

Métodos basados en *clustering* espacial para la detección de Cúmulos Estelares Abiertos

PLAN DE TRABAJO

Índice

1. Marco Teórico	1
1.1. Ciencia de Datos y Minería de Datos	2
1.2. Agrupamiento de Datos	2
1.3. Agrupamiento de Datos Espaciales	3
1.4. Cúmulos Estelares Abiertos	4
1.5. Detección de Cúmulos Estelares Abiertos mediante <i>Clustering</i> . .	4
2. Problema de Interés	5
3. Objetivo	6
3.1. Objetivos Específicos	6
4. Resultados Esperados	7
5. Metodología de trabajo	8
5.1. Metodología de Ciencia de Diseño	8
5.2. Revisiones Sistemáticas de la Literatura	8
6. Actividades	8
7. Calendarización de las Actividades	9
8. Difusión de Resultados	10
9. Materiales	10

1. Marco Teórico

La presente propuesta de plan de doctorado incluye diferentes conceptos, los cuales son introducidos en las siguientes secciones. En concreto, en la sección 1.1 se definen ciencia de datos y de minería de datos, para luego, en las secciones 1.2 y 1.3, profundizar en la actividad de agrupamiento de datos y de datos espaciales, respectivamente. Posteriormente, en la sección 1.4 se define el contexto particular de aplicación de las técnicas desarrolladas sobre el que se aplicará la tesis doctoral: la detección de cúmulos estelares abiertos. Finalmente, en la sección 1.5 se explicita el método utilizado actualmente en la literatura para el problema mencionado, donde se hace uso de técnicas de agrupamiento espacial. También se detallan las problemáticas encontradas.

1.1. Ciencia de Datos y Minería de Datos

La ciencia de datos es un campo de estudio multidisciplinario que utiliza técnicas, herramientas, algoritmos y metodologías provenientes de campos como la estadística, la inteligencia artificial y las ciencias computacionales, para la extracción de conocimiento útil y novedoso a partir de datos (Donoho 2017; Cao 2017; Sharda, Delen y Turban 2018). En este campo de estudio, la minería de datos se define como la obtención de patrones y regularidades a partir de datos que permiten generalizar tendencias que se hallan implícitas en los mismos, haciendo uso de métodos estadísticos como así también de algoritmos de aprendizaje automático (machine learning) (Aggarwal et al. 2015; Fayyad, Piatetsky-Shapiro y Smyth 1996).

Entre las actividades de minería de datos que destacan tradicionalmente por su aplicabilidad se pueden enumerar la generación de modelos de regresión, de modelos de clasificación, de modelos de asociaciones y modelos de agrupamiento (Han, Pei y Tong 2012).

1.2. Agrupamiento de Datos

El agrupamiento de datos es una de las actividades de la minería de datos cuyo objetivo es la generación de una partición del conjunto de datos, de forma tal que cada celda de la partición, denominada grupo o *cluster*, posea coherencia interna, esto es, que las entidades que pertenecen a cada *cluster* sean similares entre sí, dada una medida de similitud, al mismo tiempo que se diferencien de las entidades que pertenecen a otros grupos.

El agrupamiento de datos se ha utilizado, por ejemplo, para la segmentación de clientes para sistemas recomendadores, generando *clusters* de personas que tienen preferencias similares o comportamientos de compra similares, para luego ejecutar acciones de negocio orientados a cada grupo en particular.

La definición del agrupamiento de datos es en cierta medida ambigua, ya que permite la implementación de estrategias varias para la construcción de estas particiones utilizando además variadas medidas de similitud, en función de las características de los *clusters* a ser hallados y las características de los datos. Esto da lugar a tipos de algoritmos de agrupamiento, entre los que se destacan los algoritmos basados en distancias, los basados en densidades, los basados en grafos, entre muchos otros (Estivill-Castro 2002).

Los algoritmos de agrupamiento basados en distancias, particionan el conjunto de observaciones de la base de datos usualmente en k grupos, cada uno con un representante (usualmente el centroide o punto promedio del *cluster*). Cada elemento de la base de datos se asigna entonces al centroide más cercano, según una función de distancia dada, siendo utilizada por lo general la distancia Euclídea (Bramer 2020). Entre los algoritmos de esta familia, destacan K-means (Bock 2007; Ikotun et al. 2023) y K-Medoids (Park y Jun 2009).

Los algoritmos de agrupamiento basados en distancias suelen generar *clusters* de forma tal que dividen el espacio en celdas de Voronoi, y asignan a cada grupo los elementos de cada celda, por lo que no es factible detectar grupos de

elementos en el espacio de los datos con formas arbitrarias.

