INSTITUTO TECNOLÓGICO DE CD. GUZMÁN

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

TESIS

TEMA:

TÉCNICAS DE APRENDIZAJE AUTOMATIZADO PARA LA BÚSQUEDA DE ESTRELLAS SIMBIÓTICAS EN LA MISIÓN GAIA (DR2)

QUE PARA OBTENER EL TÍTULO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

ING. ROGELIO HERNÁNDEZ MONTES

ASESOR(A):

DRA. KARLA LILIANA PUGA NATAL

DRA. SILVANA GUADALUPE NAVARRO JIMÉNEZ

CD. GUZMÁN JALISCO, MÉXICO, AGOSTO DE 2019

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Ciudad Guzmán

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cd. Guzmán, Jal. a 15/Agosto/2019

Oficio No. DEPI/50/19

ASUNTO : AUTORIZACIÓN DE IMPRESIÓN

C. ROGELIO HERNÁNDEZ MONTES
N.C. M17290008

En cumplimiento con el documento normativo de las disposiciones para la operación de estudios de posgrado del Tecnológico Nacional de México y con base en la aprobación del Comité Tutorial comisionado para su revisión; la División de Estudios de Posgrado e Investigación le otorga la autorización de impresión de su trabajo de tesis intitulado:

"Técnicas de aprendizaje automatizado para la búsqueda de estrellas simbióticas en la misión GAIA (DR2)"

dirigido por el **Dra. Karla Liliana Puga Nathal**, desarrollado como requisito parcial para la obtención del grado de Maestro en Ciencias de la Computación, de acuerdo al plan de estudios MCCOM-2011-05.

Sin otro asunto en particular, quedo de usted.

ATENTAMENTE

DR. HUMBERTO BRACAMONTES DEL TORO
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

C.p. Archivo

Índice

Índice de ecuaciones

Índice de figuras

Índice de tablas

Capítulo 1
Introducción

El universo siempre ha maravillado a la humanidad. La curiosidad ha llevado al hombre al desarrollo de diversos instrumentos y técnicas para el estudio y comprensión de los cuerpos celestes. Los telescopios, instrumentos diseñados para esta función, han sido los principales generadores de información, misma que necesita ser procesada y analizada para obtener datos de utilidad en el campo de la astrofísica. Además existen objetos que por sus características inherentes pueden llegar a ser difíciles de clasificar, ya sea por su peculiaridad, o por su gran similitud con otros objetos, lo que causa confusión.

Respondiendo a esta realidad se presenta este proyecto de tesis, que aborda el problema con una solución computacional, que consiste en realizar la búsqueda de estrellas simbióticas a través de algoritmos de aprendizaje automatizado. Las bases de datos utilizada pertenecen a la misión espacial GAIA de la Agencia Espacial Europea (ESA por sus siglas en inglés) y específicamente la última entrega de datos denominada DR2, que fue liberada el 25 de abril de 2018 (ESA,2019).

Las técnicas que se utilizaron fueron los algoritmos de redes neuronales (ANN), Random Forest (RF), árboles de decisión (DT) y máquinas de soporte vectorial (SVM). Los resultados obtenidos en la presente investigación son alentadores para la realización de la búsqueda y clasificación de las estrellas binarias, ya que RF presentó un índice kappa de 98% y el peor resultado fue DT con 90%.

En el capítulo uno se presenta el planteamiento del problema, se aborda la necesidad de realizar una identificación correcta de estrellas simbióticas a través de técnicas de inteligencia artificial. También se plasma el objetivo de la tesis y la hipótesis que se plantea demostrar a través de este proyecto.

El capítulo dos contiene las bases teóricas sobre las cuales se sustenta el presente trabajo de tesis. Las bases teóricas han sido separadas en tres conjuntos con base al área que corresponden como criterio de separación. Los conjuntos consisten en Astrofísica, en Software y Big Data. Además de las bases teóricas se incluye en este capítulo investigaciones las cuales se encuentran relacionadas con el trabajo realizado.

El capítulo tres presenta la metodología utilizada para el desarrollo de la tesis. Información sobre la investigación, como el tipo de investigación, población y muestra usadas. Igualmente en este capítulo se encuentra la descripción y configuración de las técnicas de aprendizaje automático utilizadas.

Así mismo, en el capítulo cuatro se plasman los resultados obtenidos a través de las técnicas de aprendizaje supervisado seleccionadas. Además dentro de los resultados se anexan las esquematización de los modelos que se generaron.

El capítulo cinco incluye la conclusión final sobre la investigación realizada. Además se habla del trabajo futuro que se puede desglosar del presente trabajo de tesis.

Se anexa un glosario con los términos que se consideran necesarios aclarar su significado.

Las fuentes consultadas se encuentran al final del documento, presentadas con el formato APA 6.

1.1 Planteamiento del problema

Según Liliana Hernández en una publicación de la revista digital universitaria de la UNAM se estima que, desde hace más de diez años, la tasa de recolección de datos astronómicos es de un Terabyte por día (Hernández C., et al, 2009). Toda la información generada es almacenada en bases de datos localizadas en distintas partes del mundo. Estas bases de datos en su mayoría, están abiertas al público. Analizar toda esta información no es factible sin la ayuda de software especializado. Aun así, los astrónomos o dicho software puede cometer errores al clasificar incorrectamente las estrellas simbióticas (Symbiotics Stars “SS”), ya sea por su variabilidad tanto fotométrica como espectroscópica o porque son sistemas binarios.

El problema aquí planteado es la necesidad de identificar correctamente las estrellas simbióticas, a través de una búsqueda automática. El reto es que se necesitará el desarrollo de un software que se especialice en esto. Para ello se utilizó una base de datos abierta, con información cruda o pre-procesada. La propuesta para resolver la problemática se puede resumir en realizar la búsqueda, el análisis y la identificación de estrellas simbióticas por medio de un software que trabaje sobre bases de datos astronómicas, utilizando técnicas de Analítica de Datos.

La misión espacial GAIA, liberó en abril de 2018, una base de datos fotométrica, que contiene información de millones de estrellas. Información tan importante como su posición con gran precisión, su magnitud en varios filtros, su *paralaje*, etc. Datos que permiten determinar características físicas de estas estrellas, como su distancia o temperatura. Esta base de datos denominada (DR2), está disponible para la comunidad científica y la información plasmada en ella, puede permitir encontrar estrellas simbióticas no antes detectadas, por lo tanto se utilizarán algunos algoritmos de analítica de datos que permitan la

detección de estrellas simbióticas y con ello presentar a la comunidad científica una lista de estrellas candidatas a esta clasificación.

1.2 Objetivos

En esta sección se establece el objetivo principal del presente trabajo de investigación y se definen una serie de objetivos particulares que ayudarán al cumplimiento del mismo.

1.2.1 Objetivo general

Realizar la búsqueda de estrellas simbióticas en las bases de datos fotométricas de la misión GAIA, a través de diferentes técnicas de aprendizaje automatizado, con la intención de obtener una lista de objetos astronómicos con el perfil de estrellas simbióticas.

