Aprendizaje No Supervisado Adaptativo mediante Ensemble de Clustering y Modelos Fundacionales Tabulares para el Análisis de Datos Complejos

Ana Rodríguez-López, María Martínez-Ballesteros [1], José M. Luna-Romera [1], and José C. Riquelme [2]

Programa de Doctorado en Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla,
Av. Reina Mercedes S/N, 41012 Sevilla, España
arodriguez24@us.es

[1] Director [2] Tutor

Inicio de la tesis: agosto de 2025

Resumen Esta tesis doctoral propone el desarrollo de un *framework* de aprendizaje no supervisado adaptativo para el análisis de datos complejos, partiendo de la hipótesis de que es posible reducir la dependencia de decisiones manuales en el proceso de *clustering* mediante la combinación de mecanismos adaptativos de validación, técnicas de *ensemble* y modelos fundacionales tabulares como generadores de representaciones latentes.

Como primera contribución de esta tesis, se ha desarrollado un *framework* inicial de *ensemble* de *clustering* jerárquico basado en índices de validación interna, actualmente en revisión en *Information Fusion*, que demuestra mejoras consistentes respecto al *clustering* jerárquico clásico y evidencia limitaciones abiertas en la selección automática de criterios de validación y la estimación del número de clústeres, motivando las líneas de investigación propuestas.

A partir de esta base, la tesis tiene como objetivo general desarrollar un *framework* generalizable y reproducible de aprendizaje no supervisado que integre mecanismos adaptativos de *ensemble* de *clustering*, el análisis de la relación entre las propiedades estructurales de los datos y el comportamiento de los índices de validación interna, y la exploración del potencial de los modelos fundacionales tabulares como generadores de representaciones latentes. La propuesta será validada en dominios científicos con escasez o ausencia de etiquetas, incluyendo aplicaciones relacionadas con datos ambientales, biomédicos y/o energéticos.

Keywords: Aprendizaje no supervisado · Clustering · Ensemble · Modelos fundacionales tabulares

1. Introducción

Esta sección presenta la motivación científica de la tesis, su enfoque diferencial respecto al estado del arte y las contribuciones principales que se esperan obtener.

1.1. Contexto y Motivación

El aprendizaje no supervisado constituye una de las áreas fundamentales de la inteligencia artificial, especialmente en aquellos escenarios donde la disponibilidad de etiquetas es limitada, costosa o inexistente. Entre las principales técnicas de aprendizaje no supervisado destacan los algoritmos de *clustering*, que permiten identificar patrones, estructuras latentes y grupos de comportamiento similares en conjuntos de datos complejos, facilitando el descubrimiento de conocimiento y el análisis exploratorio de grandes volúmenes de información.

Esta capacidad resulta fundamental en numerosos dominios científicos y aplicados, como la salud, la energía o el medioambiente, donde la disponibilidad de grandes cantidades de datos contrasta con la ausencia de etiquetas fiables o de conocimiento previo suficiente para identificar las estructuras subyacentes presentes en los datos. En estos contextos, los métodos de *clustering* constituyen una herramienta esencial para apoyar la generación de hipótesis, la caracterización de fenómenos complejos y la toma de decisiones basada en datos.

A pesar de su importancia, este tipo de algoritmos presenta limitaciones metodológicas que dificultan su aplicación sistemática en conjuntos de datos complejos. Los resultados obtenidos dependen frecuentemente de decisiones tomadas por el investigador, como la elección del algoritmo de *clustering* en cuestión, la configuración de sus parámetros, la selección de criterios de validación o el criterio para determinar el número de clústeres. Estas decisiones suelen apoyarse en heurísticas o conocimiento experto, lo que dificulta la reproducibilidad de los análisis y limita la transferencia de las metodologías entre distintos dominios de aplicación.

A estas limitaciones se suman los desafíos derivados de la creciente complejidad de los datos actuales, caracterizados por alta dimensionalidad, heterogeneidad, ruido y relaciones no lineales entre variables. En este contexto, surge la necesidad de desarrollar metodologías más adaptativas, capaces de aprender representaciones informativas de los datos, descubrir automáticamente su estructura interna y reducir la dependencia de decisiones manuales durante el proceso de análisis.