En contraste, los algoritmos de agrupamiento basados en densidades, permiten definir grupos de datos según un criterio de densidad. Un grupo, entonces, está conformado por conjuntos de datos ubicados en regiones con alta densidad de observaciones, en el espacio de los datos, separados por regiones con baja densidad. Son algoritmos representantes de esta clase de algoritmos, DBSCAN (Ester et al. 1996) y OPTICS (Kriegel et al. 2011).

1.3. Agrupamiento de Datos Espaciales

Los datos espaciales son una clase particular de datos estructurados en los que se pueden diferenciar dos tipos de atributos: los atributos espaciales o contextuales, y los atributos no espaciales o de comportamiento. Los primeros permiten referenciar cada entidad con un lugar o extensión del espacio, por ejemplo, haciendo uso de coordenadas o conjuntos de coordenadas, mientras que los no espaciales aportan información sobre la entidad en cuestión en el dominio del problema (Aggarwal et al. 2015; Manjula y Narsimha 2014; Rottoli 2022).

Los datos espaciales pueden clasificarse a su vez según cómo se relacionan con el espacio en el que se encuentran. En primer lugar, los puntos representan ubicaciones en el espacio, generalmente mediante un par de coordenadas. Luego, las líneas expresan extensiones unidimensionales, tales como carreteras o ríos, generalmente mediante una sucesión de puntos o coordenadas. Por último, los polígonos representan extensiones bidimensionales, como parcelas de terrenos, regiones, provincias, entre otros (Chang 2008; Malerba 2008; Manjula y Narsimha 2014).

En la minería de datos espaciales, se busca encontrar patrones no solo entre los atributos no espaciales, como en la minería de datos tradicional, sino que también se tienen en consideración las relaciones espaciales entre las entidades en cuestión, ya sea como restricciones para la formación de patrones (por ejemplo, los *clusters* formados con agrupamiento espacial deben contener entidades espacialmente contiguas), o como parte de los patrones mismos (por ejemplo, una regla de comportamiento como “las estaciones de servicio (dato puntual) se encuentran en la intersección (relación espacial) de dos calles (datos lineales)”) (Mennis y Guo 2009; Deren Li, Wang, Yuan et al. 2016).

En cuanto al agrupamiento espacial, se distinguen tres tipos de métodos de agrupamiento: El agrupamiento clásico de datos espaciales, donde hace uso de los atributos espaciales y no espaciales de forma indistinta. La regionalización, en cambio, distingue en la conformación de grupos los espaciales de los no espaciales, utilizando los primeros para restringir la ubicación de los elementos de un *cluster*, y los segundos para asegurar su grado de similitud. Por último el descubrimiento de zonas calientes, solo hace uso de los atributos espaciales para distinguir zonas del espacio con alta concentración de datos (Brimicombe 2007; Mennis y Guo 2009; Deren Li, Wang, Yuan et al. 2016; Deren Li, Wang y Deyi Li 2015; Rottoli 2022).

1.4. Cúmulos Estelares Abiertos

Los cúmulos estelares abiertos son conjuntos de estrellas que pueden encontrarse en galaxias espirales o irregulares, las estrellas que los componen se han originado en la misma nube molecular. Otra característica de los miembros de los cúmulos estelares abiertos es que siguen una misma curva evolutiva conocida como *isocrona*, teniendo todos ellos la misma edad aunque diferenciándose en su masa, temperatura y por lo tanto en su clasificación espectral.

El conocimiento y estudio de los cúmulos abiertos permite comprender la formación estelar y la evolución de nuestra galaxia, hoy en día gracias a las distintas misiones, como lo es Gaia (**GaiaDR3**), que permite el estudio de estos cúmulos a partir de los análisis fotométricos y astrométricos. Hacia principios de la década de 2000, el inventario de cúmulos abiertos en la vecindad solar —basado en catálogos como el de (Dias, W. S. et al. 2002) y observaciones de misiones espaciales como Hipparcos (ESA) y el Telescopio Espacial Hubble— alcanzaba aproximadamente los 1100 cúmulos abiertos. Emily L. Hunt y Reffert 2023, emplearon el algoritmo HDBSCAN para realizar análisis en el catálogo de Gaia DR 3¹ buscando cúmulos estelares, analizaron 729 millones de fuentes del catálogo Gaia DR 3 con valores de $G \leq 20$ magnitudes, de dichas fuentes se lograron publicar 7167 cúmulos sin hacer distinción entre cúmulos globulares y cúmulos abiertos o incluso grupos en movimiento como son las asociaciones estelares OB.

