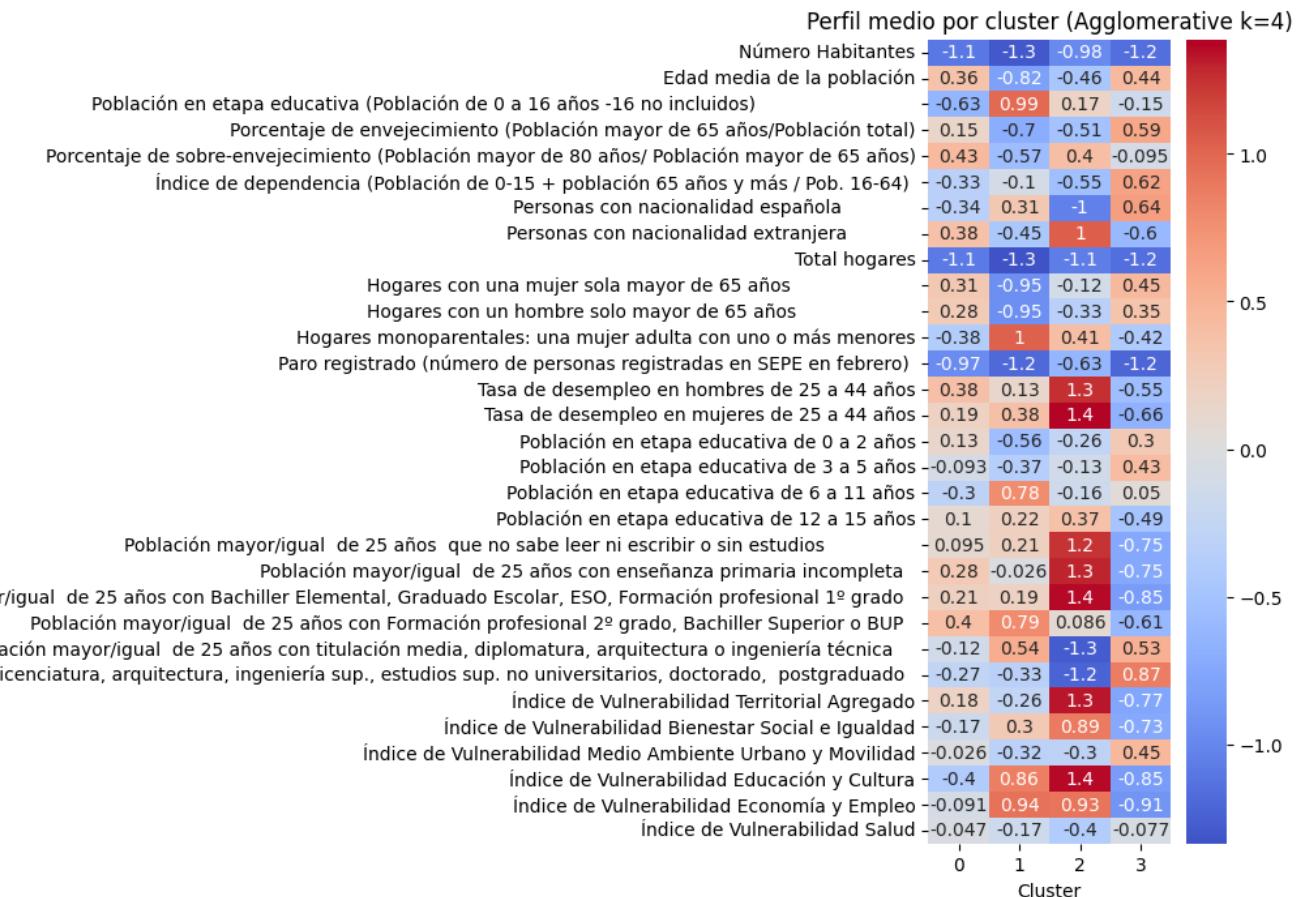
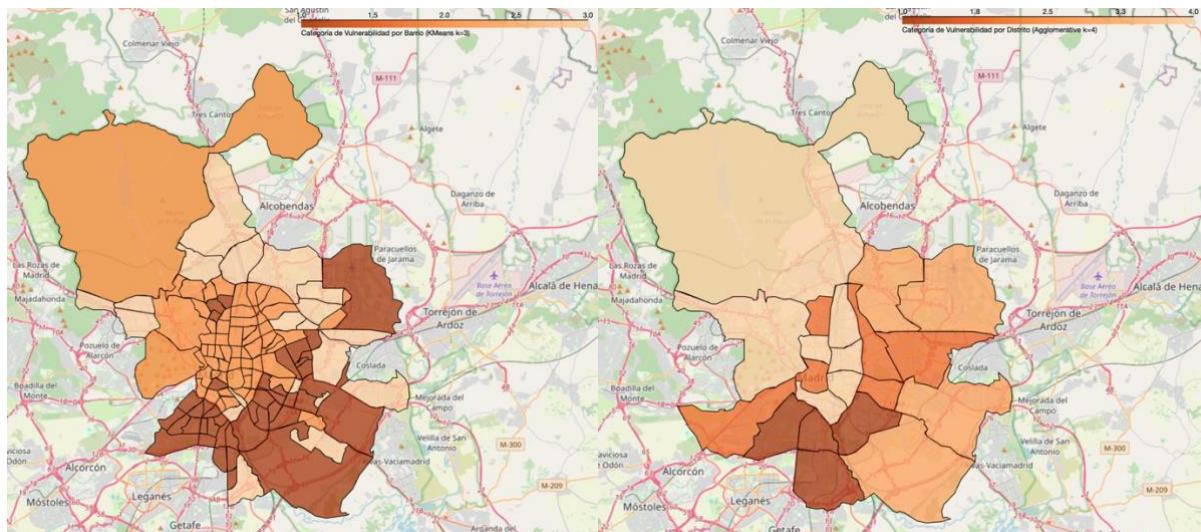


Imagen 38. Matriz de medias normalizadas modelo Agglomerative k=4 (distritos)

## 6.3. ANÁLISIS COMPARATIVO ENTRE NIVEL BARRIO Y NIVEL DISTRITO

Una vez definidos los modelos finales seleccionados – *K-means k = 3 para barrios* y *Agglomerative k = 4 para distritos* – se ha realizado una comparación visual y conceptual de los resultados obtenidos para ambos niveles territoriales.



Se observa que, a nivel barrios, el modelo permite identificar zonas pequeñas muy diferenciadas entre sí. En varios distritos aparecen barrios claramente vulnerables (en tonos oscuros) junto a otros barrios con valores mucho más bajos de vulnerabilidad (zonas claras). Es decir, dentro de un mismo distrito puede haber desigualdades internas que sólo se detectan cuando el análisis se realiza a nivel barrio.

Sin embargo, cuando se analiza la vulnerabilidad a nivel distrito, esta variabilidad se reduce, ya que el valor medio de los indicadores tiende a suavizar las diferencias internas. Algunos distritos que, como conjunto, aparecen con un nivel medio de vulnerabilidad, esconden en su interior barrios muy vulnerables o muy poco vulnerables, lo que queda oculto en el análisis agregado.

Por tanto, esto sugiere que el nivel de detalle del análisis influye en la capacidad para detectar desigualdades dentro de cada territorio. Mientras que análisis por barrios ofrece una visión mucho más precisa, el nivel distrito puede ser suficiente para una visión general pero escasa para otros fines.

## 6.4. RANKING FINAL DE VULNERABILIDAD

El ranking final para ambos niveles se ha pasado a una hoja de cálculo y se encuentra disponible en el Anexo A.



## 7. IMPLICACIONES ÉTICAS E IMPACTO SOCIAL

---

### 7.1. INTRODUCCIÓN

La ética, entendida como la reflexión crítica sobre nuestras acciones y decisiones, se convierte en un elemento central cuando el objeto de estudio involucra realidades humanas y complejas, como es el caso de la vulnerabilidad social.