1.2. Enfoque Diferencial y Contribuciones Esperadas

Esta tesis da continuidad a diversas líneas de investigación desarrolladas por el grupo Minerva, cuya trayectoria en aprendizaje no supervisado y aprendizaje profundo aplicado a dominios científicos complejos ha evidenciado dos necesidades recurrentes: obtener representaciones informativas de los datos e identificar automáticamente patrones complejos en ausencia de etiquetas.

Desde una perspectiva metodológica, gran parte de las limitaciones actuales del *clustering* pueden interpretarse como consecuencia de tres decisiones fundamentales que siguen dependiendo en gran medida del criterio del investigador: cómo representar los datos, cómo evaluar la calidad de las particiones obtenidas y cómo determinar la estructura de agrupamiento más adecuada. Aunque estos problemas han sido ampliamente estudiados, la literatura los aborda habitualmente de forma independiente, dando lugar a soluciones parciales que resultan difíciles de combinar y transferir entre dominios.

La originalidad de esta tesis reside en plantear que estas tres decisiones fundamentales no deben abordarse como problemas independientes, sino como componentes interrelacionados de un mismo proceso de descubrimiento de estructura. La hipótesis subyacente es que las propiedades intrínsecas de los datos, junto con representaciones latentes capaces de capturar información estructural relevante, contienen información suficiente para guiar de forma adaptativa la evaluación, combinación e identificación de estructuras de agrupamiento. Bajo esta premisa, la tesis propone un marco de aprendizaje no supervisado adaptativo que combina técnicas de *ensemble* de *clustering*, mecanismos de validación interna y modelos fundacionales tabulares dentro de una arquitectura unificada. Este enfoque persigue reducir la dependencia de decisiones manuales, mejorar la reproducibilidad de los análisis y facilitar la transferencia de metodologías entre dominios.

Para ello, las contribuciones previstas abordan de forma secuencial los distintos componentes necesarios para reducir la dependencia de decisiones manuales en el proceso de clustering, desde la generación y evaluación de particiones hasta el aprendizaje de representaciones y su integración en un *framework* unificado:

- Un *framework* de *ensemble* de *clustering* jerárquico basado en índices de validación interna de *clustering* (CVIs), que establece el punto de partida para el desarrollo posterior de mecanismos adaptativos de evaluación, consenso y descubrimiento de estructura, con implementación disponible en repositorio público.
- Mecanismos adaptativos capaces de relacionar las propiedades estructurales de los datos con el comportamiento de los índices de validación interna y las estrategias de consenso, permitiendo automatizar la selección de criterios de evaluación y la estimación de la estructura de agrupamiento más adecuada.
- Un estudio sistemático del potencial de los modelos fundacionales tabulares como generadores de representaciones latentes para el descubrimiento de estructura en datos complejos, analizando su capacidad para capturar información relevante para algoritmos de *clustering*.
- La integración de los avances metodológicos desarrollados en un *framework* unificado de aprendizaje no supervisado adaptativo, capaz de combinar mecanismos de representación, validación y consenso dentro de un mismo proceso de descubrimiento de estructura, validado en dominios científicos reales con escasez o ausencia de etiquetas. También disponible en un repositorio público.

2. Estado del Arte

Esta sección revisa los principales enfoques relacionados con la investigación propuesta, identificando las limitaciones que motivan las contribuciones de esta tesis.

2.1. Algoritmos de Clustering, Métodos Ensemble y Evaluación de la Calidad de las Particiones

Los algoritmos de *clustering* constituyen una herramienta fundamental para el análisis exploratorio de datos y el descubrimiento de estructura en escenarios sin etiquetas. A pesar de su amplia utilización, la selección del algoritmo, la configuración de sus parámetros y la determinación del número de clústeres continúan dependiendo en gran medida de heurísticas o conocimiento experto, comprometiendo la reproducibilidad y la transferencia de metodologías entre dominios [1][2][3].

Los métodos de *ensemble* de *clustering* combinan múltiples particiones generadas bajo distintas configuraciones con el objetivo de obtener soluciones más estables y representativas que las proporcionadas por un único algoritmo [4][5].