La importancia en el estudio de los cúmulos abiertos radica en la información evolutiva que aporta acerca de la estructura de los brazos espirales de la Vía Láctea (Dobbs y Pringle 2010), siendo relevante para determinar la evolución de la Galaxia. Cuando el cúmulo abierto es de formación reciente, entre sus miembros pueden encontrarse las estrellas TTauri (Derkink et al. 2024) y/o los YSOs (Young Stellar Objects)(Borissova et al. 2025), los cuales no terminaron de barrer de su entorno el material de la nube molecular donde se han formado.

Para el estudio de los cúmulos se analizan fuentes en el rango fotométrico comprendido entre el azul y el infrarrojo (IR). Catálogos como GAIA y APASS², del visual, para determinar las agrupaciones estelares y 2MASS³, Spitzer⁴, WISE⁵ y MSX⁶ en el IR cercano y mediano, para determinar los YSOs y las TTauri siguiendo las clasificaciones particulares de las bandas del IR para detectarlos.

1.5. Detección de Cúmulos Estelares Abiertos mediante *Clustering*

La detección de cúmulos estelares abiertos utilizando algoritmos de agrupamiento espacial se describe en líneas generales como sigue (Paíz et al. 2025; Emily L. Hunt y Reffert 2023):

¹<https://www.cosmos.esa.int/web/gaia/data-release-3>

²<https://www.aavso.org/apass>

³<https://www.ipac.caltech.edu/2mass/>

⁴<https://www.spitzer.caltech.edu/>

⁵<https://ui.adsabs.harvard.edu/abs/2011ApJ...735..112J/abstract>

⁶<https://www.eas-journal.org/articles/eas/abs/2002/03/eas04017/eas04017.html>

En primer lugar, se parte seleccionando de los catálogos de interés una región definida por un punto central y un radio. Los datos resultantes son prefiltrados según criterios dados por el dominio del problema para reducir el número del conjunto de datos.

Posteriormente, la base de datos construida es utilizada como entrada de algoritmos de agrupamiento basados en densidad tal como DBSCAN (Alfred Castro-Ginard, Jordi, Xavier Luri et al. 2018; Alfred Castro-Ginard, Jordi, Luri et al. 2019) o HDBSCAN (Campello, Moulavi y Sander 2013; Emily L. Hunt y Reffert 2023). Se seleccionan para esto sobre cinco dimensiones o parámetros astrométricos previamente normalizados (2 coordenadas de posición de las estrellas, 2 componentes de movimiento propio en coordenadas ecuatoriales celestes y paralaje). Estos atributos son tratados de manera indistinta al utilizar la distancia euclidiana como medida de similitud para la conformación de los *clusters*. Por otro lado, el ajuste de los hiperparámetros del algoritmo de agrupamiento se realiza de manera iterativa, mediante un enfoque heurístico *greedy*.

Como resultado, se obtienen *clusters* candidatos, que son posteriormente evaluados manualmente utilizando criterios empíricos para descartar falsos positivos (Emily L Hunt y Reffert 2021).

Luego, se realiza un análisis fotométrico para validar la membresía al grupo estelar de cada candidato y finalmente, se conocen propiedades como edad, absorción y estadio evolutivo, entre otros.

2. Problema de Interés

La inteligencia artificial y, en particular, la minería de datos, ha traído grandes beneficios en múltiples áreas de estudio, permitiendo el procesamiento de grandes volúmenes de datos y la obtención de patrones de conocimiento de forma práctica y certera. No obstante, si bien actualmente se han desarrollado algoritmos como Deep-Ensemble (A. Castro-Ginard et al. 2020), para descubrir gran cantidad de cúmulos en el disco galáctico o, GMM (Gaussian Mixture Models), (Sarro et al. 2014), el cual separa la población del cúmulo del ruido del campo, la aplicación de algoritmos de aprendizaje automático para el descubrimiento de cúmulos estelares abiertos aún no ha sido ampliamente explorada (Fluke y Jacobs 2020; Emily L Hunt y Reffert 2021).

Además, el procedimiento utilizado en la literatura para la determinación de cúmulos estelares abiertos utilizando métodos de agrupamiento, como HDBSCAN, posee las siguientes limitaciones:

- El algoritmo no escala con la numerosidad de la base de datos, debiendo filtrar elementos de la misma antes de su ejecución.
- Se deben estimar los hiperparámetros del algoritmo mediante estrategias heurísticas.
- Se utiliza la distancia euclidiana, tratando a los atributos de manera indistinta, incluso a aquellos que se encuentran especificados en el sistema de coordenadas ecuatoriales celestes (ascensión recta y declinación).

