UC | Facultad de Ciencias

Clasificación y derivación de propiedades de los AGNs mediante técnicas de aprendizaje automático
(Classification and Derivation of AGN Properties Using Machine Learning Techniques)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN FÍSICA DE PARTÍCULAS Y DEL COSMOS

Autor: Andrés Matres Cazorla

Directora: Amalia Corral Ramos

Septiembre - 2025

Resumen

El uso de algoritmos de Machine Learning se ha consolidado en astronomía como una herramienta esencial capaz de tratar los volúmenes de datos que ésta genera. En este trabajo se explora la aplicación de métodos supervisados y no supervisados en la caracterización de fuentes astronómicas por medio de catálogos multifrecuencias. Así, se usaron DES, VISTA, WISE y XMM-Newton como catálogos principales, además de SDSS y 2MASS en los datos de validación.

Se aplicaron técnicas de reducción de dimensionalidad por medio de una PCA, junto a métodos de clustering tales como Random Forest no supervisado y HDBSCAN para analizar la capacidad de separar galaxias, cuásares y estrellas, usando únicamente información fotométrica y flujos en rayos X. Además, se implementaron modelos de regresión basados en Random Forest para la predicción de magnitudes en el infrarrojo medio y de parámetros físicos derivados del ajuste de SEDs con CIGALE tales como la masa estelar, el SFR o la fracción de AGN.

Los resultados valoran la incorporación de flujos en rayos X como una mejora sustancial en la separación de poblaciones a costa de la reducción de estas mismas. Como remedio, se añadió el uso de límites superiores, resultando en la introducción de ruido. En regresión, la magnitud K resultó clave para la predicción en el infrarrojo medio y las propiedades como la masa o la luminosidad estelar parecen responder bien a los métodos predictivos. Este estudio confirma la utilidad de los métodos de Machine Learning en astronomía, a la vez que señala sus limitaciones y posibles mejoras con datos más profundos, numerosos y modelos más avanzados.

Palabras clave: Aprendizaje automático, Cartografiados astronómicos, Núcleos Activos de Galaxias, Galaxias, XMM-Newton, SDSS, DES, VISTA, WISE.

Abstract

The use of Machine Learning algorithms has become a consolidated and essential tool in astronomy, capable of handling the large volumes of data generated. In this work, we explore the application of supervised and unsupervised methods for the characterization of astronomical sources through multiwavelength catalogs. The main datasets employed were DES, VISTA, WISE, and XMM-Newton, complemented with SDSS and 2MASS for validation.

Dimensionality reduction techniques were applied using PCA, combined with clustering methods such as unsupervised Random Forest and HDBSCAN, to analyze the ability to separate galaxies, quasars, and stars using only photometric information and X-ray fluxes. In addition, Random Forest regression models were implemented to predict mid-infrared magnitudes and physical parameters derived from SED fitting with CIGALE, such as stellar mass, SFR, and AGN fraction.

The results highlight the inclusion of X-ray fluxes as a substantial improvement for source

separation, although at the cost of reduced sample size. As a solution, upper limits were introduced, but these led to additional noise. For regression tasks, the K-band magnitude proved to be key for mid-infrared predictions, while properties such as stellar mass and luminosity showed good consistency with predictive models. This study confirms the usefulness of Machine Learning methods in astronomy, while also pointing out their limitations and the potential for improvement with deeper, larger datasets and more advanced models.

Key words: Machine Learning, Astronomical Surveys, Active Galactic Nuclei, Galaxies, XMM-Newton, SDSS, DES, VISTA, WISE.

Índice

Agradecimientos

Quiero aprovechar esta página para agradecer todas las personas que han contribuido a que pudiera afrontar y terminar mi TFM. En primer lugar a mi familia, por darme la oportunidad de vivir en Santander para cumplir mi sueño. A mi tutora, Amalia, con quien he compartido muchas horas intentando convertir este en un trabajo respetable, así como al Dr. Giorgos Mountrichas, por haberme permitido acceder a sus datos en el ámbito de las SEDs. A mis abuelas, Teresa y Maruja, de 83 y 94 años respectivamente. Cada día que paso con ellas es una aventura que me han marcardo y me marcarán el resto de mi vida. A mis compañeros del Máster, por mantenernos unidos cuando la tormenta soplaba más fuerte. También a mis amigos, a Gema, Nerea y Esteban, por su apoyo incondicional aún cuando sin entender de inteligencia artificial consideraron dedicarme su tiempo para quejarme de lo mal que iba el código. A Cristina, por adoptarme como uno más de su familia aún cuando la mía está al otro lado del estrecho y a Fran, mi viejo compañero de batallas de la carrera y mi gran inspiración para aprender a superarme cada día.

Por último quiero citar una pequeña frase que ha acompañado mi camino a lo largo de este último año.

> *Un viaje tendrá dolor y fracaso. No son sólo los pasos adelante los que debemos aceptar. Son los tropiezos. Los juicios. El conocimiento de que fracasaremos. Que lastimaremos a quienes nos rodean. Pero si nos detenemos, si aceptamos la persona que somos cuando fallamos, el viaje termina. Ese fracaso se convierte en nuestro destino.*

1. Introducción

1.1. Motivación

Las técnicas de Machine Learning (ML) se han convertido, a lo largo de la última década, en una herramienta fundamental en la astronomía moderna. En un entorno donde es necesario analizar una cantidad ingente de datos generados por telescopios, misiones y simulaciones, cualidades como el tratamiento masivo de información son deseables e incluso necesarias. Tradicionalmente, su contraparte algorítmica suele quedarse corta tratando con grandes cantidades de datos complejos, generando un tiempo de computación que los hace mucho menos eficientes [1]. A diferencia de los algoritmos tradicionales, el aprendizaje automático se ve beneficiado por esta cualidad, poniendo un mayor énfasis en la calidad de los datos. Estos han permitido clasificar eficientemente objetos celestes, detectar anomalías, e incluso tratar de predecir eventos cósmicos. Por ejemplo, algunos algoritmos de ML han servido como instrumento para identificar exoplanetas a partir de sus curvas de luz [2], han sido usados en la clasificación de galaxias [3] o han permitido generar filtros de ruido para depurar ondas gravitacionales [4]. Teniendo en consideración que la tendencia respecto al volumen de datos astronómicos continúa creciendo exponencialmente con proyectos como el observatorio Vera C. Rubin[1] o el próximo gran observatorio de rayos-X *NewAthena*[2], los algoritmos de ML tienen por delante un ecosistema ideal que les beneficia en la exploración del cosmos.

Así pues, la principal motivación detrás de este trabajo es la exploración de las capacidades de métodos de ML supervisados y no supervisados en la construcción y validación de catálogos astronómicos. Sin embargo, los resultados y aplicaciones no deben estar restringidos solo a ellos. Este estudio además pretende contribuir con sus resultados a otros campos científico/técnicos con un entorno similar, tomando un carácter multidisciplinar.

1.2. Objetivos

Los objetivos pueden dividirse principalmente en dos puntos:

- Comprobar la fiabilidad de separar y delimitar fuentes astronómicas por medio de flujos y magnitudes fotométricas. Aún cuando las observaciones espectroscópicas son precisas y permiten obtener propiedades de los objetos astronómicos, consumen un mayor tiempo de observación que su contraparte fotométrica, generando en total muchos menos datos. Además, los datos espectroscópicos no siempre están disponibles para todas las fuentes. Así pues, uno de los objetivos es desarrollar un método de clustering que permita separar clases astronómicas sencillas (cuásares, estrellas y galaxias) mediante la disposición de un espacio que favorezca la separación de grupos.
- Con métodos supervisados, estudiar la calidad de predicción de las magnitudes en el infrarrojo medio (W1, W2 y W3) a partir de magnitudes y colores del infrarrojo cercano y óptico, además de los flujos en X. Estos permiten asegurar un método para

[1] https://rubinobservatory.org/es
[2] https://www.cosmos.esa.int/web/athena#

completar catálogos astronómicos donde puedan faltar datos en alguna de sus magnitudes. Además, aprovechar los algoritmos supervisados para ofrecer una alternativa a los métodos de cálculo de propiedades intrínsecas de galaxias tales como la masa estelar, el *star formation rate* (SFR) o la fracción de AGNs, los cuales tradicionalmente se han calculado por medio de algoritmos basados en el ajuste de SEDs.

1.3. Distribuciones espectrales de energía (SEDs) de fuentes astronómicas

Con el objetivo de caracterizar las propiedades de los objetos astronómicos, existen dos técnicas complementarias capaces de extraer información fundamental a partir de la luz emitida por los objetos celestes. Estas son **la fotometría** y **la espectroscopía**

La **fotometría** es una técnica astronómica la cual mide el flujo o intensidad de energía que emite un astro desde el cielo. La fotometría produce imágenes al someter a un instrumento fotométrico, generalmente una CCD, a la exposición de los fotones procedentes de distintas fuentes astronómicas en el cielo. A más fotones incidan sobre una región concreta, mayor intensidad es generada en el mapa y por tanto, más intensa es una fuente [5].

Aún cuando la intensidad total o intensidad bolométrica es importante para caracterizar fuentes mediante el flujo total de fotones, es más común realizar imágenes en distintas bandas del espectro. Para ello se hace uso de los filtros, los cuales permiten limitar el número de fotones a los pertenecientes a una banda conocida y por tanto limitar la capacidad de intromisión de fuentes externas a las bandas de estudio. Los filtros además pueden aportar información física cuya emisión está limitada a la sensibilidad de los instrumentos de emisión. Por ejemplo, una característica fundamental de los cuásares es la emisión en rayos X, así pues un estudio en esta banda permitirá delimitar mejor sus propiedades.

Para la medida de la intensidad en fotometría, por tradición histórica, se expresan en términos de magnitudes relativas m o absoluta M. Por nomenclatura, también puede asociarse a la magnitud en una banda el nombre del filtro, por ejemplo, la magnitud del filtro "g", se denomina g y su magnitud absoluta, G.

Para una magnitud m genérica, su expresión matemática puede definirse como:

$$m - m_{\text{ref}} = -2{,}5 \log \left(\frac{F}{F_{\text{ref}}} \right) \tag{1}$$

Donde m_{ref} es la magnitud de referencia. Históricamente, asociada al sistema *VegaMAG*, donde el flujo de referencia es el flujo de la estrella Vega, aunque actualmente existen sistemas basados en flujo como *ABMAG* y *STMAG*.

Las magnitudes relativas, como su nombre indica, reciben su nombre debido a su dependencia del instrumento de medida y por tanto son relativas la posición espacial de la medición [5]. Para asociar una medida absoluta a la fuente, se define la magnitud absoluta, por medio del módulo de distancia:

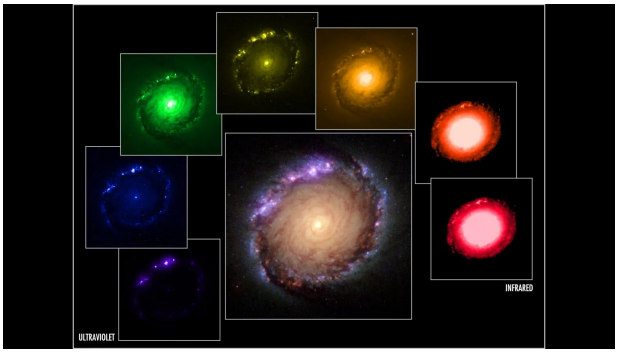


Figura 1: Imagenes multibanda referidas al mismo objetos astronómico. Fuente: https://bigthink.com/starts-with-a-bang/photometry-astronomy

$$\mu = m - M = 5\,(\log d - 1) \tag{2}$$

Donde el valor de la distancia permite recuperar el promedio de fotones perdidos debido a la expansión esférica de la radiación.

Las distintas bandas fotométricas a lo largo del espectro nos dan información sobre los fenómenos físicos asociados a las distintas estructuras internas de la fuente, ver Fig.1. Este trabajo se fundamenta en aprovechar la gran cantidad de datos fotométricos disponibles en regiones diferentes del espectro para una gran diversidad de fuentes astronómicas y optimizar la información que se puede obtener a partir de ellos.

La **espectroscopía** es la segunda técnica de medida astronómica por excelencia. Consiste en representar la intensidad de la radiación medida en función de la longitud de onda a lo largo de una banda en la imagen. Esta técnica permite generar un espectro característico el cual contiene propiedades físico-químicas. Por ejemplo, es posible medir la velocidad de rotación de una galaxia a partir del corrimiento al rojo de su espectro, además de la metalicidad de su núcleo.

Fotometría y espectroscopia parten de una naturaleza similar: ambas toman mediciones de fotones en bandas de longitud de ondas, sin embargo, la mayor diferencia radica en la información que representan. La fotometría es capaz de dar una visión espacial gracias a su intensidad asociada a sus dos coordenadas espaciales a costa de promediar los fotones recibidos en la banda asociando un único valor de intensidad. La espectroscopia por el contrario, pierde una dimensión espacial para bidimensionalizar el espectro de intensidad - longitudes de ondas.

Por supuesto, en términos prácticos, la fotometría es una técnica menos costosa debido a su menor tiempo de exposición y por tanto, por su mayor número de datos disponibles.

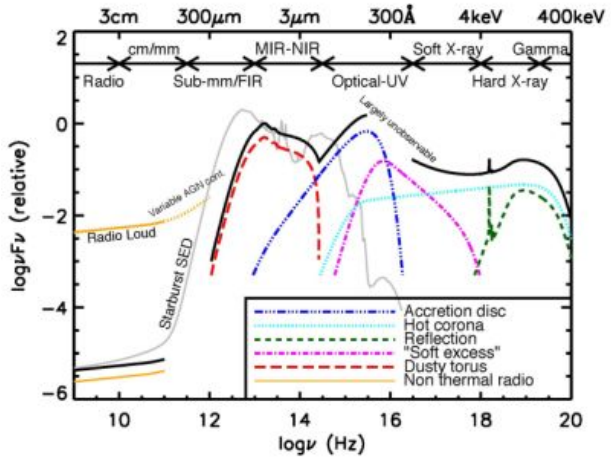


Figura 2: Ejemplo de componentes de las SEDs en AGNs. Fuente: [6]

Las **SEDs o Spectral Energy Distribution** son representaciones de la energía emitida por las fuentes astronómicas a lo largo del espectro electromagnéticos. Estas permiten caracterizar propiedades de los cuerpos astronómicos debido a la información que ellas aportan de las propiedades físicas y sus mecanismos de emisión. En general, coincidiendo más adelante con la muestra de validación escogida, las SEDs son sensibles a clases concretas de cuerpos astronómicos: estrellas, galaxias y núcleos activos de galaxias(AGNs). Los AGNs son fenómenos celestes altamente energéticos que se desencadenan en el centro de las galaxias cuando el agujero negro supermasivo central (SMBH) acreta materia procedente de sus proximidades de forma continua, afectando potencialmente a la evolución de la galaxia que lo hospeda.

A continuación, se detallan algunas de las ventajas que las SEDs pueden aportar a cada una de las clases astronómicas antes expuestas.

- **Estrellas:**
  - Gracias a la forma de la SED y la longitud de onda de emisión máxima puede derivarse su temperatura, ajustando la emisión a la de un cuerpo negro. Además, el estudio de sus líneas de emisión y absorción pueden derivar su composición química.
  - Combinando medidas de su luminosidad y temperatura puede obtenerse el radio de la estrella.
  - Aplicando modelos de evolución estelar, en combinación con otras medidas, puede derivarse su estado evolutivo.
- **Galaxias:**
  - Las SEDs son capaces de revelar información sobre la formación estelar, su población estelar, el polvo que contiene o incluso, la presencia de AGNs.

- Analizar los SEDs permiten reconstruir procesos físicos que ocurren en las galaxias, así como sus procesos evolutivos. Ejemplos de esto son las emisiones en el infrarrojo que pueden inferir la presencia de polvo en las galaxias mientras que emisiones importantes en el UV puede revelar nueva zonas de formación estelar.

- **Núcleo Activos de Galaxias (AGNs)**
  - Las SEDs de los AGNs son más complejas que las galaxias o estrellas debido a la contribución de los distintos componentes que lo forman: su disco de acreción, el toroide de polvo o la galaxia huésped (ver Fig. 2).
  - El *big blue bump* (UV-óptico) en las SEDs de los AGNs es una propiedad característica atribuido a su disco de acreción, mientras que el continuo en el infrarrojo se atribuye a emisión por parte del polvo el toroide.
  - Su modelizado permite determinar las contribuciones de los AGNs a la luminosidad total de la galaxia, y por tanto estimar la tasa de formación estelar y el impacto del AGN sobre la galaxia huésped.
  - Líneas de emisión, como *la serie Balmer del Hidrógeno*, pueden afectar a la observación de colores y ayudar en la determinación de redshifts fotométricos.
  - Los AGNs pueden clasificarse basados en las propiedades de sus SEDs, ayudando a separar las propiedades observacionales de las evolutivas.