La selección de las particiones más representativas dentro de un *ensemble* requiere criterios de calidad que no dependan de etiquetas externas. En este contexto, los CVIs internos desempeñan un papel fundamental, ya que permiten evaluar características de las particiones a partir de los propios datos [6][7]. Sin embargo, el comportamiento de los CVIs varía significativamente en función de las propiedades estructurales del conjunto de datos, como la dimensionalidad, el ruido o la forma de los clústeres, lo que dificulta la selección automática del criterio más adecuado [8]. Esta dependencia contextual de los CVIs constituye una de las limitaciones abiertas más relevantes del campo y uno de los problemas a abordar en esta tesis.

2.2. Aprendizaje de Representaciones y Modelos Fundacionales Tabulares

El aprendizaje de representaciones ha transformado el *clustering* en dominios no estructurados como imagen o texto, donde técnicas como los *autoencoders* profundos permiten proyectar los datos en espacios latentes más apropiados para el agrupamiento [9][10]. Sin embargo, su aplicación a datos tabulares presenta desafíos específicos derivados de la heterogeneidad de las variables y la ausencia de invarianzas estructurales, que limitan la efectividad de los enfoques desarrollados en otros escenarios.

En este contexto, los modelos fundacionales tabulares representan una línea emergente de investigación con un potencial todavía escasamente explorado en escenarios no supervisados. TabPFN [11] introdujo un enfoque basado en *transformers* preentrenados sobre millones de tareas sintéticas capaz de resolver tareas de clasificación tabular en una única pasada hacia delante, sin necesidad de *fine-tuning*. Su sucesor TabPFNv2 amplía considerablemente el rango operativo,

mientras que TabICL extiende estas capacidades a conjuntos de datos de mayor tamaño mediante aprendizaje en contexto [12]. Más recientemente, TabPFNv3 ha mejorado la escalabilidad y el rendimiento de esta familia de modelos, extendiendo su aplicabilidad a conjuntos de datos de mayor tamaño y modalidades tabulares más complejas [13]. Estos modelos generan representaciones latentes durante la inferencia que, en principio, reflejan información estructural aprendida durante el preentrenamiento.

Sin embargo, los *embeddings* producidos por arquitecturas *Transformer* presentan un problema conocido de anisotropía: los vectores tienden a concentrarse en un cono estrecho del espacio latente, lo que distorsiona las medidas de similitud y perjudica el rendimiento de los algoritmos de *clustering* clásicos aplicados directamente sobre ellos [14]. Trabajos recientes como ZEUS han demostrado que es posible preentrenar un modelo fundacional para inferir asignaciones de *cluster* en datos sintéticos y generalizar a datos reales, superando a algoritmos clásicos y enfoques de *deep clustering* [15], pero la aplicación sistemática de modelos como TabPFN o TabICL como generadores de representaciones para *clustering* adaptativo permanece prácticamente inexplorada.

2.3. Limitaciones Identificadas y Oportunidad de Investigación

La revisión del estado del arte identifica tres limitaciones principales: (i) la ausencia de mecanismos adaptativos para seleccionar índices de validación interna en función de las propiedades de los datos, (ii) la dificultad de inferir automáticamente la estructura de agrupamiento más adecuada, y (iii) la falta de estudios sistemáticos sobre el uso de modelos fundacionales tabulares como generadores de representaciones para clustering.

Aunque estas líneas han sido estudiadas de forma independiente, no existen marcos que las integren dentro de un mismo proceso de aprendizaje no supervisado. Esta tesis aborda conjuntamente estos tres problemas mediante la combinación de *ensemble* clustering, validación adaptativa y modelos fundacionales tabulares, con el objetivo de desarrollar metodologías más reproducibles y transferibles entre dominios.

3. Hipótesis y Objetivos

Esta sección presenta la hipótesis central de la tesis y las preguntas de investigación derivadas, así como el objetivo general y los subobjetivos específicos que estructuran su desarrollo.

3.1. Hipótesis de Partida y Preguntas de Investigación

La hipótesis central de esta tesis es que las propiedades estructurales de los datos contienen información suficiente para guiar de forma adaptativa decisiones fundamentales del proceso de *clustering*, incluyendo la evaluación de particiones, la identificación de la estructura de agrupamiento y el uso de representaciones

más informativas. Como consecuencia, es posible desarrollar metodologías de aprendizaje no supervisado más eficaces, reproducibles y transferibles que reduzcan la dependencia de decisiones manuales asumidas por el investigador.