1.2.2 Objetivos específicos

- Explorar la Base de datos que se va a utilizar de entre los datos que se han liberado de la misión GAIA.
- Caracterizar las estrellas simbióticas de acuerdo a los datos que se recuperen de las bases de datos.
- Determinar la o las técnicas más eficientes de Analítica de datos a utilizar.
- Programar las técnicas de Analítica de datos.
- Realizar las pruebas de búsqueda en la Base de datos.
- De la lista de objetos astronómicos encontrados, determinar cuáles son candidatos a estrellas simbióticas y presentarlos al experto para su verificación.

1.3 Hipótesis

A través de Técnicas de aprendizaje automatizado, se pueden encontrar estrellas simbióticas en las bases de datos de la misión GAIA (DR2).

Capítulo 2
Marco teórico

En este capítulo se explican los conceptos y definiciones básicas relacionadas con el presente trabajo de investigación, y que comprenden las áreas de: Astrofísica y Big Data.

2.1 Bases teóricas

A continuación se expone las bases teóricas en las cuales se sustenta el presente trabajo de investigación.

2.1.1 Astrofísica

La astrofísica según describe Suárez (2013) es una rama de la física, la cual se encarga del estudio de los astros, como la constitución química de las estrellas, la distribución de sus masas y tamaños en términos de la temperatura en su superficie; su formación tanto si son objetos luminosos como si no, así como la evolución que sigue cada estrella a lo largo de su vida, su clasificación de acuerdo a sus características físicas, entre otros. Para lograr su cometido utiliza distintas técnicas, a continuación se detallarán un par de ellas (fotometría y espectroscopía), y se presentarán algunos objetos de interés.

2.1.1.1 Fotometría

De acuerdo con Zamorano (s.f.) la fotometría consiste en la medición del brillo de las estrellas (y de otros objetos celestiales como pueden ser nebulosas, galaxias, etc.) en regiones específicas del espectro utilizando filtros específicos. A partir de la información recolectada por medio de fotometría pueden obtenerse datos sobre la estructura de los objetos, como su distancia, temperatura, enrojecimiento interestelar, entre otros.

Los primeros estudios sobre el brillo aparente de las estrellas fueron realizados por los griegos. Dichos estudios sirvieron para poder generar un primer sistema de clasificación. Este sistema introdujo la separación de los objetos en magnitudes siendo los objetos con magnitud uno los más brillantes, y los de magnitud seis aquellos más débiles a simple vista.

Posteriormente con la utilización de placas fotográficas se obtuvieron mediciones menos subjetivas acerca del brillo de las estrellas. Dichas placas son sensibles a radiación ultravioleta, la cual no es perceptible por el ojo humano.

Actualmente uno de los sistemas fotométricos vigentes es el sistema UBV el cual fue introducido por Harold L. Johnson y William Wilson Morgan en los años 50’s. Dicho sistema utiliza mediciones en tres bandas; ultravioleta, azul, y visible.

Esta técnica es más rápida de realizar y se ha utilizado desde hace ya más de cien años, por lo que existen registros de observaciones muy amplios en escala

temporal. Consiste en la medición de la cantidad de luz que llega a la Tierra en ciertos rangos espectrales que son aislados mediante filtros. Gracias a ella se pudo generar todo una teoría de formación y evolución estelar que a continuación se resume muy brevemente.

2.1.1.2 Espectroscopía

La espectroscopia ha jugado un papel fundamental para la ciencia, tanto para físicos, como para químicos. Es una técnica instrumental ampliamente utilizada para poder determinar la composición química cualitativa y cuantitativa de una muestra. El proceso se realiza mediante la utilización de patrones, espectros conocidos previamente de otras muestras. En el campo de la astrofísica, la espectroscopia ha sido una parte esencial para la búsqueda y comprensión de diversos objetos. Gracias a esto, se han encontrado objetos, que sólo se conocían teóricamente. Dentro de la gama de objetos que se buscan por medio de su espectro, se encuentran las estrellas simbióticas.

Además de ser utilizada para conocer la composición química de la fuente luminosa y en general el estado de la materia, el espectro revela si la estrella se está acercando o alejando de la Tierra, gracias a aplicar los conocimientos del efecto Doppler y, en el caso de objetos extendidos, permite conocer la velocidad y estructura del objeto.

La espectrografía inició en el siglo XIX, cuando J. Von Frauhofer complementó los estudios de Newton y analizó el espectro de la luz solar. Descubrió en él una gran cantidad de líneas y se dio cuenta de que ciertas características de los astros pueden ser conocidas y estudiadas analizando las propiedades de los espectros. De esta manera nació la espectroscopia moderna.

El uso de la espectroscopía según lo expuesto en Astromía (s.f) no se limita al estudio de las estrellas, también es ampliamente utilizada para el estudio de otros cuerpos celestes, como las nebulosas, las galaxias, los cuásares, los planetas y asteroides o los cometas.

2.1.1.2.1 Líneas de emisión y absorción

Las líneas espectrales se generan cuando la luz de un objeto que emite radiación en todas las longitudes de onda, pasa a través de un gas que la rodea (generalmente su atmósfera), algunos de los electrones en los átomos y moléculas del gas absorben parte de la energía. Las longitudes de onda de los fotones absorbidos son únicas para cada transición de cada tipo de átomo o molécula. La radiación que emerge de la nube de gas por tanto carecerá de aquellas longitudes de onda específicas, produciendo un espectro con líneas de absorción oscuras (NASA, 2018). Esto se ejemplifica gráficamente en la Figura 2.1 que se muestra a continuación.