Aun cuando la forma más precisa y efectiva para clasificar fuentes astronómicas sigue siendo mediante observaciones espectroscópicas, padece de ser un método observacional más costoso que la fotometría. Así, de forma alternativa, mediante fotometría, pueden construirse SEDs que abarquen vastas partes del espectro electromagnético, de forma que, mediante ajustes de SEDs "clásicos", permitan derivar propiedades fundamentales de los objetos estudiados.

En este trabajo, tal y como se indicaron en los objetivos, se utilizan catálogos fotométricos para analizar la factibilidad de separar las distintas fuentes astronómicas limitando el uso únicamente de fotometría, además de comparar la eficiencia de los métodos clásicos de ajustes de SEDs con la aplicación de técnicas de aprendizaje automático para la obtención de propiedades claves de galaxias.

2. Metodología

Este trabajo acoge dos metodologías complementarias, las cuales son reunidas en torno al aprovechamiento de catálogos astronómicos.

En primer lugar, la aplicación de **técnicas de ML no supervisado**. Los catálogos astronómicos utilizados contienen una gran cantidad de datos recogidos en sondeos a gran escala. Estos incluyen medidas fotométricas, clasificación de objetos (en algunos casos) y propiedades derivadas. El tratamiento efectivo de estos datos es esencial para responder las preguntas astrofísicas que requieren de un acercamiento más robusto capaz de manejar su volumen y complejidad. Para este labor se usaron técnicas de ML no supervisadas, es decir, se buscaron estructuras en nuestros datos aplicando técnicas de **clustering**, las cuales no necesitan etiquetación previa.

En segundo lugar, los **métodos de ML supervisados**. Estos son usados para tratar patrones, clasificar fuentes y predecir propiedades físicas con los datos disponibles. Los métodos supervisados son especialmente útiles cuando se dispone de una muestra de entrenamiento la cual se conocen las propiedades a predecir o clasificar, permitiendo al modelo adaptarse para tomar predicciones específicas, tal como el tipo de objeto o la estimación del SFR o el redshift.

Integrar ambas aproximaciones, permiten construir un entorno de trabajo capaz de fomentar la extracción de resultados científicos de manera óptima a partir de grandes muestras de datos.

2.1. Selección de muestras

En este trabajo se han formado distintas muestras procedentes de diferentes catálogos fotométricos. Debido a su naturaleza astronómica, todos los catálogos utilizados son públicos. Los catálogos resultantes han sido agrupados según las bandas fotométricas que contienen:

- **Catálogo principal**: Combinación de cuatro catálogos de distintas bandas, rayos X: **XMM-Newton**, infrarrojo cercano: **VISTA-VHS**, infrarrojo medio: **ALLWISE** y visible: **DES**; y que incluye todas las **fuentes detectadas en el infrarrojo y en el óptico en una selección de campos observados por *XMM-Newton*** [7]. Los catálogos en distintas bandas se cross-correlacionaron usando la herramienta `x-match` desarrollada durante el proyecto ARCHES[3], la cual permite cross-correlacionar un número arbitrario de catálogos mediante estadística bayesiana proporcionando probabilidad de asociación y no asociación[8]. Esto permitió asociar medidas fotométricas en varias bandas a fuentes únicas. Este catálogo incluye todas las fuentes detectadas en el infrarrojo en campos seleccionados observados con *XMM-Newton*, por lo tanto se trata de un catálogo heterogéneo pero representativo de los diferentes tipos de fuentes astronómicas (estrellas, galaxias, AGNs,...). La descripción detallada tanto de la cross-correlación como de la muestra resultante se puede encontrar en [9].

[3] http://www.arches-fp7.eu/arches/localhost_88/arches/index.html

- **Catálogo de validación**: Dado que la mayor parte de nuestro catálogo principal no contiene identificaciones, se utilizó un catálogo fotométrico, con clases de fuentes astronómicas determinadas a partir de espectroscopia óptica, en su mayoría procedente del **SDSS**[4], pero también de otros cartografiados espectroscópicos como WiggleZ[5], GAMA[6], OzDES DR1[7], 2QZ[8] y el 6dF Galaxy Survey[9]. Esta compilación de clasificaciones de fuentes astronómicas forma parte del proyecto **VEXAS** [10], que busca contrapartidas en otras longitudes de onda de fuentes infrarrojas. A diferencia del catálogo principal, aquellos objetos en los que fotometría procedente de **VISTA** no estuviese disponible, sus magnitudes son sustituidas por las de **2MASS**.
- **Catálogos de galaxias:** Varios catálogos compilados con el objetivo de determinar las propiedades de las galaxias (SFR, masa estelar, etc), con o sin un AGN en sus centros, mediante el ajuste de sus SEDs. Estos catálogos combinan fotometría en el óptico de VST-ATLAS [10], CFHTLS [11] (CanadaFranceHawaii Telescope Legacy Survey), COSMOS/Subaru Suprime-Cam [12], Dark Energy Survey (DES) y Pan-STARRS1 [13] (PS1) , VISTA en el infrarrojo cercano, Spitzer/IRAC [14] y Spitzer MIPS [15] en el medio y Herschel/PACS [16] y Herschel/SPIRE en el lejano [17] .

A continuación, se describen las principales características de los catálogos utilizados para construir las muestras de trabajo.

2.1.1. X-ray Multi-Mirror Mission - Catálogo XMM-Newton

También conocido como XMM, *XMM-Newton* es el gran observatorio de rayos-X de la agencia espacial europea (ESA). Su funcionamiento, basado en la cuenta de fotones y el funcionamiento simultáneo de todos sus instrumentos a bordo, permite la obtención de imágenes (y su fotometría asociada), espectros y curvas de luz en una misma observación. Además, debido a su amplio campo de visión de 30 arcmin (FoV), permite la detección y estudio de entre 50-100 fuentes por observación. Todo esto ha permitido que, aunque el área total observada hasta ahora sea de solo $\sim$1300 deg^2 (ver Fig. 3), se construya el mayor catálogo de fuentes de rayos-X con datos por encima de 5 KeV hasta la fecha: el

[4] https://sdss.org/
[5] https://wigglez.swin.edu.au/site/forward.html
[6] https://www.gama-survey.org/
[7] https://www.mso.anu.edu.au/ozdes/DR1
[8] https://www.2dfquasar.org/
[9] http://www.6dfgs.net/
[10] https://www.eso.org/public/teles-instr/paranal-observatory/surveytelescopes/vst/surveys
[11] https://www.cfht.hawaii.edu/Science/CFHTLS
[12] https://hsc-release.mtk.nao.ac.jp/doc/index.php/s17a-wide-cosmos
[13] https://outerspace.stsci.edu/display/PANSTARRS
[14] https://irsa.ipac.caltech.edu/data/SPITZER/docs/irac
[15] https://irsa.ipac.caltech.edu/data/SPITZER/docs/mips
[16] https://www.cosmos.esa.int/web/herschel/pacs-overview
[17] https://www.cosmos.esa.int/web/herschel/spire-overview

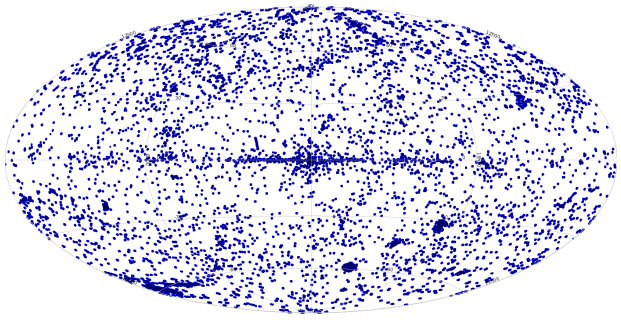


Figura 3: Mapa de apuntados de *XMM-Newton*. Fuente: XMM-SSC

XMM-Newton Serendipitous Source Catalog[18].

En particular, se ha utilizado el catálogo 4XMM, cuya decimocuarta versión (Data Release 14) fue publicada en julio de 2024. Este catálogo contiene un total de un poco más de un millón de detecciones, correspondientes a alrededor de 700.000 fuentes únicas, cubriendo energías entre **0.2 y 12 keV**.

2.1.2. Dark Energy Survey - DES

El Dark Energy Survey[19] es un proyecto internacional destinado a la investigación de la expansión acelerada del Universo y el crecimiento de estructuras a gran escala. El catálogo cubre datos profundos en el espectro visible e infrarrojo cercano en unos 5.000 $\deg^2$ de área en el hemisferio sur del cielo (1/8 del cielo, ver Fig. 4), detectando más de 300 millones de fuentes [11]. El telescopio utilizado es el Telescopio Blanco de 4 m (Cerro Tololo, Chile) equipado con la Dark Energy Camera (DECam), una cámara de 570 megapíxeles.

Su catálogo fotométrico cubre el óptico y parte del infrarrojo cercano en las siguientes bandas y con las siguientes magnitudes límite: **g** (475 nm; 24.3 mag), **r** (635 nm; 24.1 mag), **i** (775 nm; 23.3 mag), **z** (925 nm; 22.5), **Y** (1,000 nm; 21.2 mag).

En resumen, se trata de un cartografiado profundo en 5 bandas, cubriendo 1/8 del cielo, con un alcance fotométrico mucho más profundo que SDSS o 2MASS, diseñado para una cosmología de precisión.

2.1.3. VISTA-VHS

El VISTA Hemisphere Survey (VHS) es un proyecto que tiene como objetivo observar de forma uniforme, en la banda del infrarrojo cercano, el hemisferio sur, con la excepción de algunas áreas seleccionadas para observaciones más profundas [12]. Se trata de un mapeado amplio realizado con el telescopio VISTA (Visible and Infrared Survey Telescope for Astronomy) en Cerro Paranal (Chile), usando cámara VIRCAM en el infrarrojo cercano.

El área total prevista es de $\sim$ 18,000 $\deg^2$ (ver Fig. 5) en varios filtros en el infrarrojo

[18] http://xmmssc.irap.omp.eu/cat.html
[19] https://www.darkenergysurvey.org/

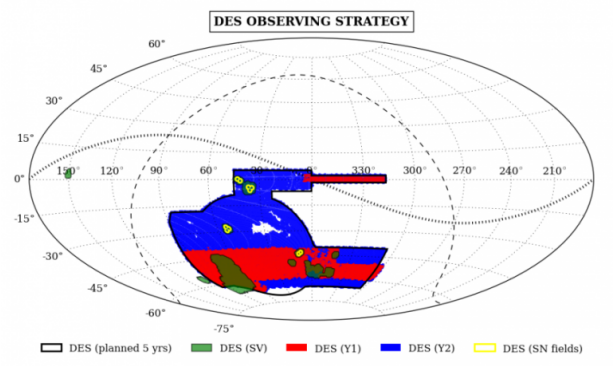


Figura 4: Mapa de cobertura de las misiones DES. Fuente: darkenergysurvey.org

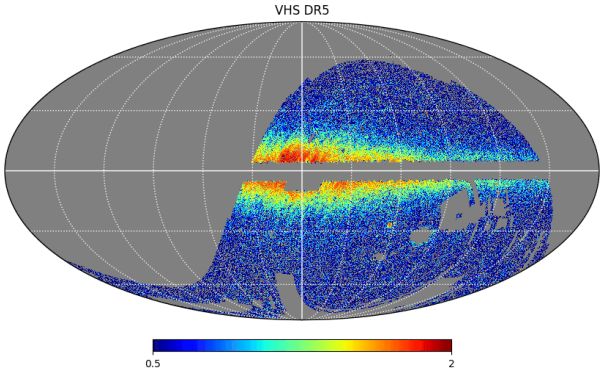


Figura 5: Densidad de objetos (por grado al cuadrado) según la proyección de Mollweide para VHS DR5. Fuente: NOIRLab

cercano. Los filtros usados en este trabajo y las magnitudes límite típicas para el VHS son: **J** (1.25 μm; 20.2 mag), **H** (1.63 μm; 19.2 mag), **Ks** (2.15 μm; 18.1 mag).

2.1.4. WISE

La misión WISE (Wide-field Infrared Survey Explorer) fue un telescopio espacial de la NASA lanzado en diciembre de 2009 con el objetivo de cartografiar todo el cielo en el infrarrojo medio, en 4 bandas: W1: 3.4 μm; W2: 4.6 μm; W3: 12 μm y W4: 22 μm.

En este trabajo se usó el catálogo público AllWISE[20], que contiene más de 700 millones de fuentes detectadas en un muestreo de todo el cielo (all-sky, cobertura $\sim$41,253 deg^2), con una profundidad fotométrica (magnitudes límite) en cada banda de: W1: 16.5 mag; W2: 15.5 mag; W3: 11.2 mag; W4: 7.9 mag.

Para construir la **muestra final multifrecuencia** solo se tuvieron en consideración las bandas **W1, W2 y W3**, es decir, las que tenían mayor profundidad fotométrica.

[20] https://wise2.ipac.caltech.edu/docs/release/allwise/

2.1.5. Sloan Digital Sky Survey- SDSS

El SDSS[21] es un proyecto de investigación iniciado en el año 2000 (con fases posteriores: SDSS-I, II, III, IV y actualmente SDSS-V) con el objetivo de cartografiar una gran fracción del cielo en el óptico. El instrumento principal es el telescopio de 2.5 m en Apache Point Observatory (Nuevo México, EE. UU.). Su base de datos pública incluye fotometría y espectroscopía para millones de estrellas, galaxias y AGNs.

El área fotométrica cubierta por el SDSS es de ∼14,555 deg^2 (aproximadamente 1/3 del cielo) en 5 filtros de banda ancha en las siguientes longitudes de onda y magnitudes límite típicas: **u** (354 nm; 22 mag), **g** (477 nm; 22.2 mag), **r** (623 nm; 22.2 mag), **i** (763 nm; 21.3 mag), **z** (913 nm; 20.5 mag).

Para construir nuestra **muestra de validación**, aprovechamos la gran cantidad de fuentes identificadas mediante espectroscopía en el SDSS, dividiéndolas en tres clases: estrellas, galaxias y AGNs.

2.1.6. 2MASS

El proyecto 2MASS (1997-2001) fue el primero en realizar un mapa completo de todo el cielo (∼41,253 deg^2) en el infrarrojo cercano. Para ello usó dos telescopios gemelos de 1.3m en Mt. Hopkins (Arizona, EE.UU.) y Cerro Tololo (Chile). La cobertura se realizó en tres bandas fotométricas con las siguientes longitudes de onda y profundidad: **J** (1.25 μm; 15.8 mag), **H** (1.65 μm; 15.1 mag) y **Ks** (2.16 μm; 14.3 mag).

En resumen, el catálogo 2MASS[22] contiene el primer mapa completo del cielo en el infrarrojo cercano (J, H, Ks), con cientos de millones de estrellas, galaxias y AGNs, sirviendo de base para gran cantidad de estudios galácticos y extragalácticos.

2.2. Muestras finales multifrecuencia

En esta sección resumimos las características principales de las muestras utilizadas en este trabajo. En la sección 3 se detallan las modificaciones realizadas a estas muestras para poder aplicarles los métodos de ML seleccionados.

- **Catálogo principal**: Compuesto por ∼ 370,000 fuentes infrarrojas en campos observados por *XMM-Newton* con fotometría en DES, VHS y WISE; filtros g, r, i, z, J, H, Ks, W1, W2 y W3, ∼20,000 de ellas detectadas por *XMM-Newton*.
- **Catálogo de validación**: Compuesto de ∼ 237,000 fuentes infrarrojas **identificadas** mediante espectroscopia óptica y con datos fotométricos en DES, VISTA/2MASS y WISE: filtros g, r, i, z, J, H, Ks, W1, W2 y W3. De ellas, unas 5000 tienen además datos en rayos-X.

[21] sdss.org
[22] https://irsa.ipac.caltech.edu/Missions/2mass.html

- **Catálogos de galaxias**: Se han utilizado varios catálogos de galaxias y AGNs, basados en fuentes de rayos X y construidos a partir de observaciones multifrecuencia en áreas determinadas del cielo a partir del cruce del catálogo COSMOS con VISTA.

2.3. Aprendizaje automático

Una distinción clave respecto al ML es la diferencia entre los métodos supervisado y no supervisado. Los métodos supervisados aprenden de datos catalogados y entrenados a través de un modelo, donde la variable predicha en la salida (por ejemplo, la clasificación de la fuente astronómica) es conocida a priori. Esta aproximación es útil para tareas como la clasificación de fuentes o la estimación del redshift de la galaxia. Los métodos no supervisados son útiles para tareas donde los datos no están catalogados y buscan descubrir patrones ocultos o grupos en los datos. A través del clustering pueden agruparse galaxias de tipos similares o detectar fenómenos astronómicos inesperados (outliers). Ambos métodos son muy potentes por sí mismos y su combinación permite explorar e interpretar el universo más eficientemente.

Ambos métodos confían en la división de los datos en dos grupos principales:

- **Las características o features:** son propias de los métodos tanto supervisados como no supervisados, comprenden todos los datos necesarios para implementar el método en los algoritmos supervisados, o para desarrollar el espacio de parámetros en los algoritmos no supervisados. En nuestro caso se trata de las magnitudes y colores derivados a partir de ellas.
- **Los objetivos o targets:** esta distinción es propia de los métodos supervisados, y comprenden aquellas propiedades de los objetos astronómicos que se quieren predecir (masa estelar, SFR, etc) o su clasificación (estrella, galaxia o AGN).