A partir de esta hipótesis, se plantean tres preguntas de investigación que estructuran el desarrollo de la tesis:

1. ¿Es posible modelar la relación entre las propiedades estructurales de un conjunto de datos y el comportamiento de distintos CVIs internos, de forma que dicha información permita automatizar decisiones tradicionalmente dependientes del experto durante el proceso de *clustering*?
2. ¿Puede la información contenida en múltiples particiones de un mismo conjunto de datos utilizarse para inferir de forma más fiable su estructura de agrupamiento, incluyendo la identificación automática del número de clústeres y la selección de soluciones más estables y representativas?
3. ¿Pueden los modelos fundacionales tabulares generar representaciones latentes que capturen información estructural relevante para el agrupamiento, facilitando la identificación de patrones complejos que no resultan evidentes en el espacio original de los datos?

La hipótesis se considerará respaldada si los mecanismos propuestos permiten seleccionar de forma adaptativa estrategias de representación, validación y consenso que conduzcan a resultados comparables o superiores a los obtenidos mediante enfoques de referencia, manteniendo su capacidad de transferencia entre distintos dominios sin requerir una configuración específica para cada uno de ellos.

3.2. Objetivo General

El objetivo principal de esta tesis es desarrollar un *framework* generalizable y reproducible de aprendizaje no supervisado adaptativo, capaz de adaptarse a diferentes tipos de datos y apoyar procesos de descubrimiento de conocimiento en escenarios con ausencia de etiquetas, mediante la combinación de técnicas de *ensemble* de *clustering* y modelos fundacionales tabulares.

3.3. Subobjetivos

- **SO1.** Desarrollar mecanismos adaptativos de *clustering* mediante técnicas *ensemble* como primera aproximación para el descubrimiento automático de estructuras.
- **SO2.** Analizar la relación entre las propiedades estructurales de los datos, los CVIs y los mecanismos de consenso para apoyar estrategias adaptativas de *clustering*.
- **SO3.** Evaluar el potencial de los modelos fundacionales tabulares como generadores de representaciones para clustering.
- **SO4.** Integrar los avances desarrollados en un *framework* unificado de aprendizaje no supervisado adaptativo.
- **SO5.** Validar las metodologías propuestas en dominios científicos caracterizados por datos complejos y escasez de etiquetas.

4. Metodología y Plan de Trabajo

Esta sección describe la estrategia metodológica y los paquetes de trabajo que estructuran el desarrollo de la tesis.

4.1. Metodología General

La investigación seguirá una metodología experimental iterativa basada en la revisión continua del estado del arte, el desarrollo y la evaluación de nuevas metodologías de *clustering*. Cada paquete de trabajo comenzará con una revisión específica de la literatura relacionada con el problema a abordar.

4.2. Visión General del Framework Unificado Propuesto

El *framework* propuesto integra tres componentes principales: (i) un módulo de *ensemble* de clustering jerárquico basado en una medida de distancia adaptativa respaldada por índices de validación interna, (ii) mecanismos adaptativos de validación y consenso guiados por las propiedades estructurales de los datos, y (iii) aprendizaje de representaciones mediante modelos fundacionales tabulares.

4.3. PT1: Framework de Ensemble de Clustering Jerárquico

Objetivo. Desarrollar un *framework* inicial de *ensemble* de *clustering* jerárquico basado en CVIs internos que mejore la calidad y estabilidad del *clustering* jerárquico clásico mediante mecanismos de selección de particiones y construcción de una medida de distancia adaptativa.

Tareas principales:

- T1.1. Diseño e implementación del *framework*: generación de particiones jerárquicas bajo distintas configuraciones, selección mediante CVIs internos y construcción de la medida de distancia adaptativa.
- T1.2. Experimentación sobre conjuntos de datos de referencia, análisis de la relación entre CVIs internos y medidas de evaluación externas.

Hito: Artículo científico que presenta el *framework*, su experimentación y validación asociadas, y las conclusiones derivadas de la investigación [16]. *(Contribución actualmente en revisión.)*

4.4. PT2: Análisis Adaptativo de CVIs y Mecanismos de Consenso

Objetivo. Analizar la relación entre las propiedades estructurales de los datos, el comportamiento de los CVIs internos y los mecanismos de consenso, con el fin de mejorar estrategias adaptativas en *clustering*.