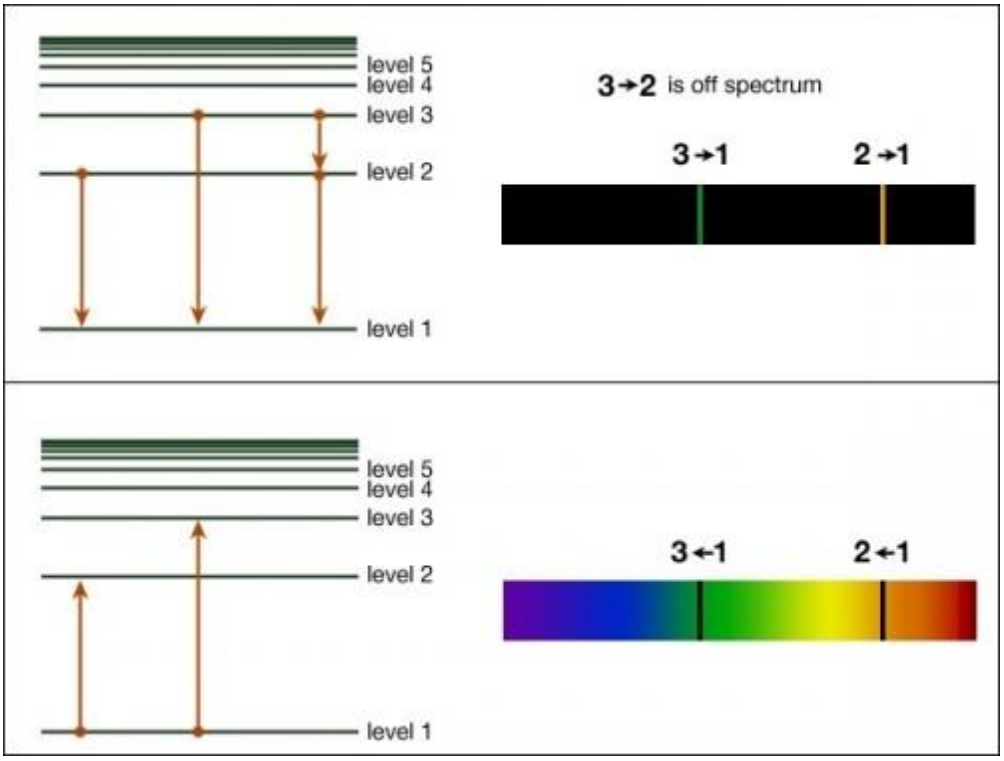


Figura 2.1 Niveles de energía de los electrones en el modelo de Bohr y como estos corresponden a la longitud de onda de líneas de absorción y emisión del espectro de un objeto (PennState, s.f.)

Así mismo cuando ese gas se encuentra a una temperatura elevada, lo que produce es una emisión en esas mismas longitudes de onda y lo que se observa son líneas de emisión.

2.1.1.3 Astrometría

La astrometría es la medición o determinación precisa de la posición de los objetos. Se usa para estudiar el movimiento de los astros y sus *paralajes*.

De acuerdo a lo expuesto en ILCE (s.f.) la astrofísica se vale de distintos métodos para la detección de objetos de interés. Uno de los métodos, de tipo astrométrico, que usa la observación con alta resolución angular es la interferometría de muy larga base o VLBI. Estos estudios permiten medir la posición de una estrella con una precisión de pocas diezmilésimas de segundo de arco. La interferometría utiliza la propiedad ondulatoria de la luz, y hace uso del fenómeno de difracción. En base a esto se tiene que la resolución angular es proporcional al diámetro del telescopio, por lo que cuanto más grande es un telescopio, mayor es su capacidad de distinguir detalles, y, como se habló en el apartado de estrellas binarias, su principal aplicación es el detectar y separar binarias astrométricas.

2.1.1.4 Ciclo de vida de una estrella

De acuerdo con Karttunen, Pountanen, y Donner (1996) y El Instituto Geográfico Nacional (s.f.), el nacimiento de una estrella o inicio de su fase en la secuencia principal inicial cuando las reacciones nucleares que transforman hidrogeno (H) en helio (He) se han iniciado y llega a un equilibrio hidrostático, es decir cuando la temperatura en el centro de la estrella llega a alrededor de 4 millones de grados y la energía generada logra detener el colapso gravitacional. A partir de este momento la evolución de la estrella dependerá de la masa que posea y del elemento que utilice como combustible. El combustible más común y básico para todas las estrellas es el hidrógeno. Sin embargo, a lo largo de su vida la estrella lo irá consumiendo. Cuando el hidrógeno se haya acabado éste se puede ver sustituido por toda una serie de elementos que serán cada vez más pesados y cuando sean consumidos se irán completando nuevas fases de la vida de la estrella. Pero cada elemento se encontrará en menor cantidad que el anterior; y será menos eficiente en la generación de energía, por lo que se infiere que las siguientes fases cada vez durarán menos.

Martínez Troya (2008), redacta en su libro que la velocidad de consumo del hidrógeno de una estrella guarda una relación directa a su masa. De esta forma estrellas con masas colosales, agotarán de una forma más rápida su hidrógeno, que estrellas con menos masa. Por tanto las estrellas con menor masa tienen una vida más larga.

Por lo tanto, para saber en qué fase de su vida se encuentra una estrella se debe analizar su “color” y su luminosidad. Estos factores dependen de su temperatura y de su tamaño.

Entre las longitudes de onda (λ) del visible, las más energéticas son interpretadas por nuestros cerebros como violeta y azul, mientras que las menos energéticas son las que corresponden al color rojo. De esta forma las estrellas con un color rojizo tendrán una temperatura muy inferior a una estrella con un color azulado.

En astronomía el color tiene una definición un tanto diferente, ya que se refiere a la diferencia entre dos magnitudes, cada una de ellas en diferente región del espectro, para definirlas se utilizan filtros especiales que permiten pasar luz correspondiente a cierto intervalo en longitud de onda. En la Figura 2.2 se muestra la curva de transmisión de algunos de los filtros más utilizados, los del sistema de Johnson o UBV (Ultravioleta, azul (Blue) y V de Visible).

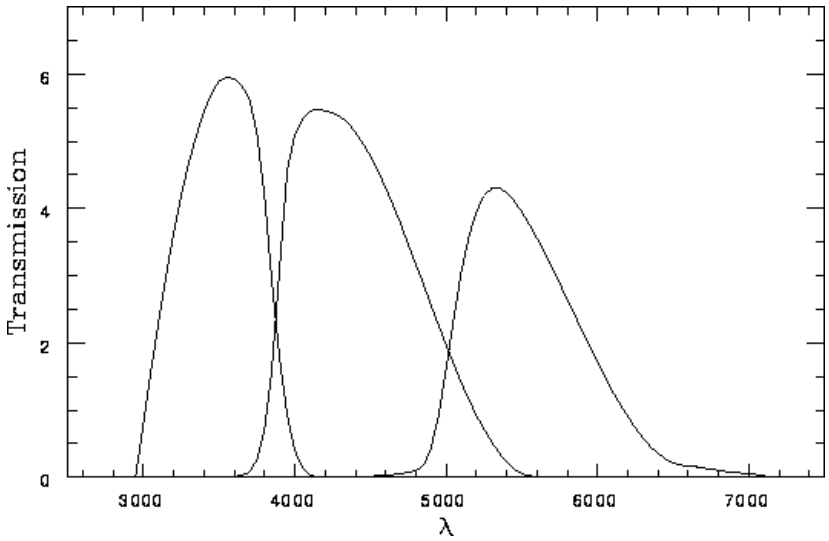


Figura 2.2 Curvas de transmisión de los filtros UBV de Johnson. (Johnson y Morgan, 1951)

Tomando en cuenta lo anterior los astrónomos trataron de encontrar alguna especie de patrón entre las propiedades de las estrellas. Por lo que a principios del siglo XX dos científicos realizaron diagramas muy parecidos (Tohmé, 2002). Hertzprung elaboró su diagrama en 1911, donde relacionaba la luminosidad de las estrellas conocidas en función de su color B-V. En 1913, Norris Russell realizó un diagrama muy parecido en el que relacionó la luminosidad de las estrellas con su tipo espectral. Al ser prácticamente iguales y desarrollados de forma independiente se denominó diagrama Hertzprung-Russell o diagrama H-R.