La desigualdad urbana es una de las problemáticas sociales más relevantes de las grandes ciudades. En el caso de Madrid, las diferencias entre barrios en términos de acceso a la educación, empleo, sanidad o condiciones de vivienda pueden determinar en gran medida la calidad de vida de sus habitantes. Este proyecto tiene como objetivo identificar y analizar los factores más significativos que contribuyen a que un barrio sea considerado vulnerable, con el fin de facilitar una toma de decisiones más justa en la distribución de recursos públicos.

No obstante, un análisis de estas características, al tratar temas sensibles relacionados con la justicia social y el espacio urbano, implica asumir una responsabilidad ética que atraviesa todo el proceso metodológico. Por ello, este capítulo se centra en identificar los riesgos derivados del estudio y explicar cómo se han abordado, con el objetivo de asegurar que tanto el desarrollo como los resultados estén alineados con una práctica ética, respetuosa y orientada al bien común.

### 7.2. MARCOS NORMATIVOS Y PRINCIPIOS ÉTICOS DE REFERENCIA

El desarrollo de este proyecto se ha guiado por una serie de principios éticos y normativas legales que garantizan un tratamiento responsable de los datos y una orientación socialmente justa del análisis. Estas referencias han servido de base para la toma de decisiones metodológicas y para evaluar los posibles riesgos asociados a la investigación.

En primer lugar, se ha tomado como marco central el **Código de Ética y Conducta Profesional de la ACM (Association for Computing Machinery)** [28], ampliamente reconocido en el ámbito tecnológico. En concreto, los artículos 1.1 y 1.2 establecen la responsabilidad de actuar en beneficio del bienestar humano y de evitar cualquier daño potencial derivado del trabajo profesional. Además, el artículo 3.1 refuerza la idea de que el bien público debe situarse como prioridad central en cualquier desarrollo tecnológico, lo cual ha sido un principio rector en todas las fases del proyecto.

Junto a ello, en el plano legal, se ha tenido en cuenta la **Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de los Derechos Digitales** [29], que establece los principios de licitud, lealtad, transparencia y minimización del tratamiento de datos, incluso cuando estos no permiten la identificación directa de personas. Se ha trabajado exclusivamente con datos públicos, agregados y anonimizados, cumpliendo con los artículos 5 y 6 de dicha ley y garantizando en todo momento el respeto a la privacidad y la protección de la información territorial.

Asimismo, se han considerado otras fuentes de referencia internacional relevantes en el ámbito de la ética de los datos y la inteligencia artificial. Entre ellas destaca el documento **“Ethics Guidelines for Trustworthy AI”** del **Grupo de Expertos de Alto Nivel en Inteligencia Artificial de la Comisión Europea** [45], donde se recogen siete requisitos clave para un desarrollo ético de la IA, incluyendo la supervisión humana, la responsabilidad, la no discriminación y la inclusión social. Estas pautas han reforzado la necesidad de garantizar que cualquier herramienta basada en datos, incluso si no aplica algoritmos inteligentes complejos, respete los principios de equidad y transparencia.

También ha sido consultada la **“Recomendación sobre la Ética de la Inteligencia Artificial”** de la **UNESCO** [46], que subraya la necesidad de considerar el impacto social, económico y medioambiental de los sistemas automatizados, así como de promover la justicia, la diversidad y el desarrollo sostenible en el uso de tecnologías digitales. Este marco ha resultado especialmente útil al tratar un fenómeno como la vulnerabilidad territorial, que afecta de forma directa a los derechos y oportunidades de las personas y comunidades.

A nivel nacional, se ha revisado el Informe del **Comité de Bioética de España** [43] sobre el uso de la inteligencia artificial y los datos masivos, en el que se advierte sobre la necesidad de adoptar un enfoque de “precaución algorítmica” y de establecer mecanismos de evaluación del impacto ético en todo proyecto de análisis automatizado que pueda afectar al bienestar colectivo. También se ha tenido en cuenta la **Carta de Derechos Digitales** publicada por el Gobierno de España, que reconoce derechos clave como la no discriminación algorítmica, la transparencia en el uso de datos y el derecho a la neutralidad tecnológica, todos ellos alineados con el enfoque adoptado en este trabajo.

Finalmente, también se han considerado fuentes normativas y estadísticas del ámbito nacional para contextualizar los umbrales sociales utilizados en el análisis, especialmente en lo relativo a indicadores económicos como el nivel de renta. Por ejemplo, el salario mínimo interprofesional (SMI) para 2024 se establece en 1.134 euros mensuales, según el **Real Decreto 145/2024** publicado el 6 de febrero de 2024 [47]. Esta información, junto con los datos disponibles del Observatorio de la Vulnerabilidad Urbana del Ayuntamiento de Madrid y los indicadores sociales por distrito de la Comunidad de Madrid, han sido fundamentales para interpretar con mayor precisión las desigualdades territoriales detectadas en el estudio [48][49].

## 7.3. RIESGOS IDENTIFICADOS Y MEDIDAS ADOPTADAS

El análisis de la vulnerabilidad territorial en una ciudad como Madrid, mediante herramientas estadísticas y técnicas de análisis de datos, implica una serie de riesgos éticos, técnicos y sociales que han sido identificados y gestionados de manera proactiva a lo largo del proyecto.

Uno de los principales riesgos detectados ha sido el de **reforzar estigmatizaciones o prejuicios preexistentes** sobre ciertos barrios o zonas urbanas. La clasificación de un área como “vulnerable” puede tener consecuencias negativas si es interpretada sin contexto, contribuyendo a alimentar estereotipos o a marginar aún más a las comunidades que viven en esas zonas. Para minimizar este riesgo se ha evitado utilizar un lenguaje peyorativo o determinista, se ha priorizado una presentación visual de los resultados que favorezca la comprensión y la no comparación directa, y se ha acompañado cada visualización de una interpretación crítica y reflexiva.

Otro riesgo relevante está relacionado con el **uso instrumental o políticos de los resultados**, en contextos que escapan al control del proyecto. Los datos generados podrían ser utilizados para justificar decisiones urbanísticas poco equitativas, procesos de gentrificación o redistribuciones de recursos que no respondan al principio de justicia social. Con el fin de prevenir este tipo de derivadas, se ha hecho un esfuerzo explícito por contextualizar el análisis como una herramienta orientada al bien común y a la mejora de la planificación urbana. En este sentido, se ha subrayado que la vulnerabilidad es un fenómeno estructural y multifactorial, y no una característica intrínseca de determinados territorios.

Desde el punto de vista técnico, otro riesgo ha sido la **possible sobreinterpretación de resultados derivados de modelos matemáticos**, especialmente en lo que respecta al *clustering* o a la creación de índices sintéticos. Aunque estas herramientas permiten identificar patrones útiles, también implican simplificaciones de la realidad que deben ser interpretadas con precaución. Para abordar esta cuestión, se ha optado por comparar distintos algoritmos de agrupamiento (*K-means*, DBSCAN, *Agglomerative*), analizando sus límites y contrastando los resultados mediante visualizaciones con PCA y perfiles medios. Además, se ha añadido un índice de vulnerabilidad compuesto a partir de variables consideradas clave, con el objetivo de facilitar la lectura sin reducir la complejidad del fenómeno.

En el ámbito de los datos, se ha considerado también el riesgo de trabajar con información desactualizada o incompleta, dado que muchas de las fuentes utilizadas provienen de estadísticas municipales agregadas. Para reducir esta incertidumbre, se han seleccionado variables con bajo porcentaje de valores nulos y se han mantenido aquellas cuya presencia fuera representativa tanto a nivel de distrito como de barrio, priorizando siempre la transparencia y la trazabilidad de las decisiones tomadas durante la limpieza y transformación de los datos.

Por último, desde un enfoque ético, se ha tenido en cuenta la **posibilidad de invisibilizar la diversidad de situaciones dentro de cada zona geográfica**. Un mismo distrito puede contener realidades socioeconómicas muy dispares que no quedan reflejadas en los datos agregados. Por ello, se ha optado por una estructura analítica flexible, que permite tanto el

análisis a nivel de barrio como la agregación por distrito, y se ha incluido una advertencia explícita sobre los límites del enfoque utilizado.