Así pues, un algoritmo no supervisado solo depende de características mientras que uno supervisado, necesita muestras de entrenamiento con características/clases a predecir conocidas.

2.3.1. Reducción de la dimensionalidad

Tras la construcción del catálogo multibanda, surge la necesidad de reducir la cantidad de variables involucradas. Esto no solo facilita la visualización y el análisis posterior, sino que también permite descubrir patrones más claros dentro del conjunto de datos. A menudo, cuando se trabaja con grandes dimensiones, se pierde perspectiva de la estructura global del sistema, y los grupos naturales que podrían existir quedan difuminados o distorsionados por el ruido o la redundancia entre variables.

La reducción de dimensionalidad tiene como objetivo transformar el espacio original en otro de menor dimensión, procurando conservar la mayor parte de la información relevante. Esta estrategia se engloba dentro del aprendizaje no supervisado, ya que actúa directamente sobre los datos sin necesidad de contar con etiquetas previas.

En este estudio se han explorado el método del **Análisis de Componentes Principales (PCA)**

Otros enfoques, como los métodos de aprendizaje sobre variedades (por ejemplo, *Locally Linear Embedding*), se han usado en trabajos previos, que utilizan una muestra similar a las usadas en este, con resultados no concluyentes [7]. Estos métodos buscan modelar la estructura geométrica de los datos cuando se sospecha que se encuentran distribuidos sobre una superficie de dimensión inferior. Sin embargo, en este trabajo se ha optado por métodos lineales que resultan más sencillos de interpretar y presentan una mayor estabilidad frente al ruido en los datos.

Todas las técnicas mencionadas han sido implementadas utilizando bibliotecas científicas en Python, concretamente `Scikit-learn`[23] y `AstroML`[24], ampliamente reconocidas en el ámbito de la astronomía.

2.3.2. Análisis de Componentes Principales - PCA

El **PCA** es una técnica estadística que permite identificar las direcciones en las que los datos presentan mayor variabilidad. A través de esta transformación lineal, se generan nuevas variables, denominadas *componentes principales*, que son combinaciones lineales de las variables originales y están ordenadas según la cantidad de varianza que explican.

La información esencial del sistema queda concentrada en la **matriz de covarianza** de los datos, la cual es simétrica y definida positiva. Gracias a estas propiedades, es posible llevar a cabo una *diagonalización* de dicha matriz: se encuentran una base ortonormal de vectores propios (autovectores) y sus correspondientes valores propios (autovalores). Estos vectores forman los ejes del nuevo sistema de referencia definido por la PCA, y los autovalores indican la varianza explicada por cada uno de ellos. Así, los vectores propios con mayor autovalor corresponden a los componentes que mejor capturan la estructura del sistema.

Además, el método permite extraer una *matriz de transformación*, la cual proyecta cualquier vector del espacio original al nuevo sistema de componentes principales. Matemáticamente, siendo $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ el vector de variables originales, la transformación PCA se expresa como: +

$$\mathbf{z} = \mathbf{W}^\top \cdot \mathbf{x}$$

donde $\mathbf{W}$ es la matriz de transformación y $\mathbf{z} = (z_1, z_2, \ldots, z_m)^\top$ representa el nuevo vector de variables transformadas.

Desarrollando esta expresión por componentes:

[23] https://scikit-learn.org/stable/
[24] https://www.astroml.org/

$$
\begin{aligned}
z_1 &= w_{11}x_1 + w_{21}x_2 + \cdots + w_{n1}x_n \\
z_2 &= w_{12}x_1 + w_{22}x_2 + \cdots + w_{n2}x_n \\
&\vdots \\
z_m &= w_{1m}x_1 + w_{2m}x_2 + \cdots + w_{nm}x_n
\end{aligned}
$$

Cada componente z_i puede interpretarse como una combinación ponderada de las variables originales, en la que los coeficientes w_{ij} indican la contribución relativa de cada variable x_j al componente z_i.

2.4. Clustering

Los métodos de clustering o búsqueda de grupos permiten agrupar datos con características similares, aun cuando el espacio de parámetro dificulte disceñir estas relaciones. Este tipo de algoritmos se basan en un agrupamiento según la distancia relativa en el espacio y por tanto definir una similitud a partir de ello.

2.4.1. Random Forest (Clustering) - URF

El algoritmo **Random Forest** (RF) es un método de aprendizaje por conjunto basado en la construcción de un gran número de **árboles de decisión**. Cada uno de ellos es entrenado usando una porción aleatoria de los datos (*bootstrap*), además de un subconjunto aleatorio de features por cada división. Aún cuando su uso más común es en ML supervisado, cuenta con una forma de ser usado en clustering.

En el método no supervisado, RF no cuenta con etiquetas reales para entrenar. En su lugar, se genera un conjunto de datos artificiales imitando su distribución marginal pero eliminando las correlaciones entre variables. A continuación, se toma un algoritmo RF de clasificación binaria para distinguir entre el conjunto de datos reales y los sintéticos. Así, la clave del método es el hecho de que RF intente discriminar ambos conjuntos mediante sus árboles de decisión. Entre los árboles, puede capturarse la estructura interna y las relaciones presente en los datos originales.

Como resultado, se obtiene una **matriz de proximidad** entre pares de observaciones: se evalúa la fracción de árboles en los que ambas poblaciones terminan en la misma hoja. Si hay una mayor proporción, existe una alta proximidad entre ellos indica que los puntos comparten características similares según el criterio de árboles. A partir de esta matriz puede construirse una matriz de distancia a partir de:

$$
D_{ij} = \sqrt{1 - P_{ij}} \tag{3}
$$

donde (D_{ij}) es la distancia entre los objetos i y j, y P_{ij}, su proximidad media. Debido a que la probabilidad de $P_{ii} = 1$, la matriz de distancia tiene además, una diagonal nula.

Este método es muy usado para la identificación de datos anómalos, tal y como se usa en [13]. En este trabajo se utilizará unicamente para construir una matriz de distancia. Esta matriz permitirá adoptar de una mejor métrica al método de separación de grupos en clustering.

2.4.2. HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [14, 15, 16] es un algoritmo no supervisado basado en DBSCAN (Density-based Spatial Clustering of Applications with Noise) [17, 18].

DBSCAN es un algoritmo de clustering basado en densidad, es decir, permite encontrar clusters a partir de sobre-densidades de puntos en el espacio de parámetros. DBSCAN requiere dos hiperparámetros principales: `eps`, el cual define la distancia máxima entre dos puntos para ser considerado parte del mismo grupo, y `min_samples`, cuyo valor define cual es la densidad mínima a partir de la cual, un conjunto de puntos pueda ser considerado denso.

El algoritmo define tres clases de puntos, esenciales para determinar grupos:

- **Puntos núcleos:** Es un punto el cual se encuentra con una densidad cuyo valor es al menos `min_samples` en una hiperesfera de radio `eps`.
- **Punto de borde:** Es un punto el cual está rodeado de puntos a la distancia `eps`, pero no es en sí un punto núcleo (no satisface la condición de densidad).
- **Punto de ruido:** Aquellos puntos que ni son puntos núcleos ni puntos de borde.

A partir de esta clasificación, se definen cada grupos a partir de delimitar aquellos puntos núcleos delimitados por puntos bordes. Finalmente se crea un grupo de ruido para los puntos de sobrantes.

HDBSCAN es una evolución de este algoritmo. Este, en vez de contar con un radio fijo `eps`, cuenta con un radio creciente. Esto es debido a que los grupos al disminuir el valor de `eps` hace que los grupos o se dividan en grupos más pequeños o se mantengan iguales. Así, haciendo un barrido creciente, se buscan aquellos grupos que persistan más a lo largo del proceso, siendo considerado los grupos óptimo [19].

2.5. Regresión

El punto clave de los algoritmos de regresión es obtener predicciones de una o varios objetivos a partir de las características o predictores.

Su mayor diferencia con los métodos algorítmicos clásico reside en que los métodos de ML de regresión puede realizar predicciones de uno o más targets mucho más rápidas, además de adaptarse mejor a las relaciones subyacentes de los grupos de datos.

En especial, en este trabajo se ha usado el método de RF de regresión.

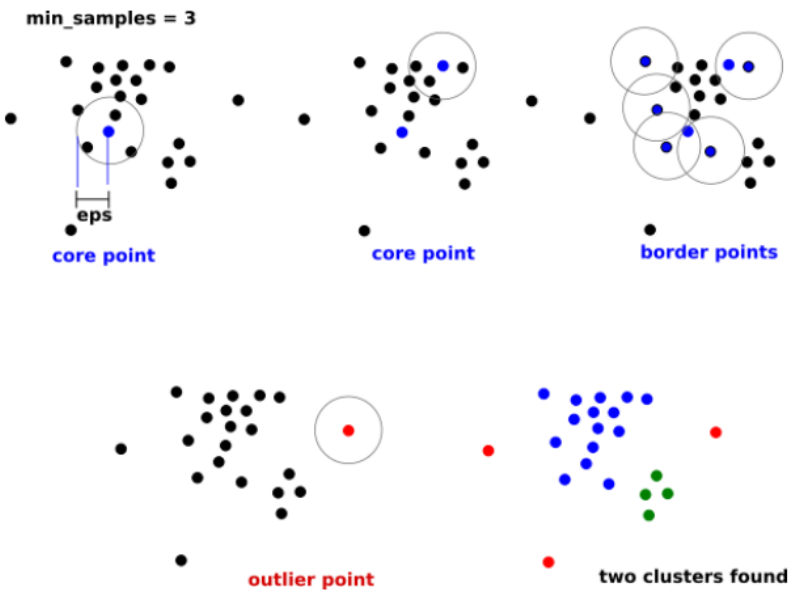


Figura 6: Imagen del proceso de separación entre puntos núcleos, borde y ruido. Fuente: [19]

2.5.1. Random Forest (Regresión)

La aplicación de RF en regresión sigue la misma lógica que el RF no supervisado: se construyen el conjunto de árboles de decisión a partir de muestras *bootstrap* de los datos originales, pero en lugar de definir una matriz de distancia, cada árbol genera una predicción numérica para el objetivo. La predicción final del modelo se obtiene a partir de un promedio de las predicciones individuales de todos los árboles.

A diferencia de los métodos paramétricos, los cuales buscan obtener una única ecuación que describa la relación entre predictores y respuesta, los modelos basados en árboles dividen el espacio de parámetros en regiones más sencillas, permitiendo captar mejor las estructuras no lineales de los datos.

En especial, es importante denotar el uso del *bagging*, el cual es clave en su funcionamiento: si cada árbol es entrenado con una muestra distinta de los datos y se introduce una carácter aleatorio en la selección de los predictores permite reducir la correlación entre los árboles, disminuyendo la varianza del modelo. De este modo el modelo es capaz de equilibrar el sesgo (desviación de las predicciones del modelo con los datos reales) con la varianza (permitiendo evitar el *overfitting* o sobreadaptación a los datos de entrenamiento [20].

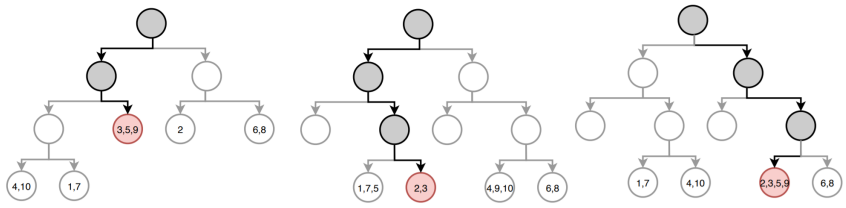


Figura 7: Predicción de objetivo por medio de árboles de decisión en un problema de regresión. Fuente: [20]

A lo largo de este trabajo, para evaluar el rendimiento de estos algoritmos, se usan los estadísticos usuales en este tipo de análisis:

- $\boldsymbol{\sigma}_{\mathrm{NMAD}}$. Es una métrica de la dispersión, muy usada en la astronomía para evaluar la precisión de predicciones. Su expresión es:

$$\sigma_{\mathrm{NMAD}} = 1{,}4826 \times \mathrm{mediana}\left(|\Delta z - \mathrm{mediana}(\Delta z)|\right) \tag{4}$$

  Donde Δz es la desviación entre el valor predicho y el valor real Es una desviación resistente a los valores anómalos debido a contar con la mediana en vez de la media.

- **Fracción de valores anómalos $\boldsymbol{\eta}$(%)**: Es una métrica que permite calcular el porcentaje de objetos astronómicos los cuales se desvían de la predicción esperada. Se define según la expresión:

$$\eta = \frac{N\left(\frac{|\Delta z|}{1+z_{\mathrm{real}}} > 0{,}15\right)}{N_{\mathrm{tot}}} \tag{5}$$

2.6. Ranking de características

Algunos métodos como la PCA o el RF cuenta con mecanismos integrados capaces de señalar las características que son más importantes para el sistema. En el caso de la PCA, por ejemplo, se utiliza la varianza explicada como medida: aquellas variables que aportan mayor varianza en los componentes principales son aquellas que mejor representan la información del sistema, y por tanto son más relevantes para este.

Random Forest supervisado usa un sistema distinto. Mediante el **MDI (Mean Decrease Impurity)** es capaz de cuantificar cuanto contribuye una característica a mejorar la pureza de los nodos entre los árboles de decisión que componen un modelo.

Esta impureza refleja cuan heterogéneo es un nodo. Si una rama en la que se ha variado una característica se mantiene más o menos constante, indica como esa característica es menos importante para la rama. Estas ramas además son ponderadas por el número de observaciones que caen en el mismo nodo, dando la importancia a aquellas ramas que mejor expliquen el total de los datos.

Como resultado se obtiene un valor numérico: cuanto mayor es el MDI de una variable, mayor es su relevancia en el modelo, ya que indica que contribuye más a separar las clases o mejorar la predicción en los árboles.

Así, gracias a estos indicadores se pueden ordenar las características según su importancia, permitiendo descartar aquellas que aporten menos al sistema.

3. Resultados

Una vez hecha la presentación de cada uno de los catálogos y los métodos que se han utilizado, es momento de presentar los resultados:

3.1. Muestra final

3.1.1. Crosscorrelación de catálogos.

Partiendo del **Catálogo Principal** descrito en la sección 2.2, la primera acción tomada ha sido actualizar el catálogo de 3XMM al 4XMM-DR14, su última versión pública. Para ello, mediante el uso de la herramienta de correlación entre catálogos de TOPCAT[25] se han localizado las coordenadas del 3XMM y las del catálogo del 4XMM, correlacionandolas con una distancia menor a 1 arcosegundo.

Una vez realizada esta actualización, el catálogo resultante contiene unas 800.000 fuentes con datos fotométricos en DES, VISTA-VHS y WISE, incluyendo probabilidades de asociación y no asociación entre las contrapartidas multifrecuencia. De esas 800.000, tan solo unas 20.000 han sido detectadas por *XMM-Newton*.

A continuación, se hizo un filtrado en calidad, seleccionando solo las asociaciones con una probabilidad superior a 2 sigmas, ya sea entre tres (DES, VHS y WISE) o cuatro catálogos (DES, VHS, WISE y XMM). Así se han obtenido dos muestras importantes para los trabajos posteriores:

- **Muestra Completa (MC)**:`catalogue_with_DES_VHS_AllWISE.fits`, con 361727 objetos con fotometría disponible del óptico al infrarrojo medio.
- **Muestra X (MX)**: `catalogue_with_DES_VHS_AllWISE_4XMM.fits` , con 15410 objetos con además flujos en rayos X.

Para evaluar los métodos de clustering, se ha seguido el siguiente enfoque:

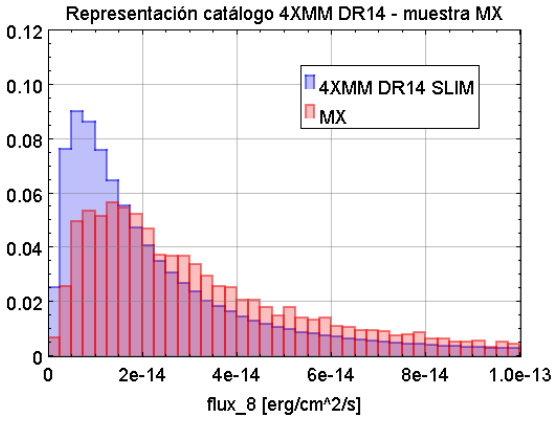
1. Aplicar técnicas de clustering solo usando datos del óptico al MIR aplicados a la MC.
2. Usando la sub-muestra MX, comprobar si la separación de fuentes mejora con la adición de datos en rayos-X.
3. Debido a que los métodos de ML tienden tendencia a mejorar con respecto a un aumento significativo de la población del catálogo, se han estimado límites superiores para las fuentes no detectadas en X en la MC para comprobar de nuevo si mejoraba la separación.