Al graficar las estrellas conocidas dentro del diagrama Hertzprung-Russell, se puede observar una gran población en dos regiones muy bien definidas y fácilmente identificables. La mayoría de las estrellas están localizadas en este par de regiones. Dentro de estos dos grupos el mayor es el que comprende las regiones de las estrellas luminosas y calientes hasta la región de las estrellas poco luminosas y frías. La zona donde se encuentra el grupo más numeroso, forma prácticamente una franja diagonal. A dicha franja se le denomina la Secuencia Principal (Martínez, 2008). Cuando la proto estrella se enciende entra en un punto de la secuencia principal, éste punto será donde le corresponde de acuerdo a su masa, mientras tenga hidrógeno la estrella permanecerá dentro de la secuencia principal. Dado que la mayor parte de las estrellas se encuentran en la secuencia principal, los astrónomos llegaron a la conclusión de que las estrellas pasan la mayor parte de su vida en esa secuencia. Cuando una estrella de masa semejante a la del Sol (o de hasta 8 Mo) abandona la secuencia principal, pasa a formar parte del segundo grupo que se aprecia en el diagrama Hertzprung-Russell como una bifurcación de la secuencia principal (hacia la esquina superior derecha) y es cuando la estrella empieza a consumir diversos elementos. Estos grupos o agrupaciones de estrellas se pueden apreciar en el diagrama H-R de Powell (2007) en la Figura 2.3 y se le conoce como la “rama de las gigantes” pues se trata de estrellas más frías pero de mucho mayor tamaño.

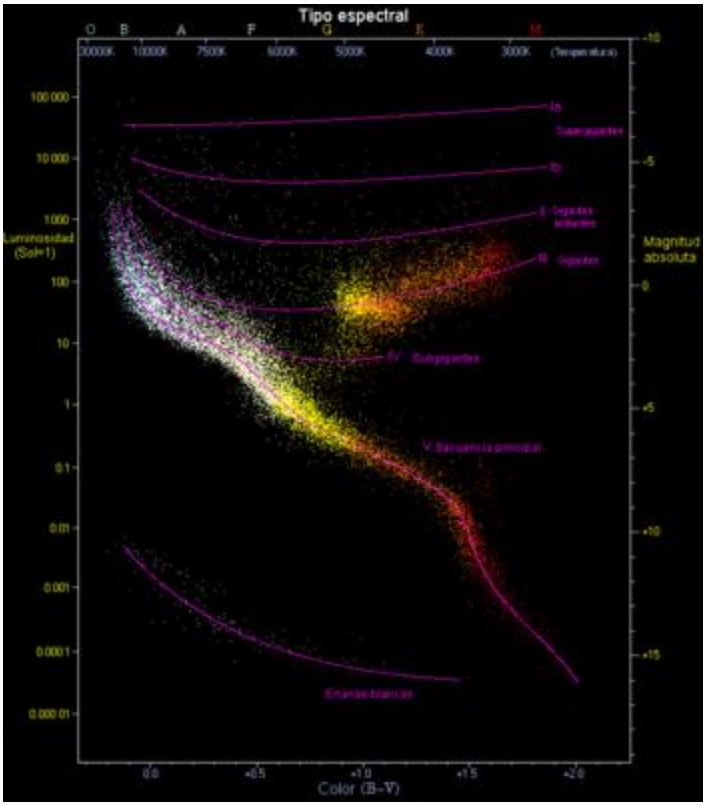


Figura 2.3 Diagrama H-R (Powell, 2007)

2.1.1.5 Estrellas binarias

El estudio de las estrellas binarias se enmarca dentro de los campos de la astronomía estelar y la dinámica estelar.

De acuerdo con las estimaciones expuestas por Echevarria (2009) la fracción de la población de estrellas binarias con respecto de estrellas solitarias es de un 80%. Por esta razón el estudio de sistemas binarios es de gran interés. Dentro de las estrellas binarias se pueden encontrar dos grandes conjuntos: estrellas binarias separadas y estrellas binarias interactivas (Echevarría Román, 2009). La principal diferencia entre los dos grupos es que las estrellas binarias separadas solo se encuentran dominadas por la fuerza de gravedad entre sus masas, haciendo que ambas estrellas giren alrededor de su centro de masa, sin embargo ésta es su única interacción, al encontrarse a grandes distancias. En cambio las estrellas binarias interactivas además de estar atrapadas dentro de la influencia gravitacional de la otra estrella, pueden llegar a tener un intercambio de masa entre ellas, adicionalmente la radiación luminosa genera otra interacción causando con ello un cambio en su comportamiento y en su evolución.

Existe otra separación válida de estrellas binarias, la cual no está ligada tanto al factor físico sino a otros aspectos, como lo es la forma de su descubrimiento. Dentro de esta separación se encuentran las estrellas binarias aparentes, este tipo de estrellas solo son binarias de tipo visual, lo que quiere

decir que no existe interacción alguna entre ellas, aunque aparentemente solo existe unos minutos de arco entre ellas, todo es un efecto de proyección.

Las binarias visuales se encuentran atrapadas dentro de la influencia gravitatoria de su compañera (Russo, s.f.), lo que quiere decir que ambas giran en torno a un centro de masa común, y como las binarias aparentes se aprecian muy cercanas. Sus movimientos mutuos orbitales son observables al telescopio, y los periodos de una órbita completa van desde cerca de un año hasta miles de años.

Las binarias astrométricas son un tipo de estrella binaria que es confundido con estrellas solitarias (Echevarría Román, 2009), ya que aparentemente en el telescopio se observa un solo objeto, no obstante su naturaleza binaria se puede inferir debido a su movimiento oscilatorio, revelando así la existencia de un cuerpo que no es visible.

Las binarias espectroscópicas son otro tipo de estrellas binarias las cuales no se pueden resolver al observarlas con el telescopio (Russo, s.f.), pero su naturaleza binaria se puede demostrar a través de su espectro. En el espectro de estas estrellas binarias se observa claramente la superposición de dos espectros diferentes, además sus líneas espectrales oscilan periódicamente en longitud de onda, lo que significa que existe una variación en su velocidad radial con la misma periodicidad y existe un desfase en su movimiento por medio periodo. Las binarias espectrales no muestran variaciones en su velocidad radial, los espectros se encuentras sobrepuestos (Echevarría Román, 2009), por lo que se infiere que la estrella binaria genera un espectro compuesto.

Las estrellas binarias fotométricas o eclipsantes son un tipo de estrella binaria que se caracteriza por que su plano de rotación es muy cercano al plano del punto de observación (Astromia, s.f.), por lo que periódicamente cada una de las estrellas de este sistema binario eclipsará la luz de su compañera, ocasionando con ello la variación periódica de su brillo.

A continuación en la Figura 2.4 se anexa la imagen recuperada de Simbad (s.f.) perteneciente a Beta Perseo o también conocida como sistema Algol. Dicha estrella se encuentra clasificada como una estrella binaria de tipo eclipsante.