## 7.4. MATRIZ DE RIESGOS DEL PROYECTO

Código	Descripción del riesgo	Tipo	Probabilidad	Impacto	Medidas adoptadas
R-01	Estigmatización de barrios considerados vulnerables	Ético / Social	Media	Alto	Uso de lenguaje neutro, anonimización, visualizaciones explicativas y no jerárquicas
R-02	Uso político o interesado de los resultados fuera del contexto del proyecto	Ético / Político	Media	Alto	Advertencias explícitas en la memoria, contextualización crítica de los resultados
R-03	Pérdida de diversidad interna al agregar datos a nivel de distrito	Técnico / Ético	Alta	Medio	Análisis también a nivel de barrio cuando es posible, y mención expresa de la limitación
R-04	Simplificación excesiva en la construcción de índices	Técnico / Metodológico	Media	Medio	Comparación de varios modelos de <i>clustering</i> , validación visual con PCA, índice construido con variables clave
R-05	Baja calidad o desactualización de algunos datos públicos	Técnico	Media	Medio	Selección de variables con menor porcentaje de valores nulos, limpieza cuidadosa de la base de datos

R-06	Falta de comprensión crítica de resultados por parte del lector no especializado	Social / Comunicación	Media	Alto	Inclusión de interpretaciones claras, mapas interactivos y explicaciones detalladas en el cuerpo del trabajo
R-07	Errores de formato o codificación en el tratamiento de datos numéricos	Técnico	Baja	Bajo	Aplicación de funciones de limpieza, conversión estandarizada y control de errores en los datos

Tabla 41. Matriz descriptiva de riesgos del proyecto

Código	Probabilidad	Impacto	Nivel del riesgo
R-01	Media	Alto	Alto
R-02	Media	Alto	Alto
R-03	Alta	Medio	Alto
R-04	Media	Medio	Medio
R-05	Media	Medio	Medio
R-06	Media	Alto	Alto
R-07	Baja	Bajo	Bajo

Tabla 42. Matriz resumen de riesgos del proyecto

## 7.5. IMPACTO SOCIAL ESPERADO

El análisis de la vulnerabilidad territorial que se desarrolla en este proyecto tiene como propósito último generar conocimiento útil para avanzar hacia una distribución más justa y equitativa de los recursos públicos en la ciudad de Madrid. En este sentido, el impacto social esperado se manifiesta en varias dimensiones complementarias.

En primer lugar, el proyecto busca **visibilizar desigualdades urbanas estructurales** que habitualmente permanecen diluidas en los discursos generales sobre el desarrollo urbano. Al identificar, mediante datos objetivos, los factores que hacen que ciertos barrios presenten mayores niveles de vulnerabilidad, se pretende aportar evidencia que contribuya a la **toma de decisiones políticas y técnicas más informadas**, orientadas a corregir desequilibrios en el acceso a servicios públicos, oportunidades educativas, empleo o condiciones de vida dignas.

Además, este trabajo tiene una clara vocación de servicio público en tanto que **pone al alcance de instituciones y ciudadanía una herramienta de diagnóstico basada en datos abiertos**, reutilizable y comprensible. Los resultados, especialmente aquellos representados visualmente en forma de mapas o rankings interpretativos, permiten facilitar la comprensión de la realidad urbana a responsables municipales, técnicos de planificación, organizaciones sociales o cualquier colectivo interesado en promover la justicia territorial.

Por otra parte, se espera que el proyecto tenga también un **efecto pedagógico y sensibilizador**. Al ofrecer un análisis detallado y crítico de las condiciones de los barrios madrileños, se contribuye a generar una ciudadanía más informada y empática con las realidades ajena. Esto es especialmente relevante en un contexto donde las desigualdades territoriales tienden a naturalizarse o invisibilizarse. La información generada puede ayudar a combatir estigmas, cuestionar prejuicios e impulsar una mayor corresponsabilidad colectiva en la defensa del derecho a la ciudad.

Desde una perspectiva metodológica, este estudio puede servir como **modelo replicable** para otras ciudades o contextos. Su diseño modular, basado en técnicas de minería de datos y visualización aplicadas a fuentes oficiales, permite adaptar fácilmente la metodología a otros entornos urbanos, reforzando así su utilidad pública.

Finalmente, cabe destacar el **potencial transformador del análisis cuando se aplica con responsabilidad**. Este proyecto no pretende ser una herramienta de clasificación o etiquetado de barrios, sino un punto de partida para repensar la ciudad desde los datos y al servicio de las personas. En este marco, se considera que el principal impacto social radica en su capacidad de **generar conversación, activar conciencia y movilizar decisiones más justas** para la construcción de un Madrid más inclusivo.

## 7.6. LIMITACIONES Y POSIBLES LÍNEAS FUTURAS DEL PROYECTO

Este proyecto, como cualquier análisis basado en datos abiertos, presenta limitaciones. Una de las más destacadas es la **escasa disponibilidad de ciertos indicadores a nivel de barrio**, lo que obliga en ocasiones a utilizar datos referidos únicamente a distritos. Esto puede ocultar desigualdades internas y limitar la precisión del análisis. Otra limitación es la necesidad de **reducir el volumen de variables** originales (más de 280), lo cual puede haber dejado fuera indicadores relevantes. Aun así, se han aplicado criterios temáticos y técnicos sólidos para garantizar la validez del análisis.

En el futuro, el proyecto podría ampliarse incluyendo nuevas fuentes de datos, desarrollando indicadores dinámicos, o integrando herramientas más avanzadas de IA, siempre respetando los principios éticos que lo han guiado desde el inicio.

## 7.7. CONCLUSIÓN

El análisis ético y social realizado en este capítulo permite afirmar que el proyecto desarrollado ha mantenido en todo momento un compromiso riguroso con los principios fundamentales de respeto, equidad y responsabilidad. El uso de datos públicos, la transparencia en la metodología aplicada y la constante reflexión sobre los efectos sociales del trabajo han permitido garantizar la integridad del estudio desde una perspectiva ética.

### 7.7.1. Viabilidad ética del proyecto

Desde el punto de vista ético, el proyecto es plenamente viable. Se han tenido en cuenta los marcos normativos vigentes, tanto a nivel nacional como internacional, y se han aplicado medidas de mitigación frente a todos los riesgos identificados. El diseño del análisis, la gestión de los datos, la interpretación de resultados y la comunicación visual se han alineado con el objetivo de no causar daño, respetar la diversidad y contribuir al bienestar colectivo.

El principio de precaución se ha aplicado especialmente en el uso de técnicas estadísticas, evitando la sobreinterpretación de modelos matemáticos y subrayando en todo momento que la vulnerabilidad es un fenómeno complejo, estructural y multifactorial. El proyecto ha buscado siempre visibilizar desigualdades sin reforzar estigmas, y generar conocimiento útil para una toma de decisiones más justa.

Además de su viabilidad, el análisis ético permite destacar otros aspectos relevantes del proyecto:

- Su valor como herramienta de diagnóstico accesible, replicable y orientada al bien común.
- Su contribución a la cultura del dato y la alfabetización estadística en clave de justicia social.

- Su potencial para sensibilizar a la ciudadanía sobre las desigualdades territoriales y fomentar una mirada más crítica y empática hacia el entorno urbano.
- La aplicación de la ética como un componente transversal del proceso de análisis, y no como un añadido final, lo que refuerza la calidad y coherencia del trabajo.

En relación con todo lo expuesto, dado el valor que presenta el proyecto y que los riesgos detectados han sido identificados y abordados de forma adecuada mediante medidas técnicas y metodológicas concretas, **podemos concluir que el proyecto no solamente es viable desde el punto de vista ético, sino que además es recomendable llevarlo a cabo**. Sus aportaciones pueden resultar especialmente relevantes para la planificación urbana, la formulación de políticas públicas más equitativas y la generación de conciencia colectiva en torno a la desigualdad.

## 8. CONCLUSIONES

---

A lo largo del presente trabajo se ha desarrollado una solución técnica para el análisis de la vulnerabilidad territorial en la ciudad de Madrid, aplicando técnicas de *clustering* no supervisado sobre un conjunto multivariable de indicadores demográficos, educativos, económicos y sociales.