Tareas principales:

- T2.1. Análisis sistemático de la relación entre las características estructurales de los datos y el comportamiento de distintos CVIs internos.
- T2.2. Estudio del impacto de distintas estrategias de consenso sobre la calidad y estabilidad de las particiones obtenidas.
- T2.3. Desarrollo de mecanismos adaptativos para la selección y combinación de CVIs y estrategias de consenso en función de las propiedades del conjunto de datos.

Hito: Artículo científico sobre las relaciones entre características estructurales de los datos, CVIs internos y mecanismos de consenso para *clustering* adaptativo.

4.5. PT3: Modelos Fundacionales Tabulares para Representación en Clustering

Objetivo. Investigar el uso de modelos fundacionales tabulares como generadores de representaciones latentes para mejorar el rendimiento, estabilidad e interpretabilidad del *clustering* en datos complejos.

Tareas principales:

- T3.1. Diseño de flujos de procesamiento para extracción de representaciones latentes en escenarios no supervisados.
- T3.2. Comparación entre *clustering* sobre características originales, sobre representaciones clásicas y sobre *embeddings* de modelos fundacionales.

Hito: Artículo científico sobre modelos fundacionales tabulares como mecanismo de representación para *clustering* adaptativo.

4.6. PT4: Framework Unificado de Aprendizaje No Supervisado Adaptativo

Objetivo. Integrar los componentes desarrollados en un *framework* adaptativo para el análisis no supervisado de datos complejos.

Tareas principales:

- T4.1. Diseño de una arquitectura modular que integre *ensemble* de *clustering*, aprendizaje de representaciones, validación adaptativa y análisis de resultados.
- T4.2. Evaluación del *framework* frente a métodos tradicionales y enfoques recientes.

Hito: Artículo científico sobre el *framework* unificado de aprendizaje no supervisado adaptativo.

Cuadro 1. Dominios de validación previstos y posibles casos de estudio.

Dominio	Posibles casos de estudio
Ambiental	Calidad del aire y distribuciones de tamaño de partículas atmosféricas, en línea con trabajos previos del grupo Minerva sobre contaminación atmosférica y series temporales multivariantes [17].
Biomédico	Identificación de perfiles clínicos y subgrupos de pacientes a partir de datos heterogéneos con disponibilidad limitada de etiquetas, incluyendo potenciales aplicaciones en reproducción asistida.
Energético	Análisis de patrones de consumo, precios de mercado y datos procedentes de sistemas SCADA para la identificación de comportamientos característicos y situaciones anómalas [18].

4.7. PT5: Transferencia y Validación Multidominio

Objetivo. Validar el *framework* en dominios científicos reales con datos complejos y escasez o ausencia de etiquetas, evaluando su capacidad de generalización y transferencia.

Tareas principales:

- T5.1. Recopilación, limpieza y análisis exploratorio de datos reales en los dominios ambiental, biomédico y/o energético, en posible colaboración con expertos de cada área.
- T5.2. Evaluación técnica de los agrupamientos obtenidos.
- T5.3. Interpretación de los clústeres y análisis de su utilidad científica y/o aplicada.

Hito: Artículo científico de aplicación en al menos uno de los dominios considerados.

4.8. Estrategia Experimental y Dominios de Aplicación

La validación experimental se realizará en dos fases complementarias. En primer lugar, se utilizarán conjuntos de datos *benchmark* procedentes de repositorios como UCI, OpenML y colecciones sintéticas controladas para analizar el comportamiento de las metodologías bajo diferentes condiciones de dimensionalidad, ruido y complejidad estructural. En segundo lugar, las propuestas desarrolladas se evaluarán en dominios científicos reales alineados con la trayectoria investigadora del grupo Minerva, seleccionados por la complejidad de sus datos y la escasez de etiquetas disponibles (Tabla 1).

4.9. Cronograma

La tesis está diseñada para desarrollarse en un periodo de cuatro años. La Tabla 2 resume la distribución prevista de actividades por semestres.