Una vez extraídas las muestras para su estudio, es necesario comprobar si estas son representativas de sus catálogos. En primer lugar, se comprobó si la muestra MX es representativa del catálogo 4XMM-DR14 del cual viene su crosscorrelación con el catálogo MC. Para ello se enfrentaron en un histograma, el flujo total en X (0.2-12 keV; `flux_8`) de ambos catálogos. La figura 8a es su resultado.

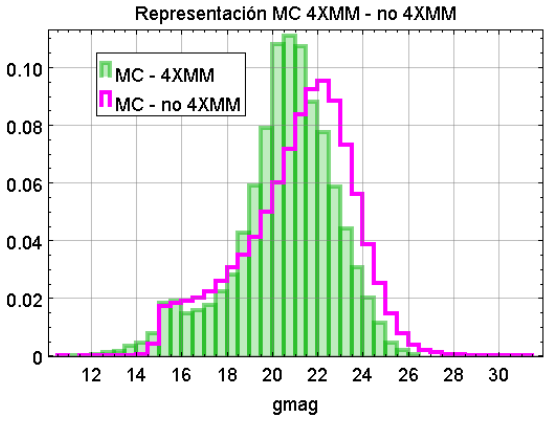
[25] https://www.star.bris.ac.uk/~mbt/topcat/

Puede observarse como la distribución de flujos de la muestra MX se separa de la del catálogo completo para flujos menores de $flux\_8 = 2\text{e-}14$ erg cm^{-2} s^{-1}. Este hecho implica que el catálogo MX reproduce bien el catálogo 4XMM-DR14 a partir de ese umbral. Esto era esperado, ya que los campos observados por *XMM-Newton* utilizados para construir el catálogo principal no incluyen las observaciones más profundas y además, al los catálogos multifrecuencia utilizados, especialmente en el MIR, no tendrían una profundidad suficiente para detectar las fuentes de rayos-X más débiles.

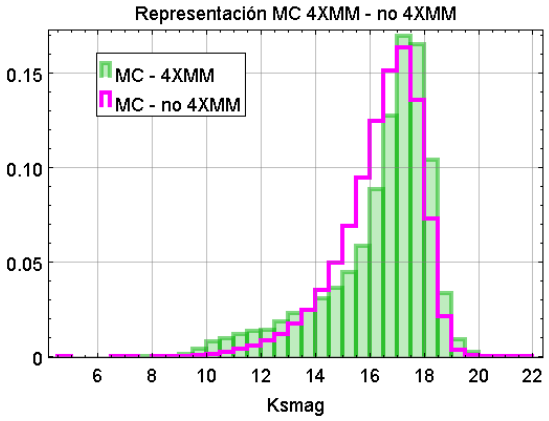
Además, se comprobó si la MC y la MX representan poblaciones de objetos similares. Para ello, se tomaron tres magnitudes en cada banda para comprobar si al cruzar con el catálogo de 4XMM ha producido un sesgo. Así, se obtuvieron las figuras 8b, 8c, 8d. Para el óptico y el MIR, vemos el mismo efecto: las fuentes detectadas en rayos-X (MX) son ligeramente más brillantes, en promedio, en estas bandas que la muestra total (MC). En cambio, la banda Ks muestra menor desviación, probablemente debido a su naturaleza más estable frente a la extinción y su relevancia en objetos de tipo AGN.



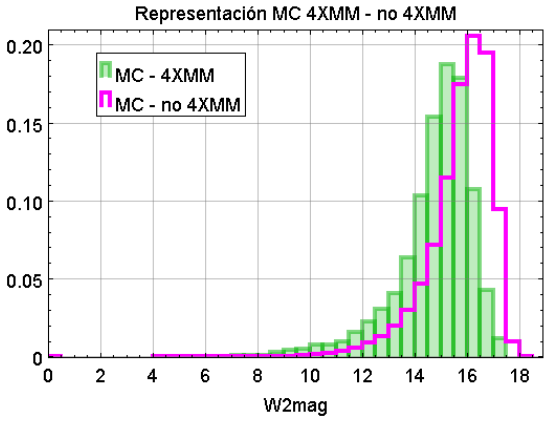
(a) Histograma flujo total en X entre catálogo 4XMM,DR14 y MX.



(b) Histograma magnitud g catálogo MC, 4XMM - no 4XMM.



(c) Histograma magnitud Ks catálogo MC, 4XMM - no 4XMM



(d) Histograma magnitud W2 catálogo MC, 4XMM - no 4XMM

Figura 8: Histogramas comparativos para comprobar la representatividad de la muestra. Los catálogos MX con el catálogo 4XMM DR14 y los datos en X de MC con los de no X de MC

3.1.2. Límites superiores

Aún cuando la población de MX es lo suficientemente representativa de la población total a estudiar, su tamaño, mucho menor que la MC, podría afectar a la eficiencia de los métodos de ML a aplicar. Sin embargo, es posible ampliar el tamaño de la muestra hasta un 2000 % estimando los flujos en rayos-X de las fuentes no detectadas mediante el cálculo de sus correspondientes límites superiores.

En un primer momento, se apostó por el uso de FLIX [21] en el cálculo de límites superiores. Esta herramienta permite adoptar límites superiores dependiendo la posición del cielo mediante un cálculo basado en modelos empíricos. Sin embargo, debido a que el volumen de datos necesarios para completar el catálogo era inmenso y que servicio se encontraba únicamente en web, impidieron que este pudiera ser usado para completar el catálogo.

Así, se llegó a la segunda opción: seleccionar aquellos valores presentes en el catálogo fotométrico que no se encuentran en el catálogo de MX. Así, se le asigna un flujo límite superior buscando aquellos flujos de MX que estén dentro del radio de 5 arco minutos. Este hecho parte de la base de que cualquiera de esos valores fotométricos habrían obtenido un flujo si este fuera mayor al valor detectado. Como no ha sido así, el valor debe ser estrictamente menor.

Esta técnica ha permitido completar el catálogo MX con datos artificiales los cuales han sido además marcados con una etiqueta booleana la cual distingue aquellos objetos que pertenecen a 4XMM (y por tanto no es un límite superior) o no pertenecen a este catálogo. Los métodos de ML contarán con esta etiqueta como una de sus características.

Conviene señalar que, si bien esta ampliación incrementa notablemente el volumen de datos disponibles, en algunos métodos puede actuar como un foco localizado de ruido, lo cual debe tenerse en cuenta en los análisis posteriores.

En la tabla 1, se indica un resumen de las características de cada catálogo generado.

Catálogo	**Condición**	**Número de datos**	**Composición de datos**
MC	$P(A \cap B \cap C) > 2\sigma$	361727	**Fotométricos:** g, r, i, z, J, H, K, W1, W2, W3
MX	$P(A \cap B \cap C \cap D) > 2\sigma$	15410	**Fotométricos:** g, r, i, z, J, H, K, W1, W2, W3 **4XMM** Banda soft (1,2,3), Banda hard (4,5), Flujo total (8)
MX + UL	$P(A \cap B \cap C \cap D) > 2\sigma + UL$	361727	Fotométricos: g, r, i, z, J, H, K, W1, W2, W3 **4XMM + Límite superior:** Banda soft (1,2,3), Banda hard (4,5), Flujo total (8) is_4XMM

Tabla 1: Resumen de las características de los catálogos utilizados. Fuente: Propio

3.1.3. Cálculo de colores

El cálculo de colores constituye un proceso fundamental para permitir una separación más nítidas de grupos en el espacio de parámetros, y, al mismo tiempo, de otorgar a la PCA

un significado físico más interpretable.

Aunque estas características no formen parte intrínsecamente de los catálogos originales, se han calculado de manera externa mediante una función debido a los numerosos métodos que se benefician de su inclusión.

Para ello la función realiza todas las combinaciones únicas de diferencia entre magnitudes fotométricas, resaltando contrastes en bandas específicas, además de permitir una mayor separación entre poblaciones de objetos.

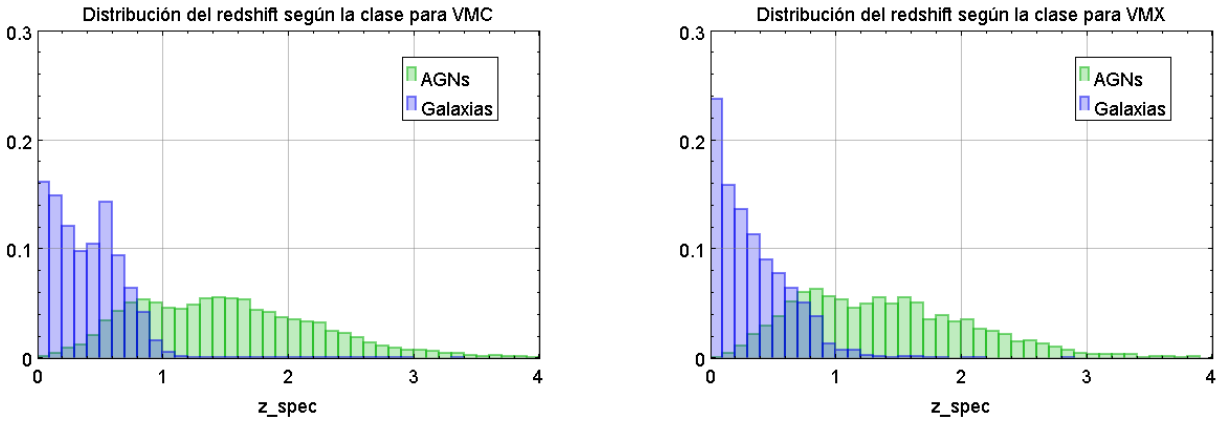
3.1.4. Datos de validación

Siguiendo la necesidad de datos de validación que corroboren los grupos sintéticos creados en el apartado de ML no supervisado, es necesario unos datos de validación alineados con los catálogos que se han usado.

A partir de esta muestra se han definido dos archivos de validación.

- **Validación de muestra completa: (VMC)**: `VEXAS_DES_VISTA_WISE_mags_zclass.fits`. Muestra extraida de los catálogos multi-frecuencia de VEXAS [22], que incluyen fotometría procedente de DES, VISTA y WISE. Seleccionamos aquellas fuentes con distancias (redshifts) y clasificaciones procedentes de espectroscopía óptica, obteniendo una muestra de 336020 fuentes, que constituyen nuestra muestra de validación para el catálogo MC.
- **Validación de muestra con X (VMX)**:`VEXAS_DES_VISTA_WISE_mags_zclass_XMM.fits` es una submuestra de la anterior cruzada con 4XMM. Contiene 4614 datos y pretende actuar como validación para el catálogo MX.

Las distribuciones de redshifs de los objetos en estas muestras de validación están representadas en la Fig. 9.

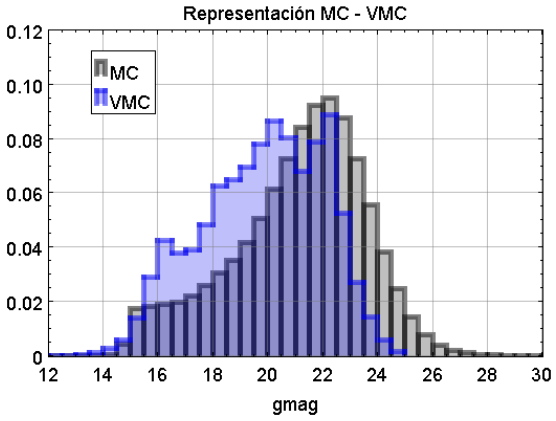


(a) Distribución en redshift de la muestra VMC - según la clase QSOs y Galaxias

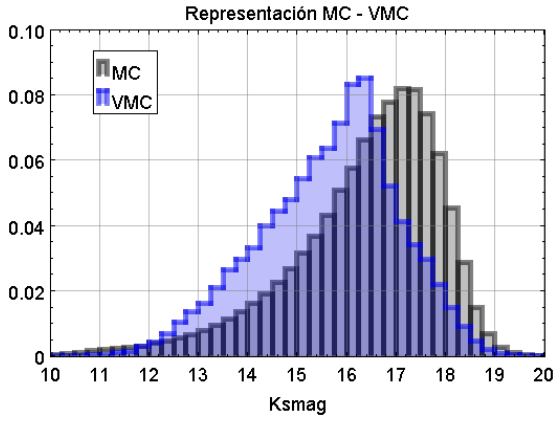
(b) Distribución en redshift de la muestra VMX - según la clase QSOs y Galaxias

Figura 9: Gráficas de representatividad en redshift para las muestras de validación VMC y VMX

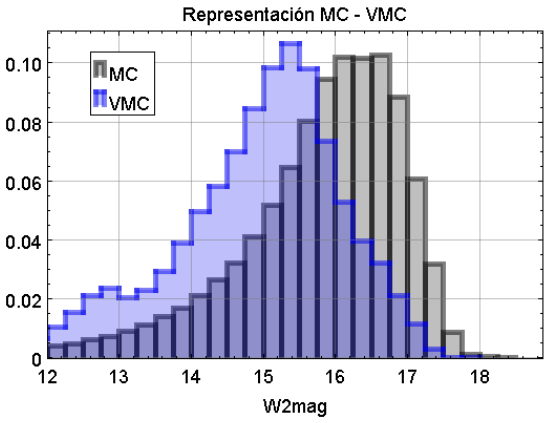
Además, se comprobó la representatividad de los datos de validación, para ello, al igual que en las figuras 8b, 8c, 8d, se obtuvieron los histogramas correspondientes 10a, 10b y 10c. Los datos de validación tienen una tendencia a representar mejor aquellos datos más brillantes, generando una discordancia entre datos de validación y la MC. Esto se debe probablemente a que sólo forman parte de la VMC aquellas fuentes con espectroscopía óptica, generalmente no disponible para las fuentes más débiles/lejanas. Se muestra también la distribución de redshift para las muestras VMC y VMX 10.



(a) Histograma magnitud g catálogo MC - VMC



(b) Histograma magnitud Ks catálogo MC - VMC



(c) Histograma magnitud W2 catálogo MC - VMC

Figura 10: Gráficas de representatividad entre los catálogos MC y su validación asociada VMC. Fuente: Propio

3.2. Clasificación de fuentes - Clustering

Una vez reunidos y limpiados los catálogos, además de asegurar unos datos de validación compatibles, se ha abordado el primer objetivo impuesto: Comprobar la posibilidad de identificar clases astronómicas usando únicamente datos fotométricos y/o flujos en X.

Por ello se seguirá el siguiente flujo de trabajo. En primer lugar se trabajará con la MC, identificando la geometría que adoptan los datos en el espacio de vectores PCs y por último, se tratará de aplicar los métodos de clustering para tratar de observar estructuras. Así mismo, a continuación, se tomará una aproximación parecida para MX y para MX+UL. Comprobaremos la geometría en el espacio de PCs además de la eficacia de las técnicas de clustering.

Los tres catálogos serán procesados por los siguientes bloques, que diferirán entre ellos por el número de datos:

1. **Reducción de dimensional (PCA)**: En este paso, reduciremos la dimensionalidad del catálogo a sus ejes principales. Este permite observar grupos visuales y contrastarlos con los marcadores de los datos de validación.
2. **Cálculo de la matriz de distancias (URF/HDBSCAN)** Los métodos de clustering necesitan definir una métrica que permita calcular distancias dentro del espacio de parámetros. Para el catálogo MX, el cual es menos poblado, se puede definir con URF una matriz de distancias adaptada al espacio de parámetros. Sin embargo, esto no es extrapolable a los catálogos más poblados como MC o MX+UL (matriz de distancias $n \times n$). Por ello, se usó en estos casos HDBSCAN, que calcula por si solo la matriz de distancia por medio de una métrica euclidiana.
3. **Determinación de los grupos sintéticos (HDBSCAN)**: HDBSCAN determina por medio de la matriz de distancia grupos de sobre-densidades. Así genera grupos delimitados los cuales son contrastados con la validación.
4. **Validación**: Los datos de validación cuentan con clases corroboradas espectroscópicamente. Esto permite aplicar los grupos sintéticos al espacio de validación y tratar de predecir cual sería el el grupo asignado. Para ello una matriz de contingencia será crucial para este hecho.

Este flujo de trabajo queda resumido en la figura 11

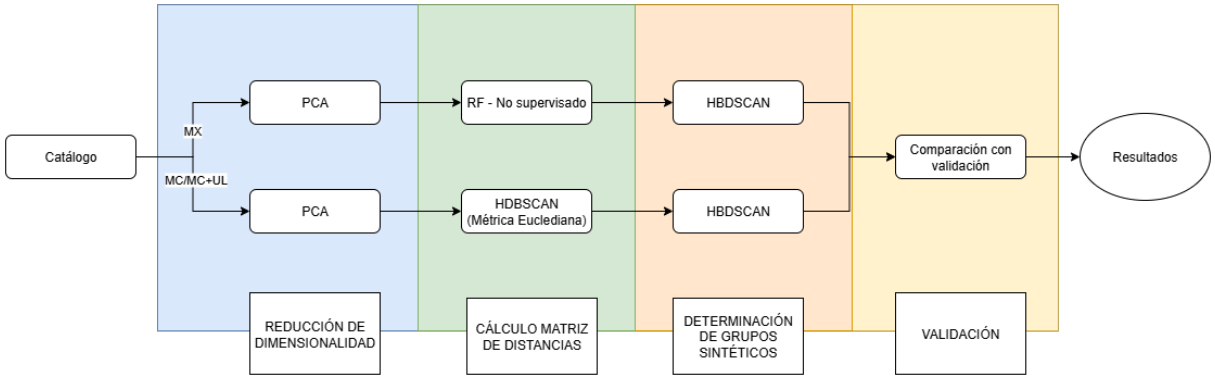


Figura 11: Esquema del flujo de trabajo desarrollado para el apartado de métodos no supervisados. Fuente: Propio

3.2.1. Exploración del catálogo MC

Este apartado tiene como objetivo aplicar el flujo de trabajo a los datos principales. Por ello, desarrollando la PCA obtenemos la figura 12.