El modelo desarrollado permite identificar agrupamientos diferenciados tanto a nivel de distritos como de barrios, destacando patrones territoriales coherentes con la estructura urbana de la ciudad. La aplicación de métricas de validación interna (*Silhouette Score* y *Calinski-Harabasz Score*) ha permitido seleccionar de forma fundamentada el número óptimo de clústeres en cada caso, optando finalmente por:

- *K-means* con  $k = 3$  para barrios, por su mayor estabilidad y coherencia visual de los agrupamientos.
- *Agglomerative Clustering* con  $k = 4$  para distritos, que ofrece una segmentación más granular adecuada al menor tamaño muestral.

Los análisis de importancia de variables confirman que las dimensiones relacionadas con el nivel educativo, el empleo, la estructura familiar y los índices agregados de vulnerabilidad (especialmente el índice IGUALA) son los principales determinantes en la diferenciación de los perfiles territoriales. Además, se ha comprobado que, en la mayoría de los casos, existe una relación consistente entre los resultados obtenidos a nivel barrio y los correspondientes a nivel distrito, aportando robustez a la metodología propuesta.

Como posibles líneas de evolución y mejora del presente trabajo se identifican las siguientes:

- Incorporación de nuevas fuentes de datos, incluyendo variables complementarias en dimensiones de salud mental, criminalidad, movilidad o accesibilidad.
- Análisis incluyendo series temporales para detectar la evolución de la vulnerabilidad territorial a lo largo del tiempo.
- Aplicación de técnicas avanzadas de *clustering* mixto o técnicas de reducción no lineal de dimensionalidad para captar estructuras complejas.
- Desarrollo de una herramienta visual interactiva que permita explorar dinámicamente los mapas generados.

- Validación externa del modelo a partir de información empírica complementaria de servicios sociales y políticas de intervención aplicadas.

## 9. OTROS MÉRITOS DEL PROYECTO

---

Además de los objetivos inicialmente planteados, el proyecto ha generado una serie de resultados adicionales que aportan un valor añadido:

- **Reutilización y adaptabilidad del sistema:** La estructura modular del código desarrollado (Python, scikit-learn, geopandas, folium, etc.) permite su reutilización directa para futuros análisis de vulnerabilidad en otros territorios o con nuevas bases de datos.
- **Uso de software libre y tecnologías open source:** Todo el procesamiento, modelado y visualización se ha realizado utilizando herramientas libres ampliamente aceptadas en la comunidad científica, facilitando su replicabilidad y accesibilidad.
- **Integración multidisciplinar:** El proyecto combina técnicas de minería de datos, análisis territorial, geovisualización y políticas públicas, integrando distintas disciplinas del ámbito socioeconómico, geográfico y tecnológico.
- **Representación interactiva de los resultados:** La generación de mapas interactivos en formato HTML permite una exploración visual dinámica de los niveles de vulnerabilidad por zonas, ofreciendo una herramienta de gran potencial para la toma de decisiones políticas o la comunicación pública.
- **Possible explotación pública de los resultados:** Dado el interés social del objeto de estudio, los resultados podrían ser compartidos como recurso de acceso abierto para su uso en el ámbito académico, institucional o ciudadano



## 10. BIBLIOGRAFÍA

---

- [1] "Panel de indicadores de distritos y barrios de Madrid. Estudio sociodemográfico - Portal de datos abiertos del Ayuntamiento de Madrid". En portada - Portal de datos abiertos del Ayuntamiento de Madrid. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=71359583a773a510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>
- [2] Ministerio de Transportes, Movilidad y Agenda Urbana, "Atlas de Vulnerabilidad Urbana," [En línea]. Accedido el 27 de enero de 2025. Disponible: <https://atlasvulnerabilidadurbana.mitma.es/#view=map1&c=indicator>
- [3] "Ministerio de Sanidad, Consumo y Bienestar Social - Portal Estadístico del SNS - Encuesta Nacional de Salud de España 2017". Ministerio de Sanidad. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://www.sanidad.gob.es/estadEstudios/estadisticas/encuestaNacional/encuesta2017.htm>
- [4] "Panel de Hogares de la ciudad de Madrid - Ayuntamiento de Madrid". Inicio - Ayuntamiento de Madrid. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://www.madrid.es/portales/munimadrid/es/Inicio/Servicios-sociales-y-salud/Servicios-sociales/Publicaciones/Panel-de-Hogares-de-la-ciudad-de-Madrid/?vgnextfmt=default&vgnextoid=6224347e00251810VgnVCM2000001f4a900aRCRD&vgnextchannel=2a26c8eb248fe410VgnVCM1000000b205a0aRCRD>
- [5] Ayuntamiento de Madrid (2023). *Informe sobre desigualdades socioeconómicas entre distritos y barrios*.
- [6] "INE - Instituto Nacional de Estadística". INE. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://www.ine.es/>
- [7] C. Hennig y T. F. Liao, "How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification", J. Roy. Statistical Soc.: Ser. C (Appl. Statist.), vol. 62, n.º 3, pp. 309–369, abril de 2013. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://doi.org/10.1111/j.1467-9876.2012.01066.x>

- [8] I. Brauksa, "Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia", *J. Econ., Bus. Manage.*, pp. 42–45, 2013. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://doi.org/10.7763/joebm.2013.v1.10>
- [9] S. Vyas y L. Kumaranayake, "Constructing socio-economic status indices: how to use principal components analysis", *Health Policy Plan.*, vol. 21, n.º 6, pp. 459–468, agosto de 2006. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://doi.org/10.1093/heapol/czl029>
- [10] R. Iglesias-Pascual, F. Benassi y C. Hurtado-Rodríguez, "Social infrastructures and socio-economic vulnerability: A socio-territorial integration study in Spanish urban contexts", *Cities*, vol. 132, p. 104109, enero de 2023. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://doi.org/10.1016/j.cities.2022.104109>
- [11] G. Piasek, I. Fernández Aragón, J. Shershneva y P. Garcia-Almirall, "Assessment of Urban Neighbourhoods' Vulnerability through an Integrated Vulnerability Index (IVI): Evidence from Barcelona, Spain", *Social Sci.*, vol. 11, n.º 10, p. 476, octubre de 2022. Accedido el 17 de noviembre de 2024. [En línea]. Disponible: <https://doi.org/10.3390/socsci11100476>
- [12] "Atlas of novel social and urban vulnerability in Spain - 300.000 Km/s". 300.000 Km/s. Accedido el 3 de febrero de 2025. [En línea]. Disponible: [https://300000kms.net/case\\_study/atlas-of-novel-social-and-urban-vulnerability-in-spain/](https://300000kms.net/case_study/atlas-of-novel-social-and-urban-vulnerability-in-spain/)
- [13] P. Garcia-Almirall, C. Cornadó y S. Vima-Grau, "Residential Vulnerability of Barcelona: Methodology Integrating Multi-Criteria Evaluation Systems and Geographic Information Systems", *Sustainability*, vol. 13, n.º 24, p. 13659, diciembre de 2021. Accedido el 3 de febrero de 2025. [En línea]. Disponible: <https://doi.org/10.3390/su132413659>
- [14] "El indice de vulnerabilidad\_CAS\_2024\_def\_v2 (Versión 1.0)". iseい-ivei. Accedido el 30 de enero de 2025. [En línea]. Disponible: [https://isei-ivei.euskadi.eus/documents/d/guest/el-indice-de-vulnerabilidad\\_cas\\_2024\\_def\\_v2?imagePreview=1](https://isei-ivei.euskadi.eus/documents/d/guest/el-indice-de-vulnerabilidad_cas_2024_def_v2?imagePreview=1)
- [15] Gómez, J. A. (2006). VIII. barrios desfavorecidos: diagnóstico de la situación española. *Exclusión social y estado de bienestar en España*, 5, 155.
- [16] J. Alguacil Gómez, J. Camacho Gutiérrez, y A. Hernández Aja, "La vulnerabilidad urbana en España. Identificación y evolución de los barrios vulnerables," *EMPIRIA: Revista de Metodología de Ciencias Sociales*, vol. 27, pp. 73–94, 2014.
- [17] C. Borrell, M. I. Pasarín, E. Díez, K. Pérez, D. Malmusi, G. Pérez, y L. Artazcoz, "Las desigualdades en salud como prioridad política en Barcelona," *Gaceta Sanitaria*, vol. 34, pp. 69–76, 2020. [En línea]. Disponible en: [CrossRef] [PubMed].
- [18] P. Bourdieu, "Capital symbolique et classes sociales," *L'Arc*, vol. 72, pp. 13–19, 1978.
- [19] R. Castel, "La dinámica de los procesos de marginalización: De la vulnerabilidad a la exclusión," en *El Espacio Institucional*, M. J. Acevedo y J. C. Volnovich, Eds. Buenos Aires: Lugar Editorial, 1991, pp. 37–54.