- Se detectan falsos positivos que posteriormente deben ser descartados utilizando criterios empíricos y análisis fotométrico adicional, agregando complejidad y trabajo manual al proceso. Esto último genera una demora en la obtención del resultado final que en algunas ocasiones perjudica a la producción científica.
- Se ignora el análisis en el rango espectral IR (de cercano a lejano) que permite revelar cúmulos estelares abiertos, ignorando por completo el óptico.

Por estos motivos, es necesario un estudio en profundidad de la aplicación de algoritmos de agrupamiento espacial para la detección de cúmulos estelares abiertos, considerando las características del dominio, los tipos y cantidad de datos de entrada, las funciones de similitud para evaluar la pertenencia de un miembro a un *cluster*, y el diseño de un método adaptable que considere estas características, superando las limitaciones del estado del arte. Una buena solución, además, implica no sólo aplicar técnicas de IA sino también de aquellas que contribuyen a mejorar la performance del sistema, tal como las relacionadas a computación de alto desempeño (HPC)(Breyer, Daiß y Pflüger 2021; Huerta et al. 2020; Kirk y Hwu 2010; Wu y Simon 2021)

3. Objetivo

Desarrollar un método adaptable, robusto y rápido de agrupamiento espacial basado en densidades para el descubrimiento de cúmulos estelares abiertos y la mejora en el conocimiento de parámetros astrométricos y físicos de cúmulos abiertos conocidos. El punto de partida son las limitaciones de las metodologías existentes aplicadas al problema a considerar.

Respecto a la problemática abordada, el objetivo es especificar un método confiable y rápido, el cual fusione los datos proporcionados por Gaia (movimiento 3D preciso y fotometría visual) con los datos IR que dan la profundidad necesaria para ver a través del polvo del disco galáctico. Se busca alcanzar un análisis completo y detallado a fin de reducir la incerteza en la membresía de algunos cúmulos y descubrir nuevos cúmulos abiertos, mejorando así la información proporcionada por los catálogos de cúmulos abiertos. Además, esta investigación permitirá mejorar el conocimiento actual de los trazadores de la estructura espiral sobre el plano de la Vía Láctea.

3.1. Objetivos Específicos

Del objetivo general se desprenden los siguientes objetivos específicos:

- *Revisar* el estado del arte de los métodos propuestos en el área de astronomía así como los aplicados a la resolución de problemáticas en diferentes áreas.
- *Diseñar* un método basado en agrupamiento espacial para la detección de cúmulos estelares, integrando como parámetros la densidad de una

región y las características astrométricas y fotométricas de sus miembros. Dicha propuesta incluirá técnicas de procesamiento de datos Espaciales Estructurados masivos, Soft Computing (Pratihari 2013; Kanwar, Kumar y Ram 2020) y Computación de Alto Desempeño.

- *Validar* el funcionamiento del método propuesto, analizando para la región considerada, la detección de:
 - Grupos estelares "aparentes" a fin de establecer si constituyen nuevos cúmulos abiertos.
 - Posibles nuevos miembros astro-fotométricos de los cúmulos pertenecientes a la región.
 - Vinculaciones entre los cúmulos abiertos y las asociaciones estelares.
- *Establecer* para los posibles miembros de un cúmulo, los parámetros de enrojecimiento, estadíos evolutivos, edad, entre otros, a través de un exhaustivo análisis de la fotometría visual e IR.
- *Contribuir* con información a los catálogos de cúmulos abiertos, por ejemplo de la constitución de los brazos espirales de la Vía Láctea, particularmente los brazos Orión-Cygnus y Carina-Sagitario.
- *Redactar* las correspondientes memorias de la Tesis planteada.

4. Resultados Esperados

En el desarrollo de la investigación se planea el estudio de métodos basados en clustering espaciales para la detección de cúmulos estelares abiertos, considerando las problemáticas detectadas en la sección 2.

- Estudio sistemático de la literatura para identificar algoritmos de agrupamiento espacial para la detección de cúmulos estelares y las soluciones tentativas que proveen a los problemas planteados. Este trabajo se encuentra actualmente en desarrollo por parte de la postulante.
- Diseño de un método de agrupamiento espacial basado en densidades para la detección de cúmulos estelares considerando como parámetros la densidad de una región y las características astrométricas y fotométricas de sus miembros.
- Especificación de una distancia métrica que se base no solo en astrometría (Gaia) sino que también utilice de entrada el color en el IR (ej. J - Ks) para filtrar estrellas de campo antes de buscar grupos.
- Validación del método mediante la aplicación del mismo a datos de cúmulos estelares abiertos conocidos.
- Validación del algoritmo que trabaje con el error solo bayesiano y considere el valor de la paralaje.