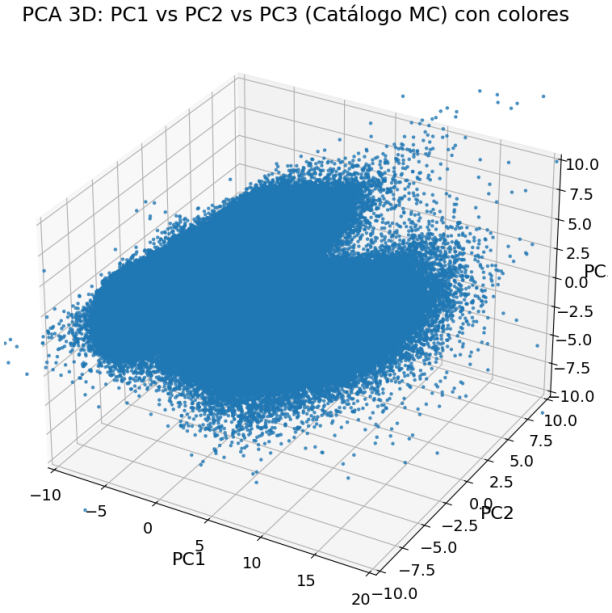


Figura 12: Representación del catálogo MC en espacio de PCA. La gran cantidad de datos dificulta observar grupos distinguibles. Fuente: Propio.

Una vez visualizado los datos, introducimos la lista de magnitudes y colores en el HDBSCAN tratando de identificar nuevos grupos. Como resultados se obtiene la figura 13

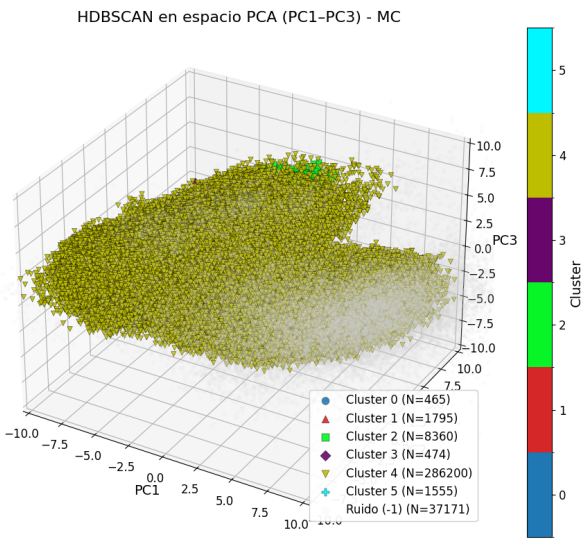
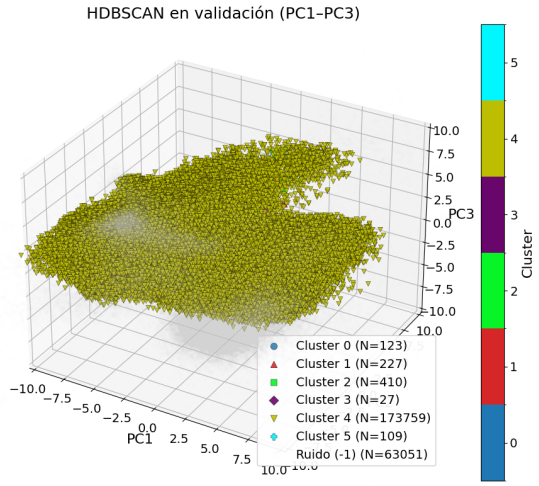
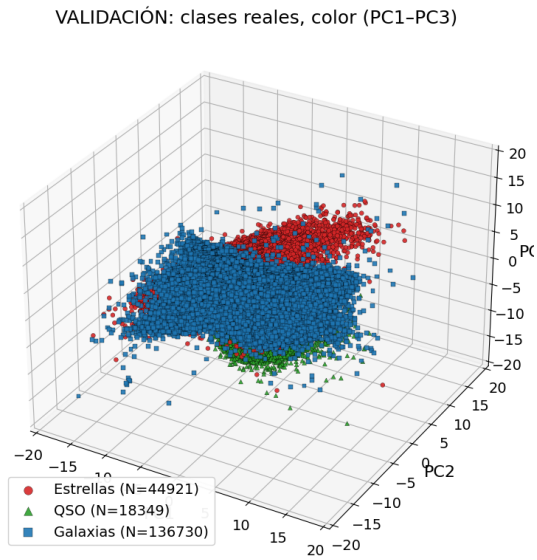


Figura 13: Representación delas predicciones de HDBSCAN para el catálogo MC en espacio de PCA. Los grupos no son representativos de la estructura interna. Fuente: Propio.

Una vez obtenido los grupos sintéticos, se aplican a los datos de validación, obteniendo las figura 21a. La figura 21b es la representación de los datos marcados con los datos de validación.



(a) Representación 3D de los grupos sintéticos aplicados a los datos de validación en espacio PC para la muestra de validación VMC



(b) Representación de las clases reales pertenecientes a los datos de validación en el espacio PC para la muestra de validación VMC

Figura 14: Resultados obtenidos al comparar los grupos sintéticos en el espacio de datos de validación con las clases reales. Fuente: Propio

Así pues, puede construirse la siguiente tabla de contingencia 2 para asociar los grupos sintéticos a los reales.

Clase	Total	Ruido (-1)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Estrellas	53364	5621 (10.53 %)	123 (0.23 %)	227 (0.43 %)	404 (0.76 %)	27 (0.05 %)	46854 (87.82 %)	108 (0.20 %)
QSO	21716	21220 (97.71 %)	0 (0.00 %)	0 (0.00 %)	2 (0.01 %)	0 (0.00 %)	494 (2.28 %)	0 (0.00 %)
Galaxias	162626	36210 (22.27 %)	0 (0.00 %)	0 (0.00 %)	4 (0.00 %)	0 (0.00 %)	126411 (77.73 %)	1 (0.00 %)
TOTAL	**237706**	**63051 (26.53 %)**	**123 (0.05 %)**	**227 (0.10 %)**	**410 (0.17 %)**	**27 (0.01 %)**	**173759 (73.10 %)**	**109 (0.05 %)**

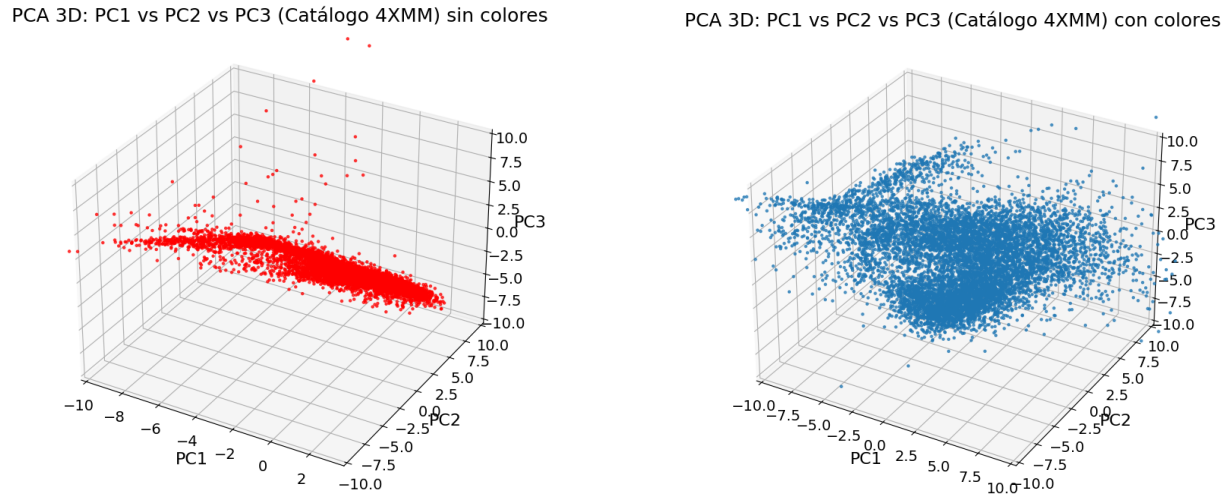
Tabla 2: Tabla de contingencia para los datos de MC para un `min_cluster_size=120`

Se puede observar que la separación obtenida no es muy efectiva. Por ello, en el siguiente apartado intentaremos comprobar la eficiencia con el grupo MX, para explorar si la inclusión de información en rayos-X mejora la separación entre los distintos tipos de fuentes.

3.2.2. Exploración magnitudes vs magnitudes y colores - Catálogo MX

Una vez obtenidos los resultados para la muestra completa, cabe preguntarse si añadir los flujos en X permite mejorar la delimitación de los *clusters* y la separación entre tipos. Como la MX cuenta con un número significativamente menor de datos, además nos permitirá comparar si los grupos de datos se separan mejor con o sin la adición de colores mediante. Por último, los datos de MX permitirán además usar el RF no supervisado para calcular la matriz de distancia y por tanto, obtener una métrica más adaptada al espacio de los datos.

Como resultados, en primer lugar como comparación en el espacio PC obtenemos las figuras 15b 15a

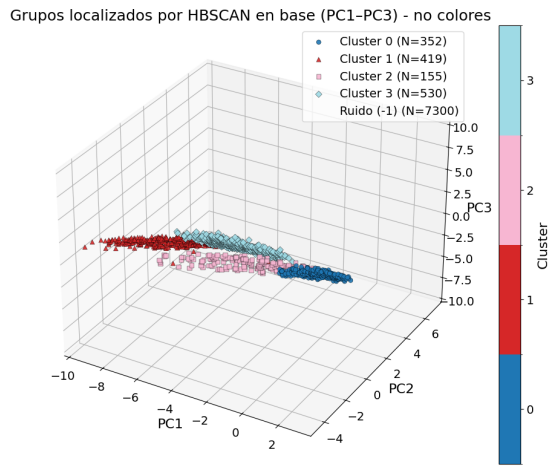


(a) Representación 3D de PCA si no se añade como características el color

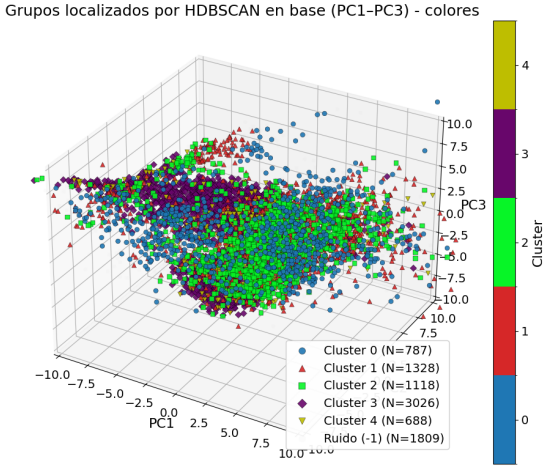
(b) Representación 3D de PCA si se añade como características el color

Figura 15: Inspección visual de diferencia entre añadir o no colores al cálculo de grupos. La PCA ha sido calculada para 7 PCs los cuales acumulan 0.995 y 0.940 de varianza acumulada respectivamente. Fuente: Propio

Desarrollando el flujo de trabajo, obtenemos los siguientes grupos de las figuras 16a y 16b.



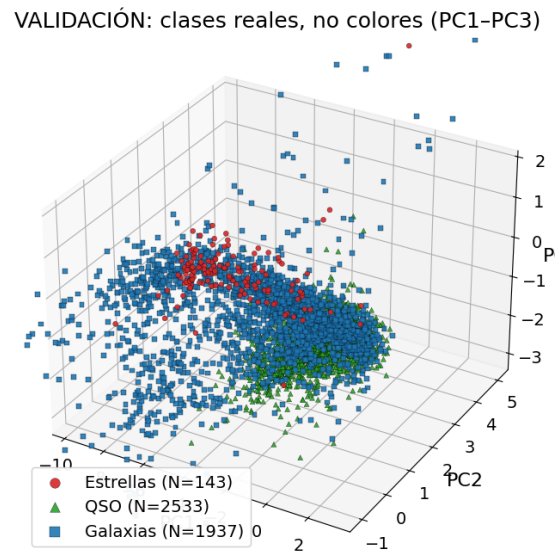
(a) Representación 3D de los grupos generados por HDBSCAN sin colores para un `min_cluster_size` = 120.



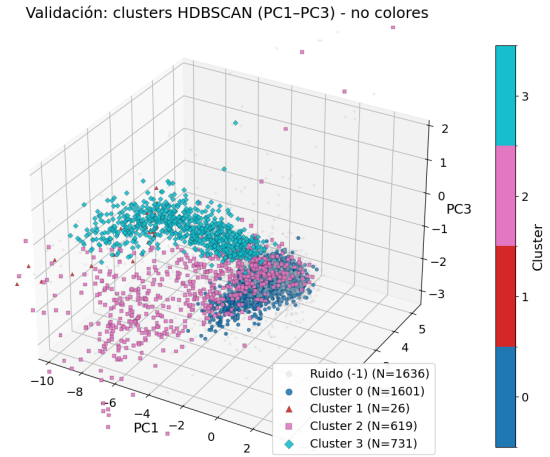
(b) Representación 3D de los grupos generados por HDBSCAN con colores para un `min_cluster_size` = 400.

Figura 16: Diferencia de los grupos generado por HBDSCAN para el catálogo 4XMM. El ruido detectado por HBSCAN ha sido ocultado para mejorar la claridad de los grupos. Fuente: propio

Si aplicamos la detección de grupos a los datos de validación en el espacio sin colores, obtenemos las figuras 17a y 17b.



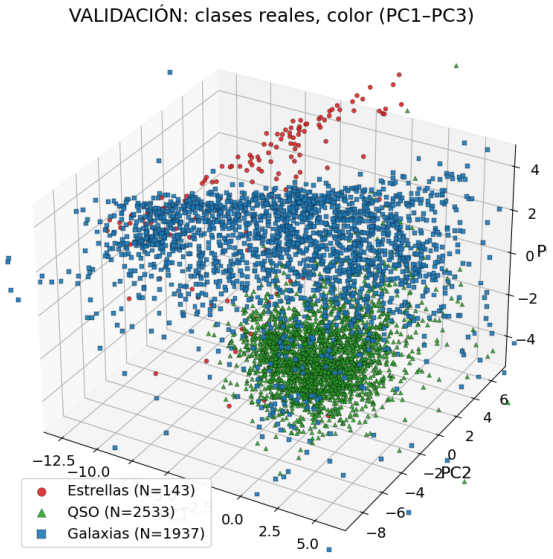
(a) Representación 3D de la clase de los datos de validación para una PCA sin colores.



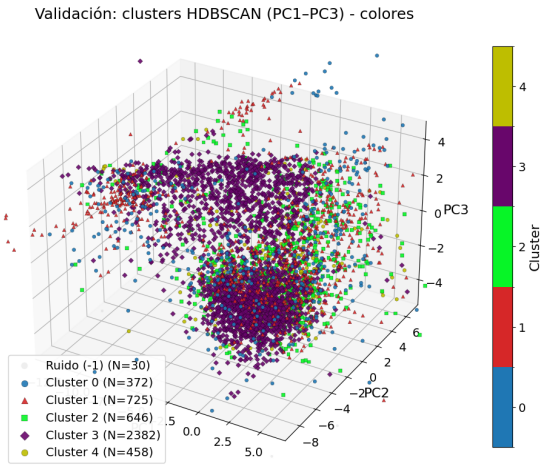
(b) Representación 3D de la predicción de grupos de HBSCAN para los datos de validación para una PCA sin colores.

Figura 17: Diferencia de los grupos generado por HBDSCAN para el catálogo 4XMM. El ruido detectado por HBSCAN ha sido ocultado para mejorar la claridad de los grupos. Fuente: propio

Hacemos lo mismo para el espacio con colores, obteniendo las figuras 18a y 18a.



(a) Representación 3D de la clase de los datos de validación para una PCA con colores.



(b) Representación 3D de la predicción de grupos de HBSCAN para los datos de validación para una PCA con colores.

Figura 18: Diferencia de los grupos generado por HBDSCAN para el catálogo 4XMM. El ruido detectado por HBSCAN ha sido ocultado para mejorar la claridad de los grupos. Fuente: propio

Donde la diferencia de grupos es notable. Al compararlo con el grupo de validación obtenemos, a modo de resumen, las tablas 3 y 4.