- [20] A. Hernández-Aja, “Áreas vulnerables en el centro de Madrid,” *Cuadernos de Investigación Urbanística*, no. 53, Instituto Juan de Herrera, Madrid, 2007. [En línea]. Disponible en: <http://polired.upm.es/index.php/ciur/article/view/268/263>. [Accedido: 19-sep-2022].
- [21] R. Castel, *Las metamorfosis de la cuestión social. Una crónica del salariado*, Barcelona: Paidós, 1995.
- [22] S. Castles y M. Miller, *The Age of Migration*, Londres: Palgrave, 1998.
- [23] R. Chambers, “Vulnerability: How the Poor Cope?,” *IDS Bulletin*, vol. 20, no. 2, Sussex: Institute of Development Studies, 1989. [En línea]. Disponible en: <https://bulletin.ids.ac.uk/index.php/idsbo/issue/view/138>. [Accedido el 3 de febrero de 2025].
- [24] J. Checa y O. Nello, “La segregación residencial y condiciones de vida. Un análisis de las desigualdades sociales en Catalunya a partir de cuatro perspectivas espaciales,” en *La reconfiguración capitalista de los espacios urbanos: Transformaciones y desigualdades*, J. M. Parreño Castellano y C. Moreno Medina, Eds. Gran Canaria: Universidad de Las Palmas de Gran Canaria, 2021, pp. 185–206.
- [25] F. Ferrando, M. Hernández-Almeida, C. Oreiro, M.-N. Seijas y J. Urraburu, “Evolución de la Segregación Socioeconómica en la Educación Pública de Uruguay”, *REICE. Rev. Iberoam. Sobre Calid., Efic. Cambio En Educ.*, vol. 18, n.º 4, pp. 143–169, septiembre de 2020. Accedido el 30 de enero de 2025. [En línea]. Disponible: <https://doi.org/10.15366/reice2020.18.4.006>
- [26] S. De la Fuente Fernández, *Análisis factorial*, Facultad de Ciencias Económicas y Empresariales, Univ. Autónoma de Madrid, Madrid, 2011. [En línea]. Disponible en: [https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/FACTORIA\\_L/analisis-factorial.pdf](https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/FACTORIA_L/analisis-factorial.pdf). [Accedido el 3 de febrero de 2025].
- [27] A. Echaves García y C. Echaves, “Jóvenes aún más precarios, crisis económica y desigualdad laboral en España,” *Cuadernos de Investigación en Juventud*, vol. 2, pp. 33–52, 2017. [En línea]. Disponible en: <https://dialnet.unirioja.es/servlet/articulo?codigo=5873997>. [Accedido el 3 de febrero de 2025].
- [28] “Código de Ética y Conducta Profesional de ACM,” 2018. Accessed: Apr. 14, 2025. [Online]. Disponible en: <https://www.acm.org/about-acm/code-of-ethics-in-spanish>
- [29] J. Del Estado, “Disposición 16673 del BOE núm. 294 de 2018,” 2018. [Online]. Disponible: <https://www.boe.es/eli/es/lo/2018/12/05/3/con>
- [30] Ayuntamiento de Madrid, “Geoportal IDEAM – Datos cartográficos de distritos,” [En línea]. Disponible en: [https://geoportal.madrid.es/IDEAM\\_WBGEOPORTAL/dataset.iam?id=422fa235-762b-11e9-861d-ecb1d753f6e8](https://geoportal.madrid.es/IDEAM_WBGEOPORTAL/dataset.iam?id=422fa235-762b-11e9-861d-ecb1d753f6e8). [Accedido: 23-abr-2025].

- [31] Code for Germany, “Madrid districts – click\_that\_hood,” GitHub, [En línea]. Disponible en: [https://github.com/codeforgermany/click\\_that\\_hood/blob/main/public/data/madrid-districts.geojson](https://github.com/codeforgermany/click_that_hood/blob/main/public/data/madrid-districts.geojson). [Accedido: 23-abr-2025].
- [32] Ayuntamiento de Madrid, “Padrón municipal – Datos estadísticos,” [En línea]. Disponible en: <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default&vgnextoid=d029ed1e80d38610VgnVCM2000001f4a900aRCRD>. [Accedido: 23-abr-2025].
- [33] J. M. Moine, “Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo”, tesis, 2013. Accedido el 23 de abril de 2025. [En línea]. Disponible: <http://hdl.handle.net/10915/29582>
- [34] J. M. Moine, A. S. Haedo, y S. E. Gordillo, “Estudio comparativo de metodologías para minería de datos,” en *XIII Workshop de Investigadores en Ciencias de la Computación*, 2011.
- [35] J. M. Moine, S. E. Gordillo, y A. S. Haedo, “Análisis comparativo de metodologías para la gestión de proyectos de minería de datos,” en *XVII Congreso Argentino de Ciencias de la Computación (CACIC 2011)*, 2011.
- [36] C. L. Hernández G and M. R. Ximena Dueñas, “Hacia una metodología de gestión del conocimiento basada en minería de datos.”
- [37] Gavilan, I. (2021). *Metodología para Machine Learning (III): SEMMA*. Recuperado de: <https://ignaciogavilan.com/metodologia-para-machine-learning-iii-semma/>
- [38] K. Jordahl, *Geopandas documentation*, 2016. [En línea]. Disponible en: <https://app.readthedocs.org/projects/geopandas-doc/downloads/pdf/latest/> . [Accedido: 23-abr-2025].
- [39] T. E. Oliphant, *Guide to NumPy*, vol. 1, p. 85, USA: Trelgol Publishing, 2006. [En línea]. Disponible en: <https://ecs.wgtn.ac.nz/foswiki/pub/Support/ManualPagesAndDocumentation/numpybook.pdf>. [Accedido: 23-abr-2025].
- [40] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [En línea]. Disponible en: <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>. [Accedido: 23-abr-2025].
- [41] E. Bisong, “Matplotlib and Seaborn,” en *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. [sin pag.]. [En línea]. Disponible en: [https://doi.org/10.1007/978-1-4842-4470-8\\_12](https://doi.org/10.1007/978-1-4842-4470-8_12). [Accedido: 23-abr-2025].
- [42] J. Hunt, “Working with Excel Files,” en *Advanced Guide to Python 3 Programming*, ser. Undergraduate Topics in Computer Science, Cham: Springer, 2019. [En línea]. Disponible en: [https://doi.org/10.1007/978-3-030-25943-3\\_21](https://doi.org/10.1007/978-3-030-25943-3_21). [Accedido: 24-abr-2025].