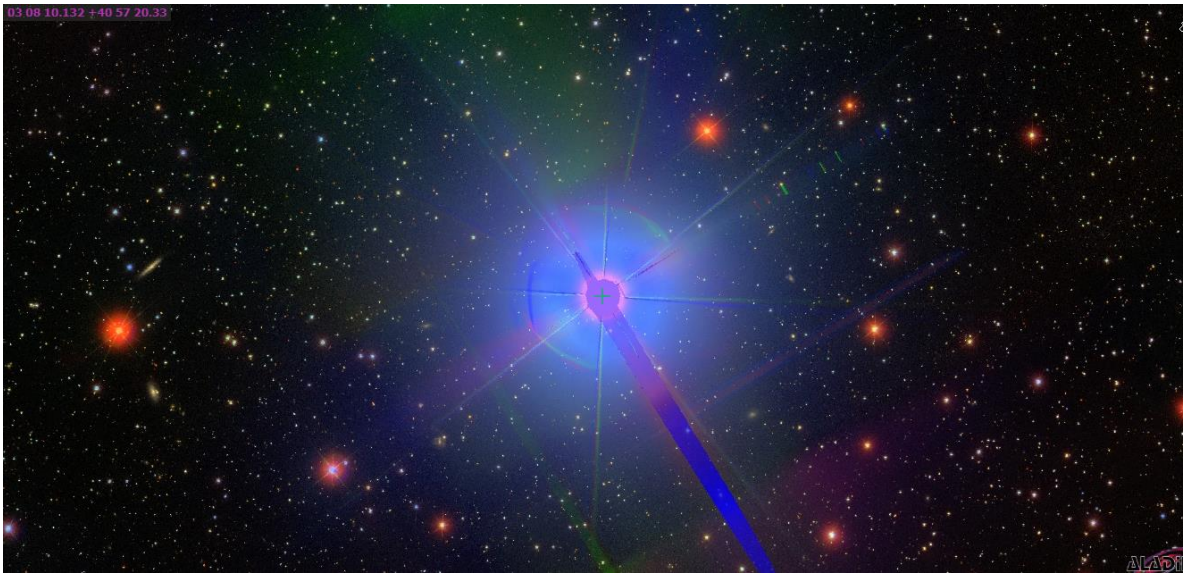


Figura 2.4 Estrella binaria eclipsante Beta Persei (Simbad, s.f.)

Por su parte en la Figura 2.5 se muestra un modelo de estrella binaria recuperada de Astromia (2018). En esta imagen se ilustra una composición binaria de tipo eclipsante.

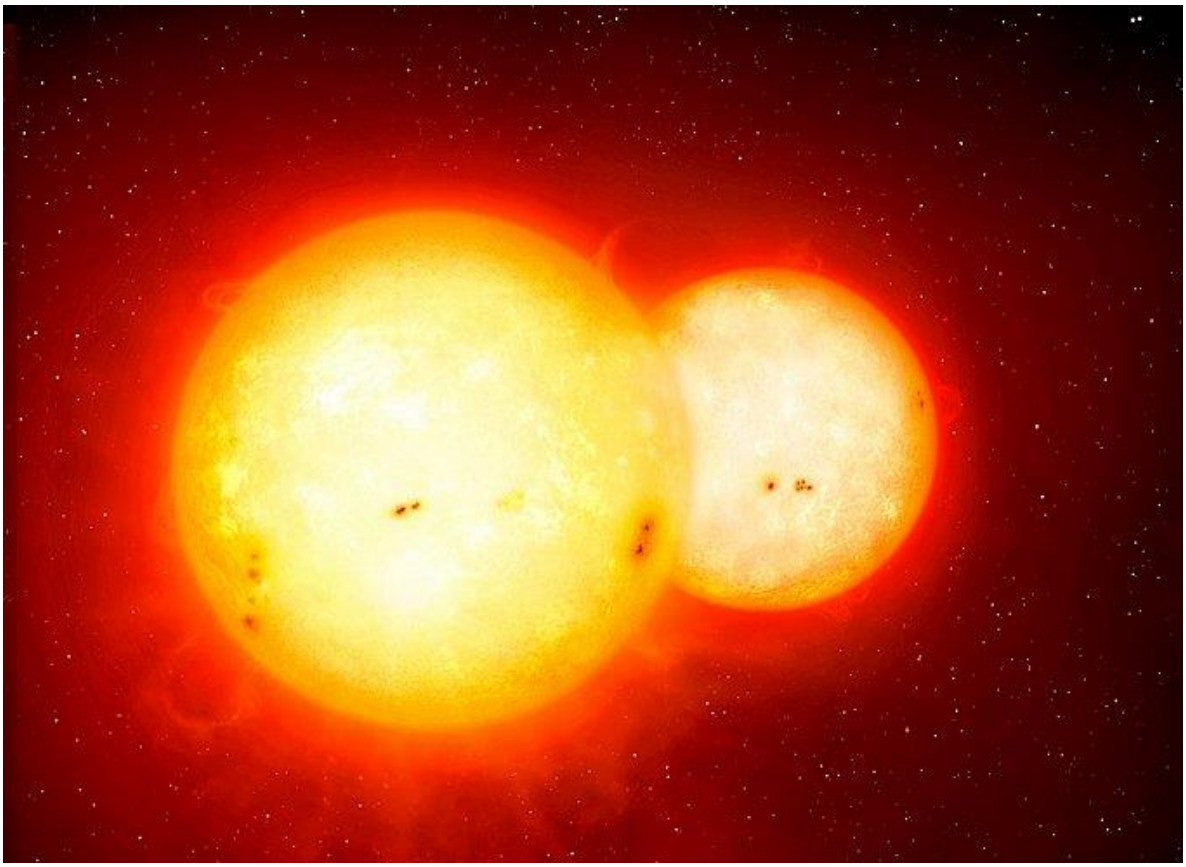


Figura 2.5 Modelo de estrella binaria (astromia, s.f.)

2.1.1.5.1 Estrellas simbióticas

Las estrellas simbióticas o sistemas simbióticos SS (Symbiotic Star), son un tipo específico de estrellas binarias, inicialmente consideradas objetos peculiares.

Generalmente se asocia a las estrellas simbióticas con un sistema formado por una estrella gigante roja y una enana blanca, sin embargo esto no tiene por qué ser siempre verdad. Con base a esto, aquellas estrellas que tengan como compañero una enana blanca serán *WD* (*White Dwarf)* simbióticas y aquellas con estrellas de neutrones o incluso hoyos negros por compañeros, se tendrán como simbióticas binarias de rayos-x (masseti 2006).

Las características del espectro de los SS comúnmente es confundido, con el de estrellas gigantes rojas y nebulosas planetarias (Kenyon & Fernandez-Castro, 1987).

En la Figura 2.6 se muestra R Aqr una estrella catalogada como simbiótica. Este sistema se encuentra conformado por una gigante roja tipo Mira y una enana blanca. En, este objeto el espectro visible se encuentra dominado por la gigante la cual es una estrella Mira de largo periodo (287 días).