Cuadro 2. Cronograma previsto de la tesis por semestres. Las columnas sombreadas corresponden a semestres completados a junio de 2026.

Actividad	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
Revisión del estado del arte	•	•	•	•	•	•	•	•
PT1: *Ensemble* de *clustering* jerárquico	•	•						
PT2: CVIs adaptativos		•	•	•				
PT3: Modelos fundacionales			•	•	•			
PT4: *Framework* unificado				•	•	•		
PT5: Validación multidominio					•	•	•	•
Redacción de tesis						•	•	•
Publicaciones		P1	P2		P3	P4		P5
Defensa								•

5. Estado Actual de la Investigación

Como primera contribución de la tesis, se ha completado el PT1 (Sección 4.3): un *framework* de *ensemble* de *clustering* jerárquico basado en una medida de distancia adaptativa, con implementación disponible en un repositorio público. Los resultados muestran mejoras consistentes respecto al *clustering* jerárquico tradicional y evidencian relaciones relevantes entre CVIs internos y métricas externas de evaluación. El trabajo correspondiente se encuentra actualmente en revisión en *Information Fusion* [16].

Las cuestiones abiertas identificadas en este trabajo, entre ellas la selección automática de criterios de validación y el papel del aprendizaje de representaciones, constituyen la base sobre la que se desarrollarán PT2, PT3 y PT4 conforme al cronograma establecido (Sección 4.9).

6. Relevancia e Impacto Esperado

Esta sección discute la importancia de la investigación propuesta tanto desde el punto de vista científico y metodológico como desde su potencial de transferencia y aplicación en problemas reales de relevancia social.

6.1. Impacto Científico y Metodológico

El aprendizaje no supervisado es, paradójicamente, una de las áreas más utilizadas y menos consolidadas de la inteligencia artificial. Su uso está ampliamente extendido en la práctica científica, pero la ausencia de etiquetas impide aplicar los mecanismos habituales de validación y comparación, lo que dificulta el progreso acumulativo del campo y genera una dependencia persistente del criterio experto. La investigación propuesta contribuye a reducir esta subjetividad: si es posible anticipar qué criterio de validación o qué representación funciona mejor en función de las propiedades de los datos, el análisis no supervisado deja de ser un arte para acercarse a una disciplina reproducible y transferible.

El interés científico de conectar modelos fundacionales tabulares con *clustering* no reside únicamente en la mejora de rendimiento que puedan ofrecer, sino en la pregunta de fondo que plantea: ¿qué aprenden realmente estos modelos sobre la estructura de los datos cuando se preentrenan para clasificación? Responder a esta pregunta en el dominio tabular, donde no existen las invarianzas naturales del texto o la imagen, tiene implicaciones que van más allá del *clustering* y pueden informar el diseño de futuros modelos fundacionales para datos estructurados.

6.2. Transferencia a la Sociedad y Aplicabilidad en Dominios Reales

Los dominios de aplicación previstos comparten una característica común: son áreas donde los datos existen en abundancia pero el conocimiento etiquetado es costoso, escaso o inviable de obtener a escala. Esto convierte al aprendizaje no supervisado adaptativo en una herramienta estratégica, no meramente técnica, para el avance científico en salud, medioambiente y energía. La disponibilidad de un *framework* abierto, documentado y validado en problemas reales puede además democratizar el acceso a estas metodologías para grupos de investigación sin grandes recursos computacionales, amplificando el impacto de las contribuciones más allá del ámbito académico inmediato.

Agradecimientos. La investigación relacionada con esta tesis doctoral ha sido financiada por el proyecto PID2023-146037OB-C21, financiado por MICIU/AEI/10.13039/501100011033. Los implicados en la misma agradecen el contrato predoctoral PREP2023-001780, financiado por MICIU/AEI/10.13039/501100011033 y por el FSE+.