- [43] Comité de Bioética de España, *Recomendaciones del Comité de Bioética de España sobre el impacto ético del uso de herramientas algorítmicas en el ámbito social y urbano*, 2023. [En línea]. Disponible en: [https://comitedebioetica.isciii.es/wp-content/uploads/2025/04/Informe\\_EEDS.pdf](https://comitedebioetica.isciii.es/wp-content/uploads/2025/04/Informe_EEDS.pdf). [Accedido: 23-abr-2025].
- [44] Gobierno de España, *Carta de Derechos Digitales*, 2021. [En línea]. Disponible en: <https://derechodigital.pre.red.es/>. [Accedido: 24-abr-2025].
- [45] High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019. [En línea]. Disponible en: [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419). [Accedido: 24-abr-2025].
- [46] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 2021. [En línea]. Disponible en: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. [Accedido: 24-abr-2025].
- [47] Gobierno de España, “Real Decreto 145/2024, por el que se fija el salario mínimo interprofesional para 2024,” BOE, núm. 31, 6 de febrero de 2024. [En línea]. Disponible en: <https://www.boe.es/eli/es/rd/2024/02/06/145>
- [48] Ayuntamiento de Madrid, “Observatorio de la Vulnerabilidad Urbana. Indicadores Sociales por Barrio y Distrito,” Área de Gobierno de Políticas Sociales, Familia e Igualdad. [En línea]. Disponible: <https://datos.madrid.es/portal/site/egob>
- [49] Comunidad de Madrid, “Indicadores Municipales de Bienestar Social,” Consejería de Familia, Juventud y Asuntos Sociales. [En línea]. Disponible: <https://www.comunidad.madrid/servicios/servicios-sociales/indicadores-sectores-atencion-social>
- [50] A. Goel et al., “Improved Twitter Sentiment Prediction through Cluster-Then-Predict Model”, *arXiv preprint*, 2015. [En línea]. Disponible: <https://arxiv.org/abs/1509.02437>
- [51] P. Liang et al., “A Rescaled Cluster-Then-Predict Approach for Enhanced Credit Scoring,” *Journal of Financial Data Science*, 2023. [En línea]. Disponible: <https://www.sciencedirect.com/science/article/pii/S1057521923005215>
- [52] S. Kim et al., “Benchmarking Cluster-Then-Predict Models to Challenge Prevailing Predictive Modeling Strategies,” *University of Hawaii ScholarSpace*, 2022. [En línea]. Disponible: <https://scholarspace.manoa.hawaii.edu/items/6cd8108a-4767-451d-a6f3-f115c8c5a40f>
- [53] C. Fraley and A. E. Raftery, “Model-Based Clustering, Discriminant Analysis, and Density Estimation”, *Journal of the American Statistical Association*, vol. 97, no. 458, 2002. [En línea]. Disponible: <https://www.tandfonline.com/doi/abs/10.1198/016214502760047131>

[54] IBM, “What is Mixture of Experts?”, *IBM Knowledge Center*. [En línea]. Disponible: <https://www.ibm.com/think/topics/mixture-of-experts>

[55] IBM, “TwoStep Cluster Node”, *IBM SPSS Modeler Documentation*. [En línea]. Disponible: <https://www.ibm.com/docs/en/spss-modeler/18.2.2>

[56] Ayuntamiento de Madrid, *Portal de Datos Abiertos*, [En línea]. Disponible en: <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=71359583a773a510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

[57] Ayuntamiento de Madrid, *Portal de Datos Abiertos*, Ranking de vulnerabilidad de los distritos y barrios de Madrid [En línea]. Disponible en: <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=d029ed1e80d38610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

[58] G. Fernández Maíllo, *VIII Informe sobre exclusión y desarrollo social en España*, Madrid: Fundación FOESSA, 2019. Disponible en: [https://www.foessa.es/main-files/uploads/sites/16/2019/06/Informe-FOESSA-2019\\_web-completo.pdf](https://www.foessa.es/main-files/uploads/sites/16/2019/06/Informe-FOESSA-2019_web-completo.pdf)

[59] Ayuntamiento de Madrid, *IGUALA. Índice de vulnerabilidad territorial agregado del Ayuntamiento de Madrid*, 2022. [En línea]. Disponible en: <https://iguala.madrid.es>

[60] Ayuntamiento de Madrid, *Portal de Datos Abiertos - Indicadores sociales, demográficos y económicos de Madrid*, 2024. [En línea]. Disponible en: <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=bb71fa5c48a0810VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

[61] Ayuntamiento de Madrid, *Estadística de Padrón municipal de habitantes por distritos y barrios*, [En línea]. Disponible en: [https://servpub.madrid.es/CSEBD\\_WBINTER/seleccionSerie.html?numSerie=030201020023\\_2](https://servpub.madrid.es/CSEBD_WBINTER/seleccionSerie.html?numSerie=030201020023_2)

[62] Ayuntamiento de Madrid, *Distritos y barrios – Igualdad Madrid*, [En línea]. Disponible en: <https://iguala.madrid.es/pages/distritos-y-barrios>

[63] Madrid Salud, *Estudio de Salud de la Ciudad de Madrid 2022 – Resumen Ejecutivo*, 2022. [En línea]. Disponible en: <https://madridsalud.es/publicacioness/estudio-de-salud-de-la-ciudad-de-madrid-2022-resumen-ejecutivo/>

[64] Ayuntamiento de Madrid, *Informe de Resultados Interdistrital – Encuesta de Calidad de Vida y Satisfacción con los Servicios Públicos de la Ciudad de Madrid. Edición 2023 – Nivel Distrito*, 2023. [En línea]. Disponible en: [https://www.madrid.es/UnidadesDescentralizadas/Calidad/Observatorio\\_Ciudad/06\\_S\\_Per](https://www.madrid.es/UnidadesDescentralizadas/Calidad/Observatorio_Ciudad/06_S_Per)

[pcion/EncuestasCalidad/EncuestaMadrides/ficheros/2023/Distritos/Informe\\_Interdistrital.pdf](#)

[65] Ayuntamiento de Madrid, *Paro registrado por mes y sexo según distrito y barrio (2017–2024)*, Serie estadística 4.1.C, Área de Información Estadística: Mercado de Trabajo. [En línea]. Disponible en: [https://servpub.madrid.es/CSEBD\\_WBINTER/seleccionSerie.html?numSerie=090404000001](https://servpub.madrid.es/CSEBD_WBINTER/seleccionSerie.html?numSerie=090404000001) 3

[66] Ayuntamiento de Madrid, *Centros Deportivos Municipales (Polideportivos)*, [En línea]. Disponible en: <https://datos.madrid.es/sites/v/index.jsp?vgnnextoid=4a5fbef4b2503410VgnVCM2000000c205a0aRCRD&vgnnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD>

[67] Ayuntamiento de Madrid, *Población en general – Servicios Sociales. Áreas de información estadística*, [En línea]. Disponible en: <https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Servicios-sociales/Poblacion-en-general/?vgnextfmt=default&vgnnextoid=487d56a3a2f59210VgnVCM2000000c205a0aRCRD&vgnnextchannel=34dfa6360e73a210VgnVCM1000000b205a0aRCRD>

[68] Ayuntamiento de Madrid, *Policía Municipal – Datos estadísticos de actuaciones*, [En línea]. Disponible en: <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnnextoid=bffff1d2a9fdb410VgnVCM2000000c205a0aRCRD&vgnnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

[69] IBM, *¿Qué es una arquitectura de datos?*, [En línea]. Disponible en: <https://www.ibm.com/es-es/topics/data-architecture>

[70] IBM, *¿Qué es un Data Warehouse?*, [En línea]. Disponible en: <https://www.ibm.com/es-es/topics/data-warehouse>

[71] H. Wickham and G. Grolemund, *R for Data Science*, O'Reilly Media, 2016.

[72] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.

[73] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2011.

[74] Ayuntamiento de Madrid, *Portal de Estadística del Ayuntamiento de Madrid*, [En línea]. Disponible en: <https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica?vgnnextchannel=8156e39873674210VgnVCM1000000b205a0aRCRD>

[75] A. Jain, “Data clustering: 50 years beyond K-means”, *Pattern Recognition Letters*, vol. 31, pp. 651–666, 2010.

[76] Scikit-learn documentation, “K-means Inertia\_ attribute,” [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.K-means.html>

- [77] Ketchen, D. J., & Shook, C. L. (1996). "The application of cluster analysis in strategic management research: An analysis and critique." *Strategic Management Journal*, 17(6), 441-458.
- [78] López-Solís, O., Hará-Sarango, A., Córdova-Pacheco, A., & Pérez-Bnceno, J. (2023). El teorema Modigliani-Miller: un análisis desde la estructura de capital mediante modelos Data Mining en pymes del sector comercio. *Revista Finanzas y Política Económica*, 15(1), 45-66.
- [79] M. Ester, H.-P. Kriegel, J. Sander y X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, en *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, 1996, pp. 226–231.
- [80] J. Han, M. Kamber y J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
- [81] B. Everitt, S. Landau, M. Leese y D. Stahl, *Cluster Analysis*, 5th ed., Wiley, 2011.
- [82] Kaufman, L. & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.
- [83] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.
- [84] Shahapure, K. R., & Nicholas, C. (2020, October). *Cluster quality analysis using silhouette score*. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 747-748). IEEE.
- [85] Wang, X., & Xu, Y. (2019, July). *An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index*. In IOP Conference Series: Materials Science and Engineering (Vol. 569, No. 5, p. 052024). IOP Publishing.