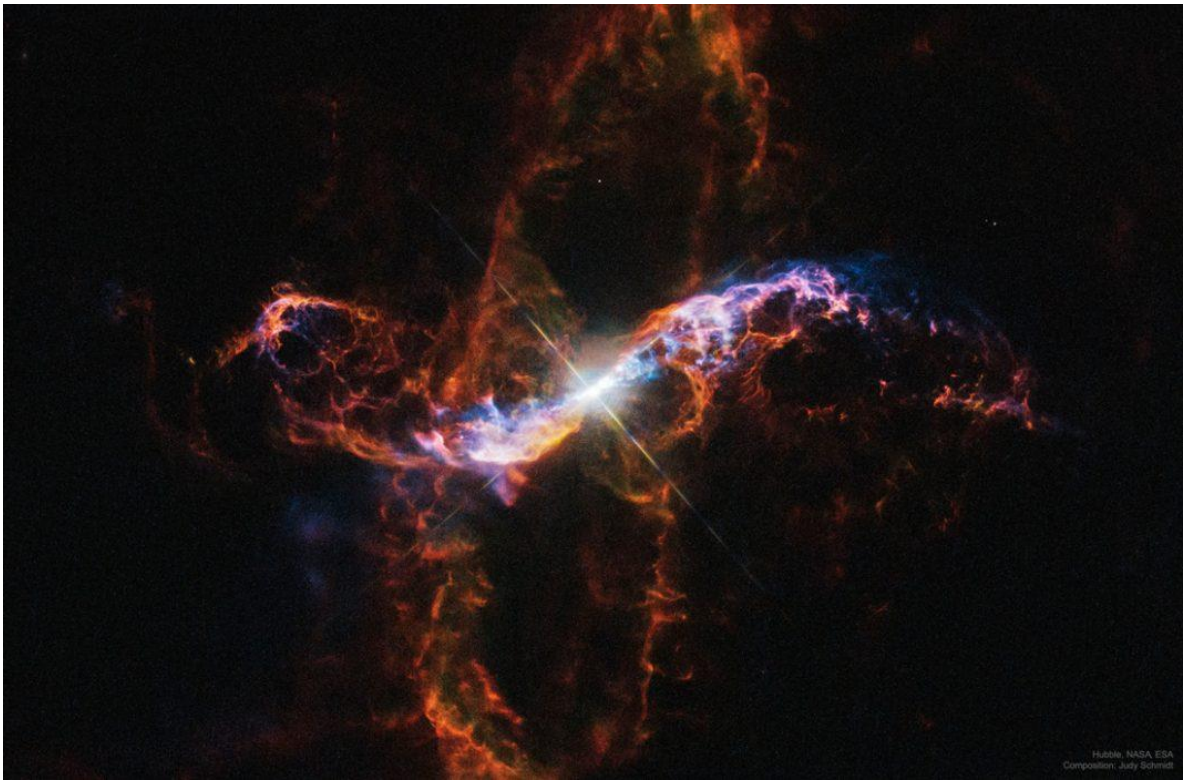


Figura 2.6 Estrella R aquarii (Schmidt, 2018)

La estrella que se muestra en la Figura 2.7 es CI cygni perteneciente a la constelación Cygnus. Esta estrella está conformada por una gigante roja, y una enana blanca (Malatesta, 2010).

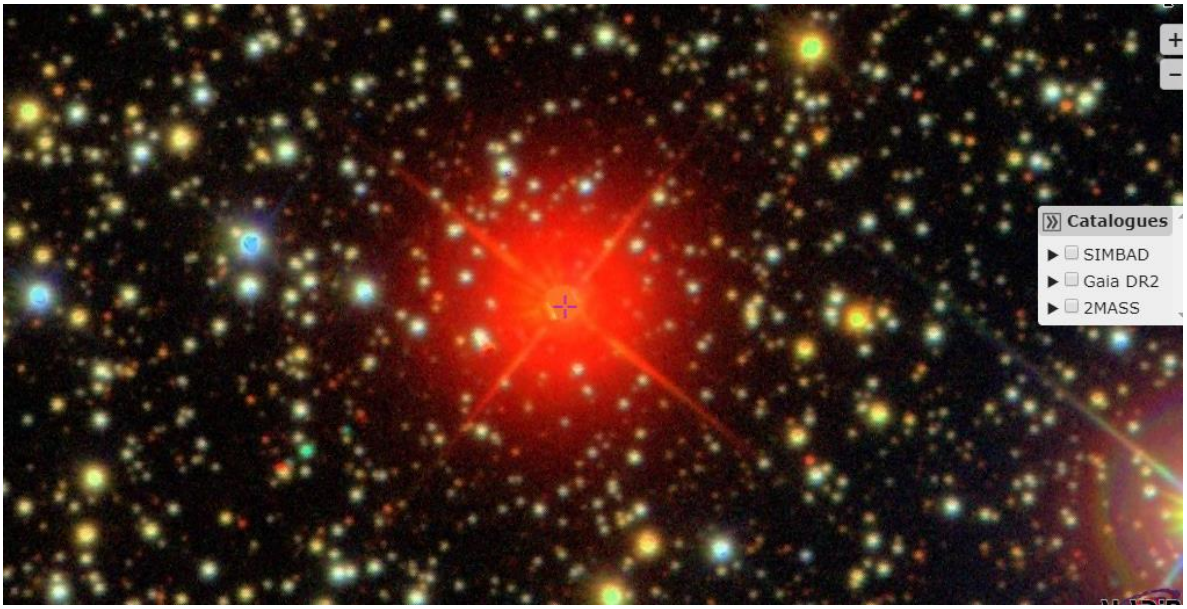


Figura 2.7 Estrella simbiótica CI cyg (Aladin, s.f.)

2.1.2 SOFTWARE

A continuación se detallarán algunas de las herramientas usadas por astrofísicos, para el estudio de diversos objetos. Dicho software también está relacionado con el trabajo de búsqueda de estrellas simbióticas.

2.1.2.1 SCISOFT

Dentro de las herramientas usadas para poder trabajar con los datos recuperados por medio de distintas fuentes, se encuentra SCISOFT (Scientific Software). SCISOFT es un proyecto creado y mantenido por la ESO (European Southern Observatory). Este proyecto proporciona una colección de distintas utilidades de interés para la astrofísica.

Las utilidades contenidas en SCISOFT son usadas para pre-procesamiento, lo que incluye pero no se limita a calibración, eliminación de ruido, análisis, entre otros. Dicho pre-procesamiento es importante ya que de esta forma se procesan los datos crudos para poder hacerlos útiles para trabajar con ellos, y obtener de ellos información de interés. Entre sus características cuenta con sistemas para análisis de datos, despliegue de imágenes directo del servidor, software de graficación, soporte para scripting y herramientas del observatorio virtual. A continuación en la Figura 2.8 se muestra la interfaz de DS9 una de las herramientas contenidas en SCISOFT.