- Ampliación del conocimiento sobre los grupos estelares presentes en el sector estudiado de los brazos espirales de la galaxia mediante el método propuesto.
- Identificación de “corrientes” (Tidal Tails) para conocer las estrellas que se están escapando del núcleo del cúmulo.
- Incremento de la cantidad de cúmulos estelares abiertos que se presentan en los catálogos relacionados con este tema.

Si bien el alcance de la tesis doctoral restringe su aplicación al ámbito de la Astronomía, se espera que el método a ser desarrollado sirva de base para la aplicación de estas técnicas en otros dominios.

5. Metodología de trabajo

En la presente sección se describen brevemente la metodología e instrumentos a utilizar para el desarrollo del trabajo propuesto.

5.1. Metodología de Ciencia de Diseño

La ciencia de diseño (Design Science) es un marco metodológico para la investigación y el diseño de artefactos que interactúen con cierto contexto para la solución de problemas dentro de los límites del mismo, y la generalización de conocimiento obtenido para ser aplicado fuera de dichos límites. La ciencia de diseño se diferencia de las ciencias aplicadas al diseñar nuevos artefactos que pueden ser luego extrapolados a nuevos contextos (Wieringa 2016).

5.2. Revisiones Sistemáticas de la Literatura

Las revisiones sistemáticas (Argimon Pallás y Jiménez Villa 2004; Cartaxo, Pinto y Soares 2020; Kitchenham et al. 2009; Rico et al. 2020) de artículos científicos siguen un método explícito para resumir la información sobre determinado tema o problema. Se diferencia de las revisiones narrativas que provienen de una pregunta estructurada y de un protocolo previamente realizado.

6. Actividades

La Propuesta está dividida en las siguientes etapas, a las cuales se asocia cada actividad y metodología, si procede. Está implícito que, en la medida que avancen las etapas, se realizarán comunicaciones científicas en revistas y congresos especializados tanto del campo de la Informática como del campo de la Astronomía.

Etapas 1

- Realización de cursos básicos de metodología y epistemología de la ciencia, y afines con la temática del plan de tesis.
- Revisión exhaustiva del Estado del Arte y realización de cursos de posgrado.
- Desarrollo de una Pasantía de Investigación en la Facultad de Ciencias Astronómicas y Geofísicas de la UNLP, con la cual, la postulante adquirirá los conocimientos básicos para investigar en el tema propuesto.

Etapas 2

- Definición de función de similitud para la conformación de cúmulos estelares abiertos.
- Implementación de primer prototipo del artefacto para la detección de cúmulos estelares abiertos.

Etapas 3

- Validación del método propuesto mediante aplicación del mismo sobre conjuntos de datos conocidos.
- Iteraciones de ajuste del diseño.

Etapas 4

- Aplicación del método sobre repositorios de datos de interés, para ampliación de base de conocimiento.
- Redacción minuciosa de las memorias de la tesis doctoral.

7. Calendarización de las Actividades

A continuación se presenta una tabla donde se consigna una calendarización aproximada de las etapas de trabajo propuestas en los cinco años de beca.

	Etapas 1	Etapas 2	Etapas 3	Etapas 4
Año 1	x			
Año 2		x		
Año 3		x	x	
Año 4			x	x
Año 5				x

8. Difusión de Resultados

Los resultados logrados a partir del trabajo realizado se difundirá en publicaciones nacionales e internacionales y workshops, tanto en la disciplina de Computación, como de Astronomía.

Actualmente, con el trabajo inicial, se realizaron las siguientes publicaciones y comunicaciones:

- “Cartografiando el Brazo de Carina-Sagitario”, autores: Rizzo, L.; Paíz L.; Elizalde Caviglia, G. & Corti, M. Presentado en la 67a Reunión Anual de la Asociación Argentina de Astronomía (RAAA), Mendoza, los días 15 - 19 de septiembre de 2025.
- “Multiwavelength study of the Carina–Sagittarius Arm: I. Astrometric and photometric search for new open clusters in the $320^\circ \leq l \leq 325^\circ$ region”, autores: Corti, M.A.; Rizzo, L.; Paíz, L.G. & Elizalde Caviglia, G. Enviado a Astronomy & Astrophysics, diciembre 2025. En Evaluación.

En preparación para enviar en marzo 2026:

- “Multiwavelength study of the Carina–Sagittarius Arm: II. Astrometric and photometric search for new open clusters in the $326^\circ \leq l \leq 330^\circ$ region”, autores: Corti, M.A.; Rizzo, L.; Paíz, L.G. & Elizalde Caviglia, G. A enviar a Astronomy & Astrophysics.