## ANEXO A

---

MAPAS INTERACTIVOS (descargar archivo HTML y abrir en navegador):

Barrios	
Modelo	Mapa
<i>K-means</i> k=4	<a href="https://drive.google.com/file/d/1srAnKTtSfiw8kHOqVydft7Jv39DHCQuM/view?usp=drive_link">https://drive.google.com/file/d/1srAnKTtSfiw8kHOqVydft7Jv39DHCQuM/view?usp=drive_link</a>
<i>K-means</i> k=3	<a href="https://drive.google.com/file/d/1odB9OqlfVgsUZQ-ttC3gqPls5_iIH5SA/view?usp=drive_link">https://drive.google.com/file/d/1odB9OqlfVgsUZQ-ttC3gqPls5_iIH5SA/view?usp=drive_link</a>
DBSCAN	<a href="https://drive.google.com/file/d/1Yruv_YWeLp8phF3r7fVYunMMwHpP2q71/view?usp=drive_link">https://drive.google.com/file/d/1Yruv_YWeLp8phF3r7fVYunMMwHpP2q71/view?usp=drive_link</a>
<i>Agglomerative</i> k=4	<a href="https://drive.google.com/file/d/1m60oRh5puCmV-gEGtegNCegFpl6aB_wG/view?usp=drive_link">https://drive.google.com/file/d/1m60oRh5puCmV-gEGtegNCegFpl6aB_wG/view?usp=drive_link</a>
<i>Agglomerative</i> k=3	<a href="https://drive.google.com/file/d/1oAJ-HRayvuCxMH7mlmarM0GXUwYzRKxm/view?usp=drive_link">https://drive.google.com/file/d/1oAJ-HRayvuCxMH7mlmarM0GXUwYzRKxm/view?usp=drive_link</a>

Tabla 43. Accesos mapas interactivos barrios

Distritos	
Modelo	Mapa
<i>K-means</i> k=4	<a href="https://drive.google.com/file/d/17EH8-D_LWwOLGNoc5G5qGPGyPMXbGsuv/view?usp=drive_link">https://drive.google.com/file/d/17EH8-D_LWwOLGNoc5G5qGPGyPMXbGsuv/view?usp=drive_link</a>
<i>K-means</i> k=3	<a href="https://drive.google.com/file/d/1h5wKmBQ-nTbwA6NHPDyO4ISHMLVTU43/view?usp=drive_link">https://drive.google.com/file/d/1h5wKmBQ-nTbwA6NHPDyO4ISHMLVTU43/view?usp=drive_link</a>
DBSCAN	<a href="https://drive.google.com/file/d/1Vz1UKn6ahbnx_HyZOr3M_k8UsKf5N37K/view?usp=drive_link">https://drive.google.com/file/d/1Vz1UKn6ahbnx_HyZOr3M_k8UsKf5N37K/view?usp=drive_link</a>
<i>Agglomerative</i> k=4	<a href="https://drive.google.com/file/d/1aXUmLLyNz14Kzt3jYlqzPi-8v1Rr3RVf/view?usp=drive_link">https://drive.google.com/file/d/1aXUmLLyNz14Kzt3jYlqzPi-8v1Rr3RVf/view?usp=drive_link</a>
<i>Agglomerative</i> k=3	<a href="https://drive.google.com/file/d/1t_bu0rlAbnn6bh6SxgODUOzzRDhx9BMn/view?usp=drive_link">https://drive.google.com/file/d/1t_bu0rlAbnn6bh6SxgODUOzzRDhx9BMn/view?usp=drive_link</a>

Tabla 44. Accesos mapas interactivos distritos

ENLACE

RANKING

FINAL

BARRIOS:

[https://docs.google.com/spreadsheets/d/1NkaWxUbjBEVurk\\_xyGwjylZLyK15Qwl4CJyEzpuF6Do/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1NkaWxUbjBEVurk_xyGwjylZLyK15Qwl4CJyEzpuF6Do/edit?usp=sharing)

zona	distrito	cluster_K-means_3	ranking_cluster_K-means_3	categoria_cluster_K-means_3
Embajadores	CENTRO	0	1	Muy vulnerable
Vista Alegre	CARABANCHEL	0	1	Muy vulnerable
Buenavista	CARABANCHEL	0	1	Muy vulnerable
Puerta Bonita	CARABANCHEL	0	1	Muy vulnerable
Comillas	CARABANCHEL	0	1	Muy vulnerable
Berruguete	TETUÁN	0	1	Muy vulnerable
Almenara	TETUÁN	0	1	Muy vulnerable
Opañel	CARABANCHEL	0	1	Muy vulnerable
San Isidro	CARABANCHEL	0	1	Muy vulnerable
Valdeacederas	TETUÁN	0	1	Muy vulnerable
Amposta	SAN BLAS-CANILLEJAS	0	1	Muy vulnerable
Aeropuerto	BARAJAS	0	1	Muy vulnerable
Hellín	SAN BLAS-CANILLEJAS	0	1	Muy vulnerable
Pueblo Nuevo	CIUDAD LINEAL	0	1	Muy vulnerable
Fontarrón	MORATALAZ	0	1	Muy vulnerable
Vinateros	MORATALAZ	0	1	Muy vulnerable
Media Legua	MORATALAZ	0	1	Muy vulnerable
Pavones	MORATALAZ	0	1	Muy vulnerable
Numancia	PUENTE DE VALLECAS	0	1	Muy vulnerable
Portazgo	PUENTE DE VALLECAS	0	1	Muy vulnerable
Palomeras Sureste	PUENTE DE VALLECAS	0	1	Muy vulnerable
Palomeras Bajas	PUENTE DE VALLECAS	0	1	Muy vulnerable
San Diego	PUENTE DE VALLECAS	0	1	Muy vulnerable
Entrevías	PUENTE DE VALLECAS	0	1	Muy vulnerable
Pradolongo	USERA	0	1	Muy vulnerable
Zofío	USERA	0	1	Muy vulnerable
Moscardó	USERA	0	1	Muy vulnerable
Almendrales	USERA	0	1	Muy vulnerable
San Fermín	USERA	0	1	Muy vulnerable
Orcasur	USERA	0	1	Muy vulnerable
Orcasitas	USERA	0	1	Muy vulnerable

Abrantes	CARABANCHEL	0	1	Muy vulnerable
Villaverde Alto, C.H. Villaverde	VILLAVERDE	0	1	Muy vulnerable
Ángeles	VILLAVERDE	0	1	Muy vulnerable
Los Rosales	VILLAVERDE	0	1	Muy vulnerable
San Cristóbal	VILLAVERDE	0	1	Muy vulnerable
Valdebernardo	VICÁLVARO	0	1	Muy vulnerable
Simancas	SAN BLAS-CANILLEJAS	0	1	Muy vulnerable
Casco Histórico de Vicálvaro	VICÁLVARO	0	1	Muy vulnerable
Casco Histórico de Vallecas	VILLA DE VALLECAS	0	1	Muy vulnerable
Canillejas	SAN BLAS-CANILLEJAS	0	1	Muy vulnerable
Arcos	SAN BLAS-CANILLEJAS	0	1	Muy vulnerable
Cortes	CENTRO	1	2	Vulnerable
Palacio	CENTRO	1	2	Vulnerable
Universidad	CENTRO	1	2	Vulnerable
Sol	CENTRO	1	2	Vulnerable
Imperial	ARGANZUELA	1	2	Vulnerable
Justicia	CENTRO	1	2	Vulnerable
Castellana	SALAMANCA	1	2	Vulnerable
El Viso	CHAMARTÍN	1	2	Vulnerable
Niño Jesús	RETIRO	1	2	Vulnerable
Recoletos	SALAMANCA	1	2	Vulnerable
Goya	SALAMANCA	1	2	Vulnerable
Fuente del Berro	SALAMANCA	1	2	Vulnerable
Guindalera	SALAMANCA	1	2	Vulnerable
Acacias	ARGANZUELA	1	2	Vulnerable
Chopera	ARGANZUELA	1	2	Vulnerable
Delicias	ARGANZUELA	1	2	Vulnerable
Palos de la Frontera	ARGANZUELA	1	2	Vulnerable
Pacífico	RETIRO	1	2	Vulnerable
Estrella	RETIRO	1	2	Vulnerable
Adelfas	RETIRO	1	2	Vulnerable
Los Jerónimos	RETIRO	1	2	Vulnerable
Ibiza	RETIRO	1	2	Vulnerable