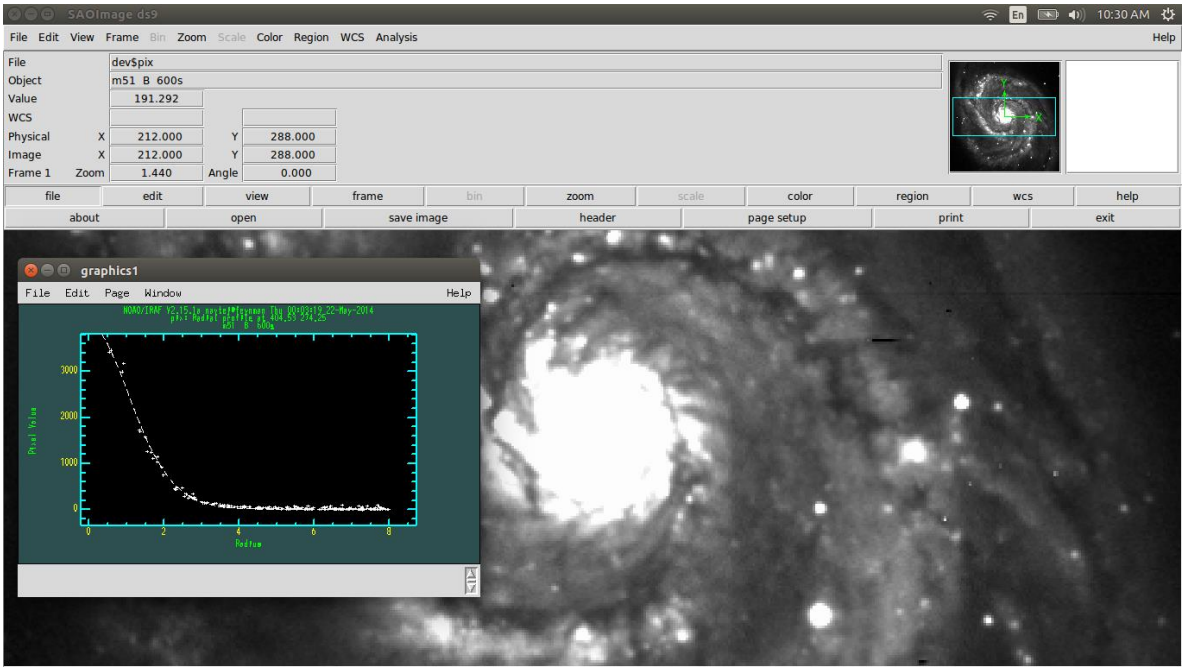


Figura 2.8 Interfaz de la utilidad ds9 perteneciente a SCISOFT ¨(captura del programa SCISOFT)

2.1.2.2 TopCat

TopCat es una herramienta desarrollada en java, dedicada a visualizar y editar información en tablas. Su objetivo es proporcionar la mayoría de las facilidades que los astrónomos necesitan para el análisis y la manipulación de *catálogos* de fuentes y otras tablas, aunque también se puede utilizar para datos no astronómicos. Comprende varios formatos diferentes de importancia astronómica (incluidos *FITS*, *VOTable* y *CDF*) y se pueden agregar más formatos.

Permite también realizar cruces de bases de datos, lo cual es muy útil cuando se desea reunir información proveniente de varios *catálogos* o bases de datos e identificar en ellos los objetos que son de interés.

Gracias a su versatilidad de graficación TopCat puede ser utilizado para emular distintos tipos de gráficos, como de densidad, o diagramas como el Hertzprung-Russell.

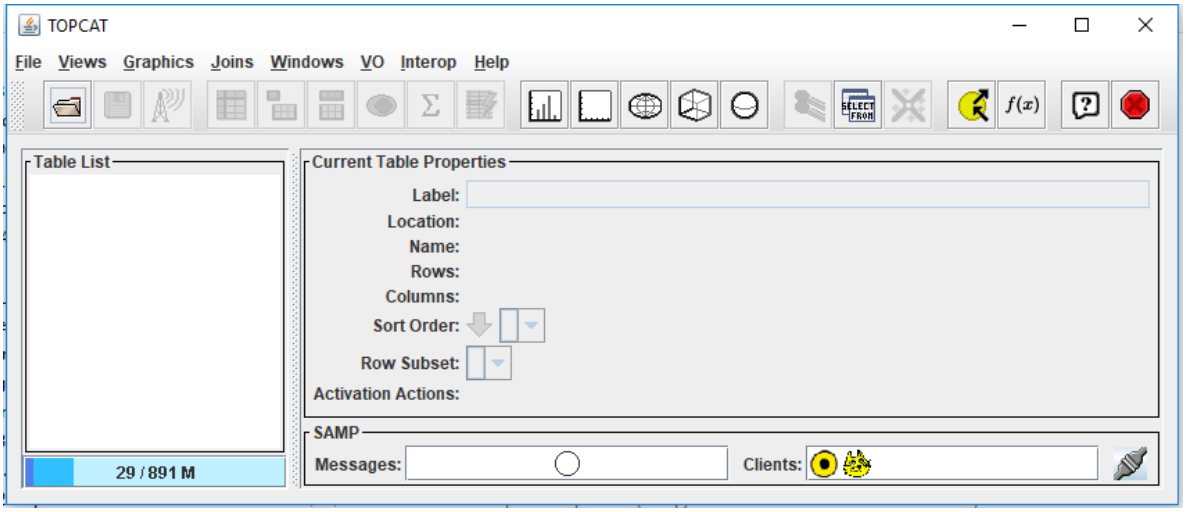
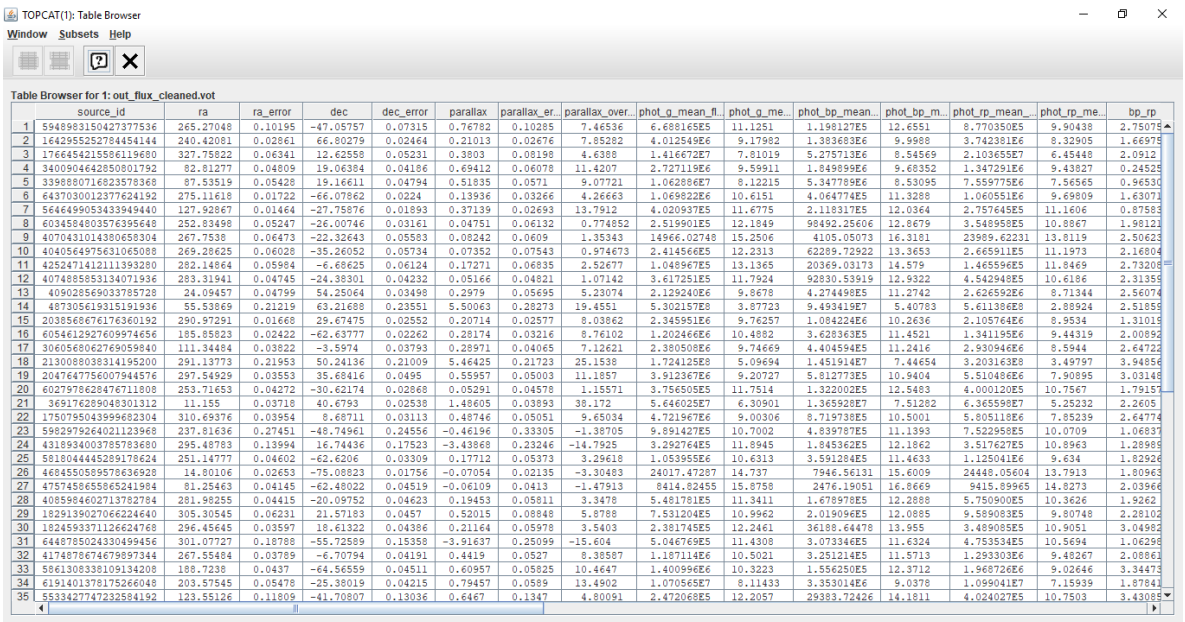