Clase	**Total**	**Ruido (-1)**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Estrellas	143	10 (6.99 %)	1 (0.70 %)	19 (13.29 %)	6 (4.20 %)	107 (74.83 %)
QSO	2533	881 (34.78 %)	1562 (61.67 %)	0 (0.00 %)	82 (3.24 %)	8 (0.32 %)
Galaxias	1937	745 (38.46 %)	38 (1.96 %)	7 (0.36 %)	531 (27.41 %)	616 (31.80 %)
TOTAL	**4613**	**1636 (35.46 %)**	**1601 (34.71 %)**	**26 (0.56 %)**	**619 (13.42 %)**	**731 (15.85 %)**

Tabla 3: Tabla de contingencia para el catálogo MX sin colores. Fuente: propio

Clase	**Total**	**Ruido (-1)**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Estrellas	143	3 (2.10 %)	40 (27.97 %)	58 (40.56 %)	24 (16.78 %)	14 (9.79 %)	4 (2.80 %)
QSO	2533	8 (0.32 %)	141 (5.57 %)	311 (12.28 %)	404 (15.95 %)	1336 (52.74 %)	333 (13.15 %)
Galaxias	1937	19 (0.98 %)	191 (9.86 %)	356 (18.38 %)	218 (11.25 %)	1032 (53.28 %)	121 (6.25 %)
TOTAL	**4613**	**30 (0.65 %)**	**372 (8.06 %)**	**725 (15.72 %)**	**646 (14.00 %)**	**2382 (51.64 %)**	**458 (9.93 %)**

Tabla 4: Tabla de contingencia para el catálogo MX con colores. Fuente: propio.

En el apartado 4, se discuten estos resultados en más detalle, pero cabe resaltar que añadir los colores constituye una mejora sustancial en la formación de grupos para HDBSCAN y su separación en la PCA.

3.2.3. Exploración de influencia de límites superiores MX+UL

Una vez explorado los datos de MX, es momento de preguntarse si añadir datos de menor calidad mejora sustancialmente los resultados. Sin embargo, el aumento de la población hace inviable el uso del URF al cálculo de matrices de distancia del orden de n^2.

En primer lugar, se obtendrán la representación en espacio PCA, la cual puede verse en la figura 19.

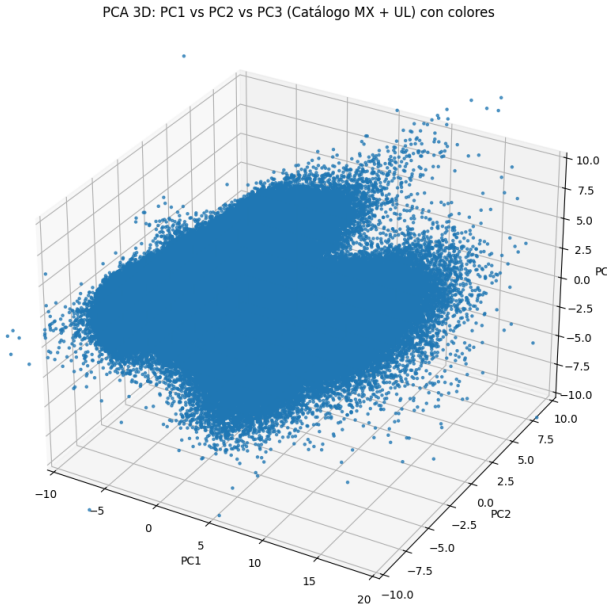


Figura 19: Representación en 3D de PCA para el catálogo MX+UL. Puede apreciarse como al añadir los límites superiores, limitan la visibilidad de los grupos de la figura 15b. Fuente: Propio.

Una vez visualizados los datos, introducimos la lista de magnitudes, colores y flujos en HDBSCAN tratando de identificar nuevos grupos. Como resultados se obtiene la figura 20.

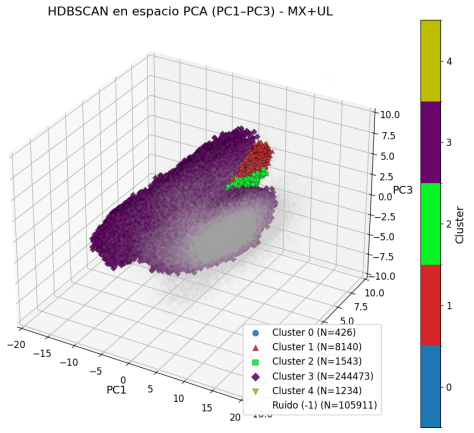
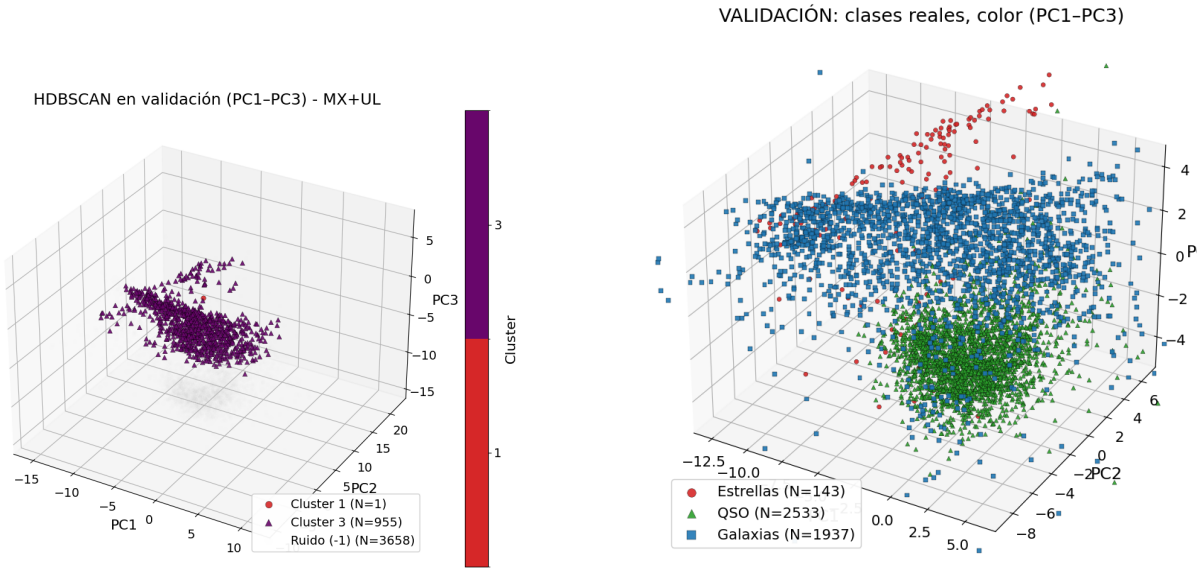


Figura 20: Representación delas predicciones de HDBSCAN para el catálogo MX+UL en espacio de PCA para un `min_cluster_size=120` Los grupos no son representativos de la estructura interna. Fuente: Propio.

Una vez obtenido los grupos sintéticos, se aplican a los datos de validación, obteniendo las figura 21a. La figura 21b es la representación de los datos marcados con los datos de validación.



(a) Representación 3D de los grupos sintéticos aplicados a los datos de validación en espacio PC.

(b) Representación de las clases reales pertenecientes a los datos de validación en el espacio PC.

Figura 21: Resultados obtenidos al comparar los grupos sintéticos en el espacio de datos de validación con las clases reales. Fuente: Propio.

Así pues, puede construirse la siguiente tabla de contingencia 5 para asociar los grupos sintéticos a los reales.

Clase	**Total**	**Ruido (-1)**	**Cluster 1**	**Cluster 3**
Estrellas	143	45 (31.47 %)	1 (0.70 %)	97 (67.83 %)
QSO	2533	2503 (98.82 %)	0 (0.00 %)	30 (1.18 %)
Galaxias	1938	1110 (57.28 %)	0 (0.00 %)	828 (42.72 %)
TOTAL	**4614**	**3658 (79.29 %)**	**1 (0.02 %)**	**955 (20.69 %)**

Tabla 5: Tabla de contingencia para el catálogo MX+UL

3.3. Propiedades de galaxia - Regresión

Este apartado permite aprovechar los catálogos desarrollados durante el apartado 3.1 con el objetivo de explorar métodos para completar datos faltantes a partir del resto de la muestra y para obtener características físicas de los objetos.

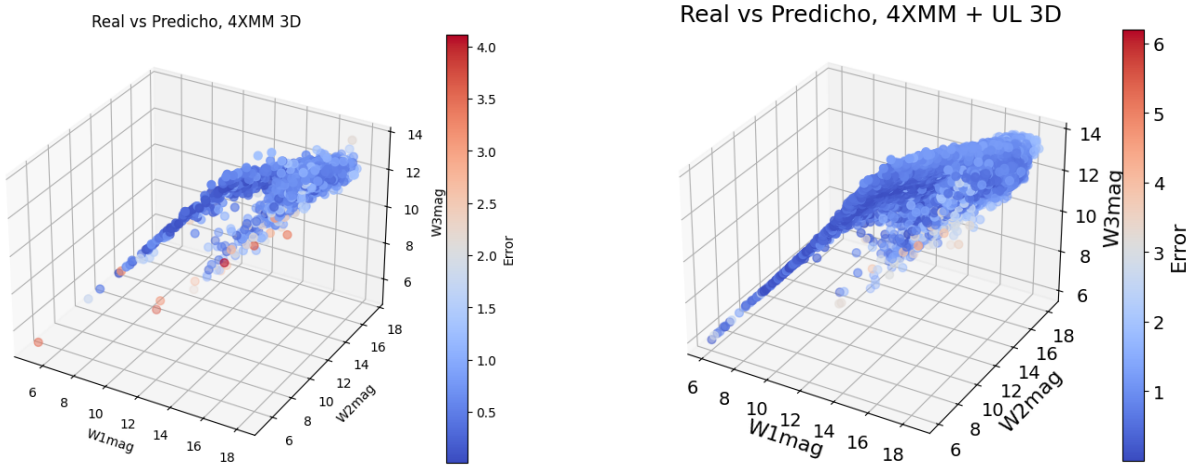
Por ello se dividirá en dos partes:

- Predecir W1, W2 y W3 (magnitudes de infrarrojo medio) a partir de las magnitudes del infrarrojo cercano, óptico y/o flujos en X
- Predecir el SFR, la masa estelar, la fracción de AGN, la luminosidad del AGN, y la luminosidad estelar a partir de magnitudes fotométricas, usando como comparación los obtenidos mediante ajuste "clásico" de SEDs, específicamente mediante `CIGALE`[26].

3.3.1. Predicción de magnitudes W1, W2 y W3

En primer lugar se exploró la estructura general de los datos de MX y MX+UL mediante la aplicación de RF. Este hecho es importante debido a los resultados obtenidos en el apartado anterior, donde se abría la posibilidad a que los datos sintéticos de límites superiores generasen dispersión y, por tanto, ruido a los resultados.

Así, se obtienen las imágenes 22a 22b.



(a) Generación de predicciones W1, W2, W3 con catálogo 4XMM

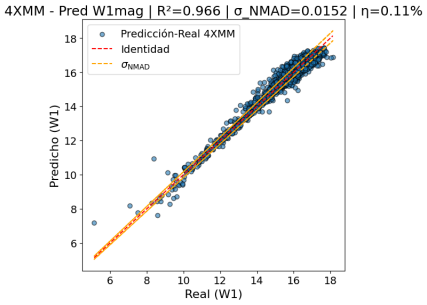
(b) Generación de predicciones W1,W2,W3 con catálogo 4XMM con datos sintéticos de límites superiores.

Figura 22: Predicciones W1, W2, W3 para catálogo con MX y MX + UL, los colores indican el error relativo con respecto al valor teórico. Fuente: Propio

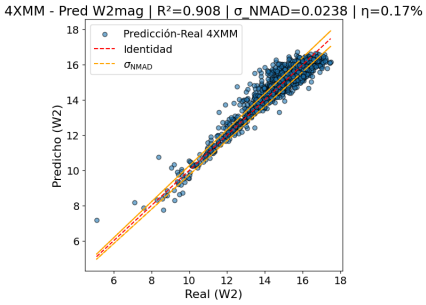
Debido a que las predicciones de RF entrelazadas también representan la correlación entre sus variables dependientes, se optó por desarrollar tres modelos por separado en el cálculo de W1, W2, W3. Así, se elimina la dependencia del modelo de optar con datos de las otras

[26] https://cigale.lam.fr/

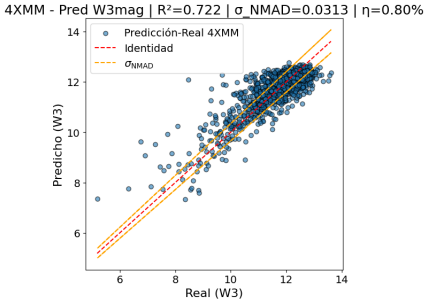
dos magnitudes a la hora de predecir la tercera. Para el catálogo MX obtenemos las figuras 23a 23b 23c.



(a) Generación W1 para catálogo 4XMM

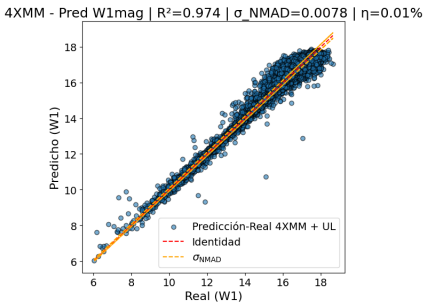


(b) Generación W2 para catálogo 4XMM

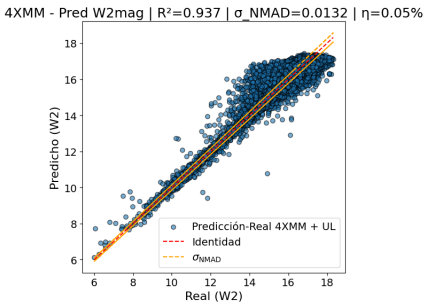


(c) Generación W3 para catálogo 4XMM

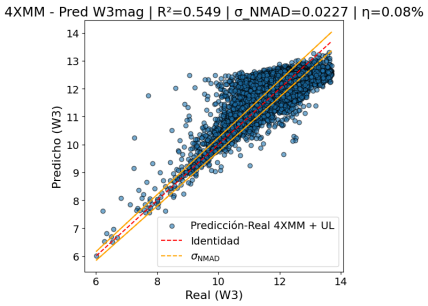
Figura 23: Representación de predicciones de W1, W2 y W3 respecto a real para el catálogo MX. Fuente: Propio



(a) Generación W1 para catálogo MX + UL



(b) Generación W2 para catálogo MX + UL



(c) Generación W3 para catálogo MX + UL

Figura 24: Representación de predicciones de W1, W2 y W3 respecto a real para el catálogo MX + UL. Fuente: Propio

Además de las figuras 24a 24b 24c para el catálogo con límites superiores.

En la tabla 6 se detalla la importancia, estimada mediante RF, de las variables principales utilizadas; y en la tabla 7 se resumen los parámetros estadísticos de cada modelo.

MX			**MX + UL**		
Rango	**Característica**	**Importancia**	**Rango**	**Característica**	**Importancia (MDI)**
1	Kmag	0.7290	1	Kmag	0.8647
2	Jmag	0.0689	2	Jmag-Kmag	0.0184
3	gmag	0.0396	3	gmag-rmag	0.0155
4	flux_8	0.0162	4	imag-zmag	0.0107
5	flux_4_5	0.0140	5	zmag	0.0073
6	gmag-rmag	0.0133	6	rmag-zmag	0.0069
7	Hmag	0.0106	7	zmag-Kmag	0.0069
8	Jmag-Kmag	0.0088	8	rmag-imag	0.0063
9	Hmag-Kmag	0.0081	9	gmag-imag	0.0055
10	zmag-Kmag	0.0079	10	Hmag	0.0050
11	imag-zmag	0.0067	11	Jmag	0.0036
12	gmag-imag	0.0066	12	flux_1_2_3	0.0035
13	zmag	0.0060	13	flux_4_5	0.0030
14	flux_1_2_3	0.0054	14	flux_8	0.0029
15	rmag	0.0049	15	zmag-Jmag	0.0027

Tabla 6: Top 15 características más importantes para el cálculo de W1, W2, W3 para cada catálogo. Fuente: Propio

	MX				MX+UL			
Magnitud	**RMSE**	$\mathbf{R^2}$	$\boldsymbol{\sigma}_{\mathrm{NMAD}}$	$\boldsymbol{\eta}$ (%)	**RMSE**	$\mathbf{R^2}$	$\boldsymbol{\sigma}_{\mathrm{NMAD}}$	$\boldsymbol{\eta}$ (%)
W1	0.3116	0.966	0.0152	0.11	0.2050	0.974	0.0078	0.01
W2	0.4650	0.908	0.0238	0.17	0.3136	0.937	0.0132	0.05
W3	0.4895	0.722	0.0313	0.80	0.3244	0.549	0.0227	0.08

Tabla 7: Parámetros estadísticos comparativos para cada magnitud en los modelos MX y MX+UL. Fuente: Propio

3.4. Modelización de SEDs

En esta sección se evalúa un método alternativo para la estimación de parámetros claves en galaxias y AGNs. Tradicionalmente, estas propiedades son obtenidas mediante el ajuste de las SEDs a varios modelos de emisión correspondientes a las diferentes componentes esperadas: emisión estelar, emisión del posible AGN y de las regiones de formación estelar, incluyendo además los efectos de la extinción. En especial, `CIGALE` es el código referencia para este hecho.