Casco Histórico de Barajas	BARAJAS	1	2	Vulnerable
Alameda de Osuna	BARAJAS	1	2	Vulnerable
El Salvador	SAN BLAS-CANILLEJAS	1	2	Vulnerable
Canillas	HORTALEZA	1	2	Vulnerable
Atalaya	CIUDAD LINEAL	1	2	Vulnerable
Colina	CIUDAD LINEAL	1	2	Vulnerable
San Juan Bautista	CIUDAD LINEAL	1	2	Vulnerable
Costillares	CIUDAD LINEAL	1	2	Vulnerable
Valdezarza	MONCLOA-ARAVACA	1	2	Vulnerable
Casa de Campo	MONCLOA-ARAVACA	1	2	Vulnerable
Vallehermoso	CHAMBERÍ	1	2	Vulnerable
El Pardo	FUENCARRAL-EL PARDO	1	2	Vulnerable
Fuentelarreina	FUENCARRAL-EL PARDO	1	2	Vulnerable
Peñagrande	FUENCARRAL-EL PARDO	1	2	Vulnerable
Ciudad Universitaria	MONCLOA-ARAVACA	1	2	Vulnerable
Argüelles	MONCLOA-ARAVACA	1	2	Vulnerable
Trafalgar	CHAMBERÍ	1	2	Vulnerable
Arapiles	CHAMBERÍ	1	2	Vulnerable
Gaztambide	CHAMBERÍ	1	2	Vulnerable
Cuatro Caminos	TETUÁN	1	2	Vulnerable
Castillejos	TETUÁN	1	2	Vulnerable
Almagro	CHAMBERÍ	1	2	Vulnerable
Ríos Rosas	CHAMBERÍ	1	2	Vulnerable
Lista	SALAMANCA	1	2	Vulnerable
Hispanoamérica	CHAMARTÍN	1	2	Vulnerable
Nueva España	CHAMARTÍN	1	2	Vulnerable
Prosperidad	CHAMARTÍN	1	2	Vulnerable
Ciudad Jardín	CHAMARTÍN	1	2	Vulnerable
Castilla	CHAMARTÍN	1	2	Vulnerable
Bellas Vistas	TETUÁN	1	2	Vulnerable
Pilar	FUENCARRAL-EL PARDO	1	2	Vulnerable

La Paz	FUENCARRAL-EL PARDO	1	2	Vulnerable
La Concepción	CIUDAD LINEAL	1	2	Vulnerable
Quintana	CIUDAD LINEAL	1	2	Vulnerable
Ventas	CIUDAD LINEAL	1	2	Vulnerable
Marroquina	MORATALAZ	1	2	Vulnerable
San Pascual	CIUDAD LINEAL	1	2	Vulnerable
Santa Eugenia	VILLA DE VALLECAS	1	2	Vulnerable
Apóstol Santiago	HORTALEZA	1	2	Vulnerable
Pinar del Rey	HORTALEZA	1	2	Vulnerable
Aravaca	MONCLOA-ARAVACA	2	3	Poco vulnerable
El Plantío	MONCLOA-ARAVACA	2	3	Poco vulnerable
Valverde	FUENCARRAL-EL PARDO	2	3	Poco vulnerable
Valdemarín	MONCLOA-ARAVACA	2	3	Poco vulnerable
El Goloso	FUENCARRAL-EL PARDO	2	3	Poco vulnerable
Mirasierra	FUENCARRAL-EL PARDO	2	3	Poco vulnerable
Atocha	ARGANZUELA	2	3	Poco vulnerable
Legazpi	ARGANZUELA	2	3	Poco vulnerable
Piovera	HORTALEZA	2	3	Poco vulnerable
Palomas	HORTALEZA	2	3	Poco vulnerable
Horcajo	MORATALAZ	2	3	Poco vulnerable
El Cañaveral	VICÁLVARO	2	3	Poco vulnerable
Valdefuentes	HORTALEZA	2	3	Poco vulnerable
Valderrivas	VICÁLVARO	2	3	Poco vulnerable
Ensanche de Vallecas	VILLA DE VALLECAS	2	3	Poco vulnerable
Butarque	VILLAVERDE	2	3	Poco vulnerable
Rejas	SAN BLAS-CANILLEJAS	2	3	Poco vulnerable
Rosas	SAN BLAS-CANILLEJAS	2	3	Poco vulnerable
Timón	BARAJAS	2	3	Poco vulnerable
Corralejos	BARAJAS	2	3	Poco vulnerable

Tabla 45. Ranking final barrios

ENLACE	RANKING	FINAL	DISTRITOS:
<a href="https://docs.google.com/spreadsheets/d/1QEdjXztmoMBY0rQDHfsB1YQFwGUDWJLsOgKtCuOCfIA/edit?usp=sharing">https://docs.google.com/spreadsheets/d/1QEdjXztmoMBY0rQDHfsB1YQFwGUDWJLsOgKtCuOCfIA/edit?usp=sharing</a>			

zona	distrito	cluster_agglo m_4	ranking_cluster_agglo m_4	categoria_cluster_aggl om_4
CARABANC HEL	CARABANC HEL	2	1	Muy vulnerable
PUENTE DE VALLECAS	PUENTE DE VALLECAS	2	1	Muy vulnerable
USERA	USERA	2	1	Muy vulnerable
VILLAVERDE	VILLAVERDE	2	1	Muy vulnerable
LATINA	LATINA	0	2	Vulnerable
TETUÁN	TETUÁN	0	2	Vulnerable
MORATALAZ	MORATALAZ	0	2	Vulnerable
CENTRO	CENTRO	0	2	Vulnerable
SAN BLAS-CANILLEJAS	SAN BLAS-CANILLEJAS	0	2	Vulnerable
CIUDAD LINEAL	CIUDAD LINEAL	0	2	Vulnerable
HORTALEZA	HORTALEZA	1	3	Poco vulnerable
VILLA DE VALLECAS	VILLA DE VALLECAS	1	3	Poco vulnerable
VICÁLVARO	VICÁLVARO	1	3	Poco vulnerable
BARAJAS	BARAJAS	1	3	Poco vulnerable
FUENCARRA L-EL PARDO	FUENCARRA L-EL PARDO	3	4	Muy poco vulnerable
CHAMBERÍ	CHAMBERÍ	3	4	Muy poco vulnerable
MONCLOA-ARAVACA	MONCLOA-ARAVACA	3	4	Muy poco vulnerable
CHAMARTÍN	CHAMARTÍN	3	4	Muy poco vulnerable
RETIRO	RETIRO	3	4	Muy poco vulnerable
ARGANZUELA	ARGANZUELA	3	4	Muy poco vulnerable
SALAMANCA	SALAMANCA	3	4	Muy poco vulnerable

Tabla 46. Ranking final distritos

## ANEXO B

---

CÓDIGO

<https://colab.research.google.com/drive/1DoqkyAEmDEMAAdHBmerBu2Wc9INb9nkg5?usp=sharing>

BARRIOS:

CÓDIGO

<https://colab.research.google.com/drive/12IzEqNxTxnK9hU6343FSUoTfmky0vnof?usp=sharing>

DISTRITOS:

DOCUMENTOS

NECESARIOS

PARA

EJECUTAR

CÓDIGO:

<https://drive.google.com/drive/folders/1nXXSqvJglqzkwYuMoPyLQqRb74Zm0Qy?usp=sharing>