9. Materiales

Los materiales a utilizar para el desarrollo de la tesis son:

- Guías para la realización de revisiones sistemáticas de la literatura en el campo de la Ingeniería de Software (Kitchenham et al. 2009; Cartaxo, Pinto y Soares 2020).
- Datos provenientes de:
 - **GAIA Data Release 3**: El catálogo de la misión del satélite GAIA en su versión 3 correspondiente al año 2022, incluye datos astrométricos y fotométricos que resultan fundamentales para el análisis de estrellas, galaxias y cúmulos. Cuenta con datos sobre la posición, el movimiento propio, las magnitudes fotométricas en la banda G, así como velocidad radial entre otros campos de gran relevancia para poder realizar los estudios fotométricos y astrométricos de las fuentes observadas. Los datos obtenidos en el catálogo corresponden al rango visual y parte del rojo La mayoría de estos datos son cuantitativos de tipo decimal. (Gaia Collaboration et al. 2023). En el transcurso de la beca (aproximadamente en el año 2027) Gaia habrá generado posiblemente su último catálogo, el DR 4 y también será consultado en el desarrollo de la Tesis.

- **SPITZER:** El catálogo Spitzer proporcionado por el telescopio espacial del mismo nombre, aporta datos obtenidos en el rango del infrarrojo, debido a la longitud de onda se logra obtener información atravesando el polvo cósmico y permitiendo alcanzar objetos que se encuentran más lejanos u oscurecidos por el material del medio interestelar. El telescopio cuenta con tres instrumentos que se encargan de brindar diferentes tipos de datos, *IRAC Infrared Array Camera* que aporta imágenes en las cuatro longitudes de onda del infrarrojo 3.6, 4.5, 5.8, 8 microns, *IRS Infrared Spectrograph* el cual ofrece datos espectrales en lugar de imágenes y *MIPS Multiband Imaging Photometer for Spitzer* que aporta imágenes en las longitudes de onda en 24, 70 y 160 micrones. El conjunto de valores obtenidos son de gran utilidad para determinar qué tipo de fuentes se están analizando.
- **Wide-field Infrared Survey Explorer:** Este catálogo correspondiente a la misión de WISE, indicado por su nombre también aporta datos de valores en el infrarrojo, en el rango de 3.4, 4.6, 12, y 22 micrones. El conjunto de datos que proporciona para las cuatro bandas tiene una sensibilidad de 0.08 a 6 mJy, yendo de 3.4 a 22 micrones.
- **MSX PSC:** El catálogo obtenido de esta misión buscaba analizar fuentes brillantes en el infrarrojo medio, con un rango que abarca las seis bandas espectrales 4.29, 4.35, 8.28, 12.13, 14.65 y 21.3 micrones de longitud.
- **2MASS:** Catálogo que contiene valores comprendidos en el rango del infrarrojo cercano en las bandas de 1.25, 1.65 y 2.16 micrones de longitud de onda.
- **ASteCA:** *Automated Stellar Cluster Analysis* posibilita realizar un análisis de datos automatizado sin la intervención del usuario. Con su empleo aplicado al grupo de estrellas en estudio pueden obtenerse edad, enrojecimiento, módulo de distancia, metalicidad, binariedad y masa total de cada cúmulo abierto que se investigue.
- **pyUPMASK:** Es un código escrito en lenguaje Python que usa librerías de Astropy⁷ y un método de agrupado que extrae de la librería Scikit-learn⁸. En la selección de miembros estelares realiza la corrección por borde para puntos que están localizados cerca del límite del cúmulo. Luego, modela con una función Gaussiana el centro de cada cúmulo y la probabilidad de pertenencia al mismo, y si para esa estrella no se obtiene cierto valor de percentil, ella será rechazada como miembro del cúmulo.
- Recursos informáticos de las universidades y proyectos de investigación a los que pertenecen tanto la tesista como los directores de tesis y de beca: UTN FRCU, UNSL y UNLP.