Nuestro objetivo es usar los resultados de la aplicación de `CIGALE` a las muestra de galaxias y AGNs descrita en la sección 2.2, procedente de [23], para construir muestras de entrenamiento y validación a las que poder aplicar algoritmos de ML supervisado.

Para ello se toman como features magnitudes fotométricas, que mediante un RF multivariable permiten obtener la tasa de formación estelar (SFR), la fracción de AGN, la luminosidad estelar y la luminosidad del AGN.

3.4.1. Datos y construcción del dataset

Los datos utilizados provienen de una **cross-correlación** previa al trabajo entre los catálogos `COSMOS` y `UltraVISTA`[23]. Estos están repartidos en dos subcatálogos, uno que agrupa a las galaxias, y otro que agrupa a las galaxias con detección en X propia de AGNs. El primer paso fue combinar ambos subcatálogos ya que estos representaban parte de un catálogo más grande. Así, de su unión se obtiene el catálogo `COSMOS_UltraVISTA_Merged`. En la figura 25 se representa la distribución del redshift normalizada para los catálogos `COSMOS_ULtraVISTA_gals` y `COSMOS_UltraVISTA_XrayAGN`.

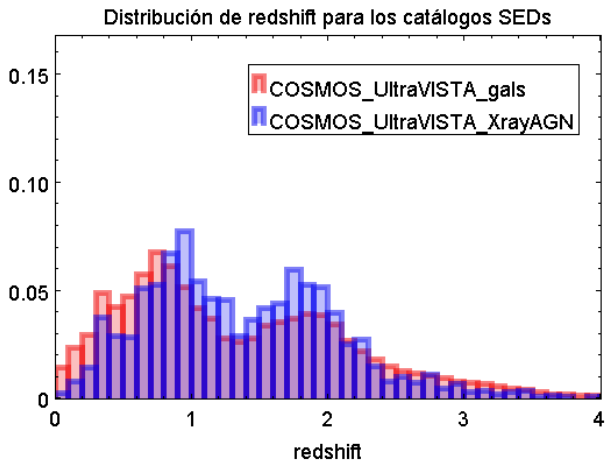


Figura 25: Distribución normalizada del redshift de los dos catálogos de galaxias usados para las SEDs. Fuente: Propio

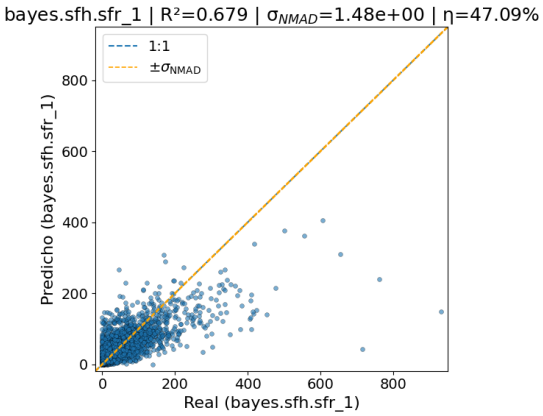
3.4.2. Preprocesamiento

Durante la limpieza de datos del catálogo surgieron dos problemas importantes: En primer lugar, las magnitudes fotométricas estaban divididas según el instrumento que la midió,

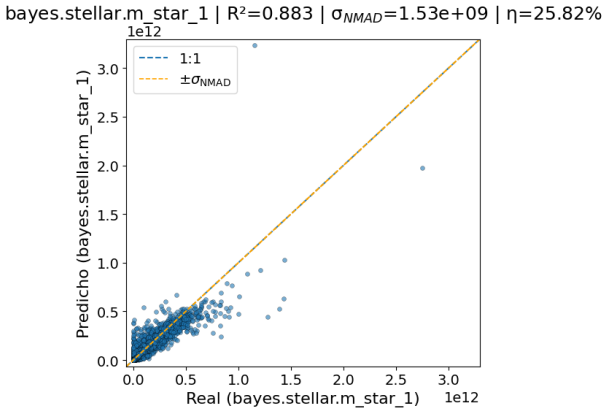
y por tanto varios instrumentos podían medir la misma magnitud; además como segundo problema, de carácter más técnico, en lugar de valores ”vacíos”/`NaN`, algunas magnitudes fotométricas marcaban la ausencia de medida con un valor “placeholder” de -9999. Estos dos incidentes consiguieron ser subsanados siguiendo los pasos detallados en el Apéndice A.

3.4.3. Aplicación del Random Forest

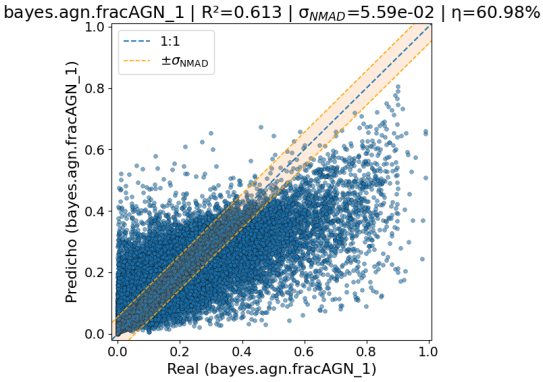
Una vez preparado el catálogo limpio, se aplicó el modelo de Random Forest para la predicción de los cinco parámetros de interés. Además de las predicciones, el algoritmo proporciona un **ranking de importancia de features**, que permite identificar los filtros fotométricos que contribuyen más significativamente a la estimación de cada parámetro. Los resultados se presentan en el conjunto de figuras 26 y el ranking en la figura 27.



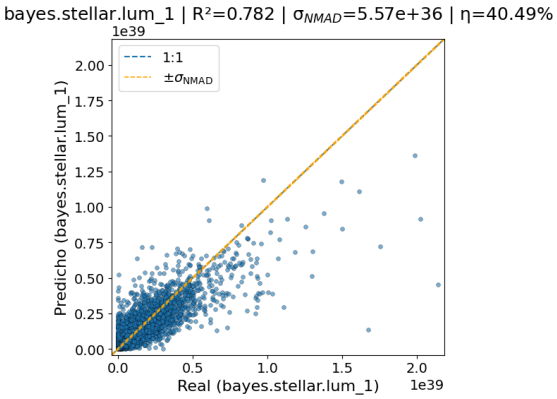
(a) Predicción de SFR para catálogo combinado de Cosmos_Ultravista



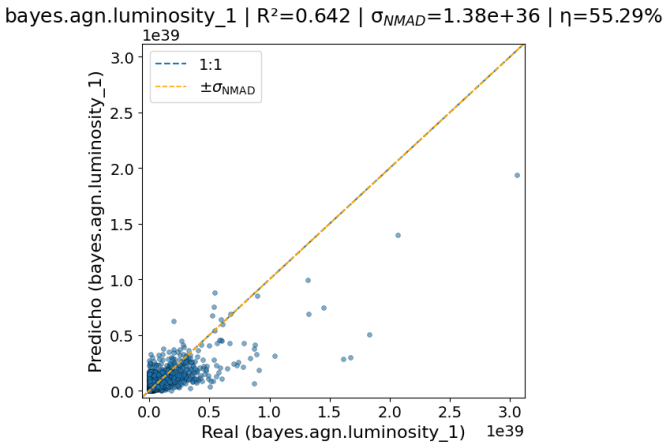
(b) Predicción de $M_\star$ para catálogo combinado de Cosmos_Ultravista



(c) Predicción de fracción AGN para catálogo combinado de Cosmos_Ultravista



(d) Predicción de luminosidad estelar para catálogo combinado de Cosmos_Ultravista



(e) Predicción de luminosidad AGN para catálogo combinado de Cosmos_Ultravista

Figura 26: Representación de predicciones para SFR, $M_\star$, fracción AGN, luminosidad estelar y luminosidad del AGN respecto a los valores reales en el catálogo 4XMM+UL. Fuente: propio.

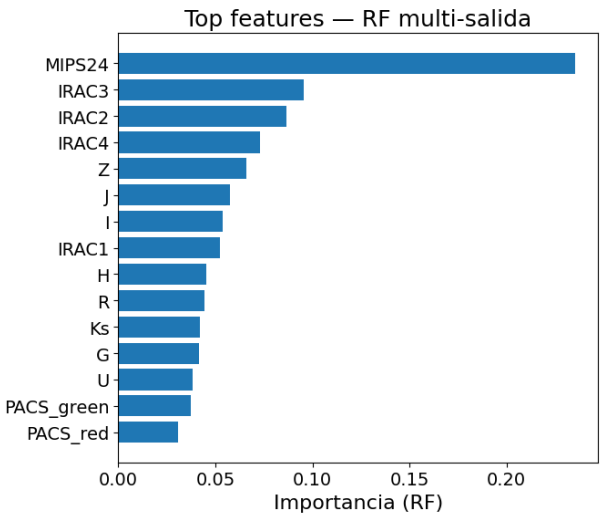


Figura 27: Ranking de importancia para las magnitudes en orden decreciente. Fuente: Propio

Finalmente los resultados están recogidos en la tabla 8 a modo de tabla de resumen de resultados.

Propiedad	**Target**	**RMSE**	$\mathbf{R^2}$	σ_{NMAD}	η (%)
SFR [$\mathrm{M_\odot\,yr^{-1}}$]	bayes.sfh.sfr_1	$1{,}37 \times 10^{1}$	0.679	$1{,}48 \times 10^{0}$	47.09
Masa estelar $M^\star$ [$\mathrm{M_\odot}$]	bayes.stellar.m_star_1	$2{,}34 \times 10^{10}$	0.883	$1{,}53 \times 10^{9}$	25.82
Fracción AGN	bayes.agn.fracAGN_1	$9{,}97 \times 10^{-2}$	0.613	$5{,}59 \times 10^{-2}$	60.98
Luminosidad estelar [W]	bayes.stellar.lum_1	$3{,}56 \times 10^{37}$	0.782	$5{,}57 \times 10^{36}$	40.49
Luminosidad AGN [W]	bayes.agn.luminosity_1	$2{,}49 \times 10^{37}$	0.642	$1{,}38 \times 10^{36}$	55.29

Tabla 8: Métricas de predicción obtenidas con Random Forest para los cinco *targets* seleccionados de `CIGALE`. Fuente: propio.

4. Discusión

En este apartado se analizan los resultados del apartado 3.

4.1. Muestra final

4.1.1. Datos de validación

Uno de los elementos más sensibles en el apartado de clustering es el número y la naturaleza de los datos de validación. El número de datos de validación depende de los catálogos con los que se han cruzado, ya que como máximo se podrán obtener un número de datos completos igual a la población del catálogo más pequeño.

Esto tiene una relevancia significativa tanto para el catálogo de entrenamiento como el de validación. Un catálogo menos poblado reduce significativamente el desempeño de los métodos de Machine Learning: gran parte de las estructuras que puedan existir pueden detectarse como outliers o incluso ser incluidas en estructuras distintas debido a no contar con la suficiente densidad. A cambio, añadir más catálogos en diferentes bandas permite una mayor profundidad, separando mejor las clases. Al final es una cuestión pura de compromiso.

Sin embargo, la naturaleza de los datos es fundamental para el estudio de los catálogos. Esto puede verse con el contenido de la población de VMX. Debido a que los mayores emisores de rayos-X son los QSO, al cruzar los catálogos fotométricos con 4XMM sesgan la población a favor de un mayor número de galaxias y sobre todo QSO. Esto se distingue con el contenido de la población: aun cuando VMC cuenta con unos 237000 datos (53 354 estrellas, 21 716 QSO y 162 626 galaxias) VMX cuenta con 4613 de los cuales son 143 estrellas, 2533 QSO y 1937 galaxias. Así pues, constituye una bajada en la población de estrellas desde el 22.44 % al 3.10 % sesgando los resultados según la clase.

Este sesgo abre la posibilidad de realizar estudios específicos a la clase favorecida, así pues, un estudio de AGNs estará muy favorecido por datos fotométricos con los que además se cuentan con flujos en rayos-X.

Esto pone un compromiso fundamental en astronomía, tal y como se demostrará en apartados posteriores: aumentar la profundidad y calidad de la muestra implica necesariamente reducir la variedad de esta misma.

4.2. Clustering

El apartado de clustering ha permitido estudiar a través de una reducción de variables las estructuras internas de los distintos catálogos disponibles para el estudio. Además, mediante el uso del mismo método de clustering para los tres catálogos ha permitido estudiar mejoras sustanciales en dos apartados claves: separación de clases astronómicas y la mejora en delimitación de grupos sintéticos.

4.2.1. Exploración catálogo MC

Este apartado ha consistido en la exploración de un catálogo puramente fotométrico, sin contar con los flujos en rayos X. Los datos han sido adaptados al espacio PCA, quedando limitado en una forma de dos lóbulos principales. Sin embargo, el gran número de datos, y, sobre todo la profundidad, ha impedido observar claramente estructuras internas más relevantes (Figura 12)

Al acudir a los datos de validación, parte de la estructura interna parece aflorar. Uno de los lóbulos, el diagonal, cuenta con similitud con el lóbulo marcado por las estrellas mientras el segundo lóbulo, parece ser una combinación tanto de cuásares como de galaxias.

El uso de HDBSCAN, junto a una métrica euclidiana en el espacio (la cual debería ser favorecida por la linealidad que aplica la PCA), no ha conseguido separar adecuadamente los datos, generando un grupo central el cual abarca la mayoría de los datos y varios subgrupos de poca relevancia (Figuras 13, 21a). Esta disposición coincide con los resultados de [7], donde sus grupos sintéticos generados por GMM se repartían de una forma muy similar.

Los resultados, sin embargo, no son inesperados: la selección de magnitudes para MC solo cuenta con datos en el espectro visible e infrarrojo, bandas espectrales en las que las clases espectroscópicas que proporcionan la validación no son tan relevantes. Así pues, puede existir una distinción asignando cada lóbulo a las clases estrellas (diagonal) y galaxias/QSO.

Así, es lógico preguntarse si es posible mejorar esta separación entre las clases añadiendo nuevas características por medio de las bandas. Una propuesta podría ser el redshift que permite delimitar una clara separación entre estrellas y galaxias/quasares, mientras que la adición del flujo en X debería mostrar una mejor separación entre estrellas y, sobre todo galaxias y QSO, como se analiza en el siguiente apartado.

Debido a que uno de los objetivos es distinguir grupos de objetos astronómicos únicamente con datos fotométricos, descartando el redshift el cual tiene un origen espectroscópico, la implementación del catálogo de 4XMM queda completamente justificada.

4.2.2. Exploración magnitudes vs magnitudes y colores - Catálogo MX

Las diferencias entre los datos de los catálogos MC y MX son evidentes: el catálogo MC cuenta con 20 veces más datos a cambio de una mayor profundidad del catálogo MX. Esta diferencia radica en la disminución sustancial de los tiempos de computación permitiendo la oportunidad de aplicar URF para el cálculo de la matriz de distancias.

Así se obtienen las imágenes 15a y 15b. Se ha decidido hacer el estudio de colores con el catálogo MX debido a la claridad que muestran los sondeos iniciales. El aumento de la profundidad permiten observar estructuras internas más diferenciadas que además son favorecidas por la adición de colores fotométricos.

Si se comparan con los datos de MC, los dos lóbulos parecen separarse de mejor manera, mostrando las poblaciones de estrellas, QSO y galaxias aún más separadas. Al comparar

con los datos de validación, sobre todo en el grupo con colores, los dos lóbulos son divididos en tres grupos. Uno poco poblado que es el de las estrellas (debido principalmente al sesgo del cruce con 4XMM), un grupo de galaxias y un cúmulo esférico de cuásares.

Las tablas de contingencia 3,4, revela un hecho sorprendente. El grupo que no cuenta con colores agrupa más de la mitad de los cuásares en el cluster 0, un clúster 1 constituido casi únicamente con estrellas, y un clúster 2 y 3 mixtos. Es importante resaltar que la disminución sustancial de los datos de validación puede afectar a estos porcentajes ya que existe un sesgo que beneficia sobre todo a los QSO. En cuanto a la tabla con colores, en primer lugar hay que indicar que ha sido necesario subir el parámetro `min_cluster_size=300` indicando que la adición de colores provoca una dispersión mayor entre los datos. Además, gran parte de los datos detectados por el ruido parecen ser obviados por HDBSCAN reduciendo su población en un 35 %. Sin embargo, las tendencias parecen retroceder a los resultados obtenidos por MX, un clúster central que acumula gran parte de los datos (en este caso clúster 3) y grupos satélites con poca aportación con las clases espectroscópicas.

Así, el siguiente paso es tomar un catálogo más poblado con la misma profundidad. Por ello se optó por la técnica de límites superiores.