Referencias

1. Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I., Akinyelu, A.A.: A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Eng. Appl. Artif. Intell.* **110**, 104743 (2022). https://doi.org/10.1016/j.engappai.2022.104743
2. Ran, X., Xi, Y., Lu, Y., Wang, X., Lu, Z.: Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artif. Intell. Rev.* **56**, 8219–8264 (2023). https://doi.org/10.1007/s10462-022-10366-3
3. Gharehchopogh, F.S., Ibrikci, T.: A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery. *Adv. Eng. Inform.* **62**, 102647 (2024). https://doi.org/10.1016/j.aei.2024.102647
4. Boongoen, T., Iam-On, N.: Cluster ensembles: A survey of approaches with recent extensions and applications. *Comput. Sci. Rev.* **28**, 1–25 (2018). https://doi.org/10.1016/j.cosrev.2018.01.003
5. Golalipour, K., Akbari, E., Hamidi, S.S., Lee, M., Enayatifar, R.: From clustering to clustering ensemble selection: A review. *Eng. Appl. Artif. Intell.* **104**, 104388 (2021). https://doi.org/10.1016/j.engappai.2021.104388

6. Rodríguez-López, A., Riquelme-Santos, J.C., Galán-Sales, F.J., Luna-Romera, J.M.: Diseño y evaluación de una nueva medida de distancia para clustering jerárquico. In: Advances in Artificial Intelligence, CAEPIA 2024, *Lect. Notes Comput. Sci.* Springer, Cham (2024). https://doi.org/10.1007/978-3-031-62799-6
7. Hassan, A.A., Ahmed, A.M.: From A-to-Z review of clustering validation indices. *Neurocomputing* **597**, 127985 (2024). https://doi.org/10.1016/j.neucom.2024.127985
8. Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J.: K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **622**, 178–210 (2023). https://doi.org/10.1016/j.ins.2022.11.139
9. Guo, X., Gao, L., Liu, X., Yin, J.: Improved deep embedded clustering with local structure preservation. *Expert Syst. Appl.* **168**, 114380 (2021). https://doi.org/10.1016/j.eswa.2020.114380
10. Chen, Y., Zhang, D., Liu, X.: Exploring structural components in autoencoder-based data clustering. *Eng. Appl. Artif. Intell.* **133**, 108170 (2024). https://doi.org/10.1016/j.engappai.2024.108170
11. Hollmann, N., Müller, S., Eggensperger, K., Hutter, F.: TabPFN: a transformer that solves small tabular classification problems in a second. In: 11th International Conference on Learning Representations (ICLR 2023). OpenReview.net (2023). https://doi.org/10.48550/arXiv.2207.01848
12. van Breugel, B., van der Schaar, M.: Why tabular foundation models should be a research priority. *arXiv* 2405.01147 (2024). https://doi.org/10.48550/arXiv.2405.01147
13. Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Manium, M., Hoo, S.B., Bühler, M., Garg, A., et al.: TabPFN-3: Technical Report. arXiv preprint arXiv:2605.13986 (2026). https://doi.org/10.48550/arXiv.2605.13986
14. Ethayarajh, K.: How contextual are contextualized word representations? Revisiting anisotropy in transformers. In: Proceedings of EMNLP-IJCNLP 2019, pp. 55–65. ACL (2019). https://doi.org/10.18653/v1/D19-1006
15. Marszałek, P., Kuśmierczyk, T., Wydmański, W., Tabor, J., Śmieja, M.: ZEUS: Zero-shot embeddings for unsupervised separation of tabular data. In: Advances in Neural Information Processing Systems (NeurIPS 2025). *arXiv* 2505.10704 (2025). https://doi.org/10.48550/arXiv.2505.10704
16. Rodríguez-López, A., Luna-Romera, J.M., Martínez-Ballesteros, M., Riquelme-Santos, J.C.: Ensemble-Based Distance Framework for Hierarchical Clustering. *Information Fusion* (under review)
17. Jiménez-Navarro, M.J., Lovrić, M., Kecorius, S., Nyarko, E.K., Martínez-Ballesteros, M.: Explainable deep learning on multi-target time series forecasting: an air pollution use case. *Results Eng.* **24**, 103290 (2024). https://doi.org/10.1016/j.rineng.2024.103290
18. Jiménez-Navarro, M.J., Troncoso-García, A., Troncoso, A., Martínez-Álvarez, F., Martínez-Ballesteros, M.: Explainable deep learning with embedded feature selection for electricity demand forecasting. In: 2024 International Conference on Smart Systems and Technologies (SST), pp. 152–159. IEEE (2024). https://doi.org/10.1109/sst61991.2024.10755283