⁷<http://www.astropy.org>

⁸<http://scikit-learn.org>

Referencias

- Aggarwal, Charu C et al. (2015). *Data mining: the textbook*. Vol. 1. 3. Springer. DOI: 10.1007/978-3-319-14142-8.
- Argimon Pallás, Josep María y Josep Jiménez Villa (2004). “Métodos de investigación clínica y epidemiológica”. En: *Cap 3*, págs. 16-26.
- Bock, Hans-Hermann (2007). “Clustering methods: a history of k-means algorithms”. En: *Selected contributions in data analysis and classification*, págs. 161-172.
- Borissova, Jura et al. (sep. de 2025). “Young Stellar Objects in the Carina Nebula: Near-infrared Variability and Spectroscopy”. En: 170.3, 135, pág. 135. DOI: 10.3847/1538-3881/adeb7d. arXiv: 2507.20086 [astro-ph.SR].
- Bramer, Max (2020). *Principles of data mining*. Springer. DOI: 10.1007/978-1-4471-7493-6.
- Breyer, Marcel, Gregor Daiß y Dirk Pflüger (2021). “Performance-Portable Distributed k-Nearest Neighbors using Locality-Sensitive Hashing and SYCL”. En: *Proceedings of the 9th International Workshop on OpenCL. IWOCL '21*. Munich, Germany: Association for Computing Machinery. ISBN: 9781450390330. DOI: 10.1145/3456669.3456692. URL: <https://doi.org/10.1145/3456669.3456692>.
- Brimicombe, Allan J (2007). “A dual approach to cluster discovery in point event data sets”. En: *Computers, environment and urban systems* 31.1, págs. 4-18. DOI: 10.1016/j.compenvurbsys.2005.07.004.
- Campello, Ricardo JGB, Davoud Moulavi y Jörg Sander (2013). “Density-based clustering based on hierarchical density estimates”. En: *Pacific-Asia conference on knowledge discovery and data mining*. Springer, págs. 160-172.
- Cao, Longbing (2017). “Data science: a comprehensive overview”. En: *ACM Computing Surveys (CSUR)* 50.3, págs. 1-42. DOI: 10.1145/3076253.
- Cartaxo, Bruno, Gustavo Pinto y Sergio Soares (2020). “Rapid reviews in software engineering”. En: *Contemporary Empirical Methods in Software Engineering*. Springer, págs. 357-384.
- Castro-Ginard, A. et al. (2020). “Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc”. En: *Astronomy Astrophysics* 635, A45. DOI: 10.1051/0004-6361/201937386.
- Castro-Ginard, Alfred, C Jordi, X Luri et al. (2019). “Hunting for open clusters in Gaia DR2: the Galactic anticentre”. En: *Astronomy & Astrophysics* 627, A35.
- Castro-Ginard, Alfred, C Jordi, Xavier Luri et al. (2018). “A new method for unveiling open clusters in Gaia-New nearby open clusters confirmed by DR2”. En: *Astronomy & Astrophysics* 618, A59.
- Chang, Kang-Tsung (2008). *Introduction to geographic information systems*. Vol. 4. McGraw-Hill Boston. ISBN: 9781260920710.
- Derkink, Annelotte et al. (ago. de 2024). “Disk Evolution Study Through Imaging of Nearby Young Stars (DESTINYS): PDS 111, an old T Tauri star with a young-looking disk”. En: 688, A149, A149. DOI: 10.1051/0004-6361/202348555. arXiv: 2406.04160 [astro-ph.EP].

- Dias, W. S. et al. (2002). “New catalogue of optically visible open clusters and candidates”. En: *AA* 389.3, págs. 871-873. DOI: 10.1051/0004-6361:20020668. URL: <https://doi.org/10.1051/0004-6361:20020668>.
- Dobbs, C. L. y J. E. Pringle (2010). “Age distributions of star clusters in spiral and barred galaxies as a test for theories of spiral structure”. En: *Monthly Notices of the Royal Astronomical Society* 409.1, págs. 396-404. DOI: 10.1111/j.1365-2966.2010.17323.x.
- Donoho, David (2017). “50 years of data science”. En: *Journal of Computational and Graphical Statistics* 26.4, págs. 745-766. DOI: 10.1080/10618600.2017.1384734.
- Ester, Martin et al. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise”. En: *kdd*. Vol. 96. 34, págs. 226-231.
- Estivill-Castro, Vladimir (2002). “Why so many clustering algorithms: a position paper”. En: *ACM SIGKDD explorations newsletter* 4.1, págs. 65-75. DOI: 10.1145/568574.568575.
- Fayyad, Usama, Gregory Piatetsky-Shapiro y Padhraic Smyth (1996). “From data mining to knowledge discovery in databases”. En: *AI magazine* 17.3, págs. 37-37. DOI: 10.1609/aimag.v17i3.1230..
- Fluke, Christopher J y Colin Jacobs (2020). “Surveying the reach and maturity of machine learning and artificial intelligence in astronomy”. En: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.2, e1349.
- Gaia Collaboration et al. (2023). “Gaia Data Release 3: Summary of the content and survey properties”. En: 674, A1, A1. DOI: 10.1051/0004-6361/202243940. arXiv: 2208.00211 [astro-ph.GA].
- Han, Jiawei, Jian Pei y Hanghang Tong (2012). *Data mining: concepts and techniques*. Morgan kaufmann. ISBN: 978-0-12-381479-1.
- Huerta, E. A. et al. (2020). “Convergence of artificial intelligence and high performance computing on NSF-supported cyberinfrastructure”. En: *Journal of Big Data* 7.1. ISSN: 2196-1115. DOI: 10.1186/s40537-020-00361-2. URL: <http://dx.doi.org/10.1186/s40537-020-00361-2>.
- Hunt, Emily L y Sabine Reffert (2021). “Improving the open cluster census-I. Comparison of clustering algorithms applied to Gaia DR2 data”. En: *Astronomy & Astrophysics* 646, A104.
- (2023). “Improving the open cluster census. II. An all-sky cluster catalogue with Gaia DR3”. En: *Astronomy & Astrophysics* 673, A114. DOI: 10.1051/0004-6361/202346285.
- Ikotun, Abiodun M et al. (2023). “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data”. En: *Information Sciences* 622, págs. 178-210.
- Kanwar, Sandeep Singh, Sanjeev Kumar y Mangey Ram, eds. (2020). *Soft Computing: Techniques in Engineering Sciences*. Vol. 1. De Gruyter Series on the Applications of Mathematics in Engineering and Information Sciences. Berlin, Boston: De Gruyter. ISBN: 9783110628234.
- Kirk, D. y W. Hwu (2010). *Programming Massively Parallel Processors: A Hands-on Approach*. 1st. San Francisco, USA: Morgan Kaufmann Publishers Inc. ISBN: 0123814723.