4.2.3. Exploración de influencia de límites superiores MX+UL

La adición de límites superiores ha permitido aumentar la población del catálogo en X a los mismos números que el catálogo MC. En un primer momento se apostó por `FLIX` para este hecho, sin embargo, los tiempos de espera de cada límite superior la hizo una herramienta inviable para el número de datos a completar.

Así es como se llegó a la segunda solución, la cual es una aproximación menos realista, pero que sirve bien para demostrar la influencia en los métodos con catálogos con límites superiores. Así se obtuvieron la figura 19. Esta imagen permite ver cómo este catálogo recupera la forma de dos lóbulos centrales propia del catálogo MC adquiriendo, en consecuencia, los problemas derivados.

Al aplicar HDBSCAN, en la figura 20, los grupos divididos parecen formar sublóbulos de tamaños no despreciables. Sin embargo, en la tabla 5, algunos de los clusters sintéticos no aparecen, debido principalmente a la diferencia de tamaño poblacional entre datos de entrenamiento y validación. Este hecho abre la posibilidad de estudiar en el futuro cómo afectan el cálculo de límites superiores en la validación, pero por límite de tiempo esta opción no ha podido ser explorada.

4.2.4. Métodos de clustering: Aplicación de resultados.

Aunque los resultados en los métodos de clustering no han resultado concluyentes, estos abren la puerta a ser implementados junto a técnicas híbridas, donde la identificación de grupos basados en similitudes entre los datos permitan combinarse con técnicas de ML supervisado mejorando sus clasificaciones o predicciones. Estas técnicas híbridas ya han sido probadas obteniendo buenos resultados [24]

4.3. Propiedades de galaxia - Regresión

4.3.1. Predicción de magnitudes W1, W2 y W3

Tal y como se indicó en los objetivos de este trabajo, la predicción de las magnitudes que conforman la banda infrarroja media, permite afinar una herramienta capaz de completar la falta de estas magnitudes. Además, al compartir la misma estructura que las construcciones de modelos con los datos de SEDs permite desarrollar una familiaridad con este proceso, resultando más sencillo afinar los métodos o comprender sus resultados.

Así, se desarrollaron tres modelos basados en RF supervisado para cada una de las magnitudes W1, W2 y W3, obteniendo resultados, tanto para el catálogo de MX como el de MX+UL. Estos modelos han sido representados por un diagrama predicción-real. Este permite asociar la calidad del modelo a la búsqueda de su relación lineal 1:1. Así, métricas como el R^2, el cual mide la linealidad de los datos, toman una importante relevancia, ya que cuanto más cercano esté este valor a 1, más lineal será esta relación. Además, la adición del RMSE, permite medir la dispersión de los datos, reflejando si el modelo cuenta con una mayor varianza.

Adicionalmente, se han calculado dos parámetros claves: σ_{NMAD} el cual representa la desviación media normalizada, la cual permite concretar un valor de varianza resistente a datos anómalos. Además, se introduce el valor η que indica el porcentaje respecto al total de datos anómalos.

Como consecuencia, la tabla 7 indica que los modelos para W1 y W2 mantienen una relación lineal, no tanto como W3 la cual parece necesitar en su modelo más características. La adición de límites superiores ha reducido la fracción de datos anómalos, añadiendo una relación más lineal que empeora para W3.

También, gracias a los indicadores de `feature_importance`, permite hacer un ranking de importancia de características en cada modelo (Figura 6). Así se demuestra que la característica más importante para ambos modelos es la magnitud K, la cual es la magnitud más cercana al infrarrojo medio. A continuación se presentan la magnitud J y g. Sin embargo, las magnitudes propias parecen perder importancia al añadir los límites superiores, haciendo que algunos colores en estas bandas dominen más que incluso las magnitudes.

4.4. Modelización de SEDs

La modelización de SEDs por medio de RF supervisado ha seguido la misma estructura que en el apartado anterior. Sin embargo, debido a la necesidad del uso de `CIGALE` para validar las predicciones, se ha optado por tomar catálogos usados en otros trabajos.

Así, se tomaron los mismos datos que en el trabajo de [23], donde se realiza el análisis con `CIGALE` para un cross-match de COSMOS-ULTRAVISTA. Contando con sus resultados, se han incorporado como predictores cada una de las magnitudes explicadas en el apartado de Metodología, obteniendo las figuras 26. Entre los resultados, el mejor modelo es el de masa estelar, el cual obtiene una mejor linealidad y una menor población de datos anómalos. A

continuación mejora la luminosidad estelar (la cual está relacionada con la masa debido al Mass to Light ratio), y después el SFR, la fracción de AGN y la fracción de luminosidad del AGN.

Los resultados, sin embargo, no son del todo satisfactorios. Al usarse un modelo conjunto de los cinco objetivos se han obtenido resultados menos lineales que en la predicción de magnitudes en el infrarrojo medio. Sin embargo, los buenos resultados en la función de masa y luminosidad estelar abre la puerta a mejora con las características que podrían delimitar mejor los demás targets.

Además, se realizó el mismo proceso para cada catálogo por separado, obteniendo la tabla 9. La mejora en la unión de catálogos es evidente, en general el catálogo de AGNs es muy inferior en población al de galaxias. Este hecho hace que las métricas de galaxia dominen en los resultados. Sin embargo, añadir el de AGN mejora consistentemente las métricas de Fracción de AGN y Luminosidad AGN.

Propiedad	**Target**	**AGN+Galaxias**			**AGN**			**Galaxias**		
		RMSE	R^2	η (%)	RMSE	R^2	η (%)	RMSE	R^2	η (%)
SFR [$M_\odot\,yr^{-1}$]	bayes.sfh.sfr_1	$1{,}37 \times 10^{1}$	0.679	47.09	$4{,}49 \times 10^{1}$	0.115	75.33	$1{,}37 \times 10^{1}$	0.650	46.50
Masa estelar $M^\star$ [$M_\odot$]	bayes.stellar.m_star_1	$2{,}34 \times 10^{10}$	0.883	25.82	$1{,}86 \times 10^{11}$	0.326	65.67	$2{,}06 \times 10^{10}$	0.896	26.59
Fracción AGN	bayes.agn.fracAGN_1	$9{,}97 \times 10^{-2}$	0.613	60.98	$1{,}82 \times 10^{-1}$	0.279	78.00	$9{,}91 \times 10^{-2}$	0.606	60.97
Luminosidad estelar [W]	bayes.stellar.lum_1	$3{,}56 \times 10^{37}$	0.782	40.49	$1{,}68 \times 10^{38}$	0.204	76.00	$3{,}60 \times 10^{37}$	0.752	40.39
Luminosidad AGN [W]	bayes.agn.luminosity_1	$2{,}49 \times 10^{37}$	0.642	55.29	$1{,}84 \times 10^{38}$	0.200	66.67	$2{,}66 \times 10^{37}$	0.513	55.50

Tabla 9: Resumen de métricas de predicción para los cinco *targets* seleccionados de `CIGALE`, evaluados en los tres *datasets*: AGN+Galaxias, AGN y Galaxias. Fuente: propio.

5. Conclusiones

Una vez terminada la discusión, a modo de resumen se expondrá las conclusiones de cada uno de los objetivos.

- Gracias al estudio de los catálogos con diferentes bandas, se ha demostrado que un aumento de la población de catálogos sin una mejora en la profundidad, impide que los grupos sean fácilmente separables en los métodos de clustering.
- La adición de flujos en X ha mejorado sustancialmente la separación en los datos de validación. Aun cuando el método de HDBSCAN parece mejorar con esto, los resultados no pueden ser declarado como suficientes debido al sesgo interno obtenido del cruce de los datos con 4XMM resultando en una bajada sustancial de la población de los datos de validación.
- Una solución al problema del tamaño poblacional ha sido añadir nuevos datos en X por medio de la técnica de límites superiores. Este hecho sin embargo, ha resultado contraproducente, la adición de límites superiores han añadido ruido a la representación dificultando la mejora que añadía los datos en X.
- La predicción de magnitudes de infrarrojo medio está dominada por el valor de la magnitud K del infrarrojo cercano. La adición de límites superiores en rayos-X mejoran las predicciones, aun cuando W3 parece exhibir un comportamiento menos lineal.
- Los métodos de predicción basado en los resultados de `CIGALE` han permitido una aproximación lineal y más rápida del cálculo tradicional de características propias de galaxias y AGNs. La adición de métodos más refinados como redes neuronales pueden permitir una mejora de los modelos que pueden coexistir con los ya tradicionales.

5.1. Trabajo futuro

Teniendo en cuenta que cada uno de los apartados recorre campos del ML muy diferentes entre sí, no se ha podido realizar un estudio exhaustivo de cada uno de estos objetivos. Sin embargo, este trabajo ha permitido realizar un estudio global relacionando campos de ML en principio separados.

En los métodos de clustering, la mejora de la adición de flujos en rayos-X es evidente. Añadir características como el redshift podría mejorar aún más la separación de los grupos, además del uso de métodos de clustering combinados con algoritmos de clasificación.

En predicción, los modelos obtenidos permiten demostrar una buena relación entre modelo-real, aún cuando esta puede mejorarse con técnicas que, además, pueden ser más interpretativas que RF. Un aumento del número y la calidad de los datos además deben refinar el modelo, disminuyendo la población de datos anómalos. Parece que la técnica de límites superiores puede ser refinada aún más, con lo que aplicando una estimación más rigurosa, debería presentar resultados más refinados.

En las predicciones de SEDs, por falta de tiempo solo ha podido estudiarse un único catálogo. Asociando nuevos catálogos donde se ha usado `CIGALE`, además de incrementar significativamente el tamaño de las muestras, puede abrir una puerta a la comparación de resultados según el catálogo.

Finalmente, en un futuro próximo se prevé preparar para publicar en `GitHub` el código desarrollado a lo largo de este trabajo, con el objetivo de facilitar su consulta y posible reutilización.

Referencias

[1] Guangping Li, Zujia Lu, Junzhi Wang, and Zhao Wang. Machine learning in stellar astronomy: Progress up to 2024. arXiv preprint arXiv:2502.15300, 2025. (Citado en la página 1).

[2] Joshua N. Winn, Matthew J. Holman, Guillermo Torres, Peter McCullough, Christopher Johns-Krull, David W. Latham, Avi Shporer, Tsevi Mazeh, Enrique Garcia-Melendo, Cindy Foote, Gil Esquerdo, and Mark Everett. The transit light curve project. ix. evidence for a smaller radius of the exoplanet xo-3b. The Astrophysical Journal, 683(2):1076, aug 2008. (Citado en la página 1).

[3] C. H. A. Logan and S. Fotopoulou. Unsupervised star, galaxy, QSO classification. Application of HDBSCAN. , 633:A154, January 2020. (Citado en la página 1).

[4] Rich Ormiston, Tri Nguyen, Michael Coughlin, Rana X. Adhikari, and Erik Katsavounidis. Noise reduction in gravitational-wave data via deep learning. Physical Review Research, 2(3):033066, July 2020. (Citado en la página 1).

[5] Juan Fabregat. Apuntes de astrofísica observacional (versión 3). Apuntes, s.f. Recuperado de https://www.uv.es/fabregaj/apuntes/aoi_v3.pdf. (Citado en la página 2).

[6] V.A. Masoura. The coevolution of the AGNs and their host galaxies. PhD thesis, University of Athens, 09 2022. (Citado en la página 4).

[7] Pablo Gómez Nicolás. Classification of astronomical sources through machine learning techniques, 2019. Trabajo de Fin de Grado en Física, Universidad de Cantabria. (Citado en las páginas 6, 12, and 40).

[8] F. X. Pineau, S. Derriere, C. Motch, F. J. Carrera, F. Genova, L. Michel, B. Mingo, A. Mints, A. Nebot Gómez-Morán, S. R. Rosen, and A. Ruiz Camuñas. Probabilistic multi-catalogue positional cross-match. , 597:A89, January 2017. (Citado en la página 6).

[9] G. Mountrichas, V. A. Masoura, A. Corral, and F. J. Carrera. Comparative analysis of the SFR of AGN and non-AGN galaxies, as a function of stellar mass, AGN power, cosmic time, and obscuration. , 683:A143, March 2024. (Citado en la página 6).

[10] V. Khramtsov, C. Spiniello, A. Agnello, and A. Sergeyev. Vexas: Vista extension to auxiliary surveys. data release 2: Machine-learning based classification of sources in the southern hemisphere. , 651:A69, July 2021. (Citado en la página 7).

[11] Fermilab. The dark energy survey: Survey and operations. https://www.darkenergysurvey.org/the-des-project/survey-and-operations/. (Citado en la página 8).

[12] Astro Data Lab – NOIRLab. Vista hemisphere survey (vhs). Astro Data Lab, Community Science and Data Center (CSDC), NSF NOIRLab, 2020. (Citado en la página 8).

[13] Dalya Baron and Dovi Poznanski. The weirdest sdss galaxies: results from an outlier detection algorithm. Monthly Notices of the Royal Astronomical Society, 2016. (Citado en la página 14).

[14] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, Advances in Knowledge Discovery and Data Mining, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. (Citado en la página 14).

[15] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans. Knowl. Discov. Data, 10(1), July 2015. (Citado en la página 14).

[16] Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selection. In 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), page 223228. IEEE, September 2020. (Citado en la página 14).

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In D. W. Pfitzner and J. K. Salmon, editors, Second International Conference on Knowledge Discovery and Data Mining (KDD'96). Proceedings of a conference held August 2-4, pages 226–331, January 1996. (Citado en la página 14).

[18] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. CM Transactions on Database Systems (TODS), pages 42(3), 19, 2017. (Citado en la página 14).

[19] A. Lojas and otros. A machine learning algorithm for noisy datasets. Astronomy Astrophysics, 633:A154, 2020. (Citado en las páginas 14 and 15).

[20] Joaquín Amat Rodrigo. Random forest con python, 2020. Última actualización: agosto 2025. (Citado en las páginas 15 and 16).

[21] XMM-Newton Survey Science Centre at IRAP. Flix: Sensitivity estimator for xmm-newton data. http://flix.irap.omp.eu, 2025. Last accessed on 12 de septiembre de 2025. (Citado en la página 20).

[22] C. Spiniello and A. Agnello. VEXAS: VISTA EXtension to Auxiliary Surveys. Data Release 1. The southern Galactic hemisphere. , 630:A146, October 2019. (Citado en la página 21).

[23] G. Mountrichas and I. Georgantopoulos. The properties of supermassive black holes and their host galaxies for type 1 and 2 active galactic nuclei in the eFEDS and COSMOS fields. , 683:A160, March 2024. (Citado en las páginas 35 and 42).

[24] Rishabh Soni and K. Mathai. An Innovative Cluster-then-Predict Approach for Improved Sentiment Prediction, pages 131–140. Springer, 01 2016. (Citado en la página 41).

A. Limpieza de dataset COSMOS-VISTA

Para la limpieza del catálogo se siguieron los siguientes pasos.

1. Se sustituyeron por `NaN` todos los valores anómalos, definidos como aquellos que comprenden el intervalo $-100 < m < 100$. De este modo se eliminan tanto los placeholders como valores de magnitud físicamente inconsistentes.
2. Para cada filtro fotométrico observado por varias cámaras, se aplicó un criterio jerárquico:
   - Si todas las cámaras tenían valores distintos de `NaN`, se seleccionó la cámara con menor número de `NaNs` entre las candidatas.
   - Si coexistían cámaras con y sin valores, se priorizó aquella con menor porcentaje de `NaNs` .
   - Si todas las cámaras carecían de datos, la magnitud correspondiente se marcó como `NaN`.
3. Finalmente, se identificaron los filtros con mayor proporción de `NaNs`, que fueron descartados del análisis. La Tabla 10 recoge el porcentaje de valores ausentes por filtro, ordenados a lo largo del espectro.

Columna	**NaNs**	**Columna**	**NaNs**
Features		*Features*	
SPIRE250	91.39 %	R	0.85 %
SPIRE350	91.36 %	I	0.70 %
SPIRE500	91.35 %	Z	0.62 %
Ks	20.64 %	PACS_red	0.33 %
H	20.43 %	MIPS24	0.29 %
J	20.31 %	PACS_green	0.27 %
U	9.12 %	IRAC4	0.16 %
G	3.04 %	IRAC2	0.00 %
IRAC3	1.84 %	IRAC1	0.00 %
Targets			
bayes.sfh.sfr_1	0.06 %	bayes.stellar.m_star_1	0.06 %
bayes.agn.fracAGN_1	0.06 %	bayes.stellar.lum_1	0.06 %
bayes.agn.luminosity_1	0.06 %		

Tabla 10: Porcentaje de datos vacíos por columna en `Merged_catalog_final.csv`. El número total de filas es $N = 296{,}893$. En rojo, las columnas que se han optado por eliminarse para aumentar la población de los datos. Fuente: propio

Este procedimiento permitió optimizar a la vez el tamaño de la muestra final y la cobertura multi-longitud de onda, identificando aquellos filtros que solo están disponibles para un número muy limitado de fuentes. Así pues, se descartaron las magnitudes `SPIRE` cuya baja cobertura reducía significativamente el tamaño del catálogo.