- Kitchenham, Barbara et al. (2009). "Systematic literature reviews in software engineering—a systematic literature review". En: *Information and software technology* 51.1, págs. 7-15. DOI: 10.1016/j.infsof.2008.09.009.
- Kriegel, Hans-Peter et al. (2011). "Density-based clustering". En: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.3, págs. 231-240. DOI: 10.1002/widm.30.
- Li, Deren, Shuliang Wang y Deyi Li (2015). *Spatial data mining*. Springer. ISBN: 978-3-662-56936-8. DOI: 10.1007/978-3-662-48538-5.
- Li, Deren, Shuliang Wang, Hanning Yuan et al. (2016). "Software and applications of spatial data mining". En: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6.3, págs. 84-114. DOI: 10.1002/widm.1180.
- Malerba, Donato (2008). "A relational perspective on spatial data mining". En: *International Journal of Data Mining, Modelling and Management* 1.1, págs. 103-118.
- Manjula, Aakunuri y G Narsimha (2014). "A review on spatial data mining methods and applications". En: *International Journal of Computer Engineering and Applications* 7.1.
- Mennis, Jeremy y Diansheng Guo (2009). "Spatial data mining and geographic knowledge discovery—An introduction". En: *Computers, Environment and Urban Systems* 33.6, págs. 403-408. DOI: 10.1016/j.compenvurbsys.2009.11.001.
- Paíz, Leonardo G et al. (2025). "A New Open Cluster Census in the Region of Cygnus OB2 with Gaia Data". En: *International Journal of Astronomy and Astrophysics* 15.2, págs. 171-196.
- Park, Hae-Sang y Chi-Hyuck Jun (2009). "A simple and fast algorithm for K-medoids clustering". En: *Expert systems with applications* 36.2, págs. 3336-3341.
- Pratihaar, Dilip Kumar (2013). *Soft computing: fundamentals and applications*. Alpha Science International, Ltd.
- Rico, Sergio et al. (2020). "Guidelines for conducting interactive rapid reviews in software engineering—from a focus on technology transfer to knowledge exchange". En: URL: 10.5281/zenodo.4327725.
- Rottoli, Giovanni Daián (2022). "Formulación de procesos para una ingeniería de explotación de información espacial". Tesis doct. Universidad Nacional de La Plata.
- Sarro, L. M. et al. (2014). "Cluster membership probabilities from Gaia-type data: A qualitative and quantitative comparison of methods". En: *Astronomy Astrophysics* 563, A45. DOI: 10.1051/0004-6361/201322744.
- Sharda, Ramesh, Dursun Delen y Efraim Turban (2018). *Business intelligence, analytics, and data science: a managerial perspective*. 4.^a ed. Pearson. ISBN: 978-0-13-463328-2.
- Wieringa, Roel (2016). *Design science methodology for information systems and software engineering*. Springer. ISBN: 978-3-662-52446-6. DOI: 10.1007/978-3-662-43839-8.
- Wu, K. y H.D. Simon (2021). "High-Performance Computational Intelligence and Forecasting Technologies". En: Wiley. Cap. 22. ISBN: 9781119482086.