Figura 2.9 Interfaz principal de TopCat (Captura del programa TopCat)

A continuación en la Figura 2.10 se muestra el apartado para desplegar los datos contenidos en TopCat permitiendo realizar operaciones sobre dicha información. La información que se muestra depende de la tabla generada por los datos proporcionados al software, en la figura 2.8 se tiene una muestra de las estrellas simbióticas recuperadas de la base de datos de GAIA DR2. El contenido de este archivo se explicará un capítulo más adelante. Se incluye aquí tan solo para mostrar cómo despliega las tablas TopCat.



	source_id	ra	ra_error	dec	dec_error	parallax	parallax_er...	parallax_over...	phot_g_mean_fl...	phot_g_me...	phot_bp_mean...	phot_bp_m...	phot_rp_mean_...	phot_rp_me...	bp_rp
1	5948983150427377536	265.27048	0.10195	-47.05757	0.07315	0.76782	0.10285	7.46536	6.688165E5	11.1251	1.198127E5	12.6551	8.770350E5	9.90438	2.75075
2	1642955525784454144	240.42081	0.02861	66.80279	0.02464	0.21013	0.02676	7.85282	4.012549E6	9.17982	1.383683E6	9.9988	3.742381E6	8.32905	1.66975
3	1766454215586119680	327.75822	0.06341	12.62558	0.05231	0.3803	0.08198	4.6388	1.416672E7	7.81019	5.275713E6	8.54569	2.103655E7	6.45448	2.0912
3	3400904642850801792	82.81277	0.04809	19.06384	0.04186	0.69412	0.06078	11.4207	2.727119E6	9.59911	1.849699E6	9.68352	1.347291E6	9.43827	0.24525
3	3398880716823578368	87.53519	0.05428	19.16611	0.04794	0.51835	0.0571	9.07721	1.062886E7	8.12215	5.347789E6	8.53095	7.559775E6	7.56565	0.96530
6	6437030012377624192	275.11618	0.01722	-66.07862	0.0224	0.13936	0.03266	4.26663	1.069822E6	10.6151	4.064774E5	11.3288	1.060551E6	9.69809	1.63071
7	5646499053433949440	127.92867	0.01464	-27.75876	0.01893	0.37139	0.02693	13.7912	4.020937E6	11.6775	2.118317E5	12.0364	2.757645E5	11.1606	0.87583
6	6034584803576395648	252.83498	0.05247	-26.00746	0.03161	0.04751	0.06132	0.774852	2.519901E5	12.1849	98492.25606	12.8679	3.548958E5	10.8867	1.98121
4	4070431014380658304	267.7538	0.06473	-22.32643	0.05583	0.08242	0.0609	1.35343	14966.02748	15.2506	4105.05073	16.3181	23989.62231	13.8119	2.50623
4	4040564975631065088	269.28625	0.06028	-35.26052	0.05734	0.07352	0.07543	0.974673	2.414566E5	12.2313	62289.72922	13.3653	2.665911E5	11.1973	2.16804
4	4252471412111393280	282.14864	0.05984	-6.68625	0.06124	0.17271	0.06835	2.52677	1.048967E5	13.1365	20369.03173	14.579	1.465596E5	11.8469	2.73208
4	4074885853134071936	283.31941	0.04745	-24.38301	0.04232	0.05166	0.04821	1.07142	3.617251E5	11.7924	92830.53919	12.9322	4.542948E5	10.6186	2.31359
4	4090285690337857280	24.09457	0.04799	54.25064	0.03498	0.2979	0.05695	5.23074	2.129240E6	9.8678	4.274498E5	11.2742	2.626592E6	8.71344	2.56074
4	4873056193151591936	55.53869	0.21219	63.21688	0.23551	5.50063	0.28273	19.4551	5.302157E5	3.87723	9.493419E7	5.40783	5.611386E8	2.88924	2.51859
2	2038568676176360192	290.97291	0.01668	29.67475	0.02552	0.20714	0.02577	8.03862	2.345951E6	9.76257	1.084224E6	10.2636	2.105764E6	8.9534	1.31015
4	6054612927609974656	185.85823	0.02422	-62.63777	0.02262	0.28174	0.03216	8.76102	1.202466E6	10.4882	3.628363E5	11.4521	1.341195E6	9.44319	2.00892
3	3060568062769059840	111.34484	0.03822	-3.5974	0.03793	0.28971	0.04065	7.12621	2.380508E6	9.74669	4.404594E5	11.2416	2.930946E6	8.5944	2.64722
3	2130088038314195200	291.13773	0.21953	50.24136	0.21009	5.46425	0.21723	25.1538	1.724125E3	5.09694	1.451914E7	7.44654	3.203163E3	3.49797	3.94856
3	2047647756007944576	297.54929	0.03553	35.68416	0.0495	0.55957	0.05003	11.1857	3.912367E6	9.20727	5.812773E5	10.9404	5.510486E6	7.90895	3.03148
3	6027978628476711808	253.71653	0.04272	-30.62174	0.02868	0.05291	0.04578	1.15571	3.756505E5	11.7514	1.322002E5	12.5483	4.000120E5	10.7567	1.79157
1	3691762859048301312	11.155	0.03718	40.6793	0.02538	1.48605	0.03893	38.172	5.646025E7	6.30901	1.365928E7	7.51282	6.365598E7	5.25232	2.2605
4	1750795043999682304	310.69376	0.03954	8.68711	0.03113	0.48746	0.05051	9.65034	4.721967E6	9.00306	8.719738E5	10.5001	5.805118E6	7.85239	2.64774
3	5982979264021123968	237.81636	0.27451	-48.74961	0.24556	-0.46196	0.33305	-1.38705	9.891427E5	10.7002	4.839787E5	11.1393	7.522958E5	10.0709	1.06837
4	4318934003785783680	295.48783	0.13994	16.74436	0.17523	-3.43868	0.23246	-14.7925	3.292764E5	11.8945	1.845362E5	12.1862	3.517627E5	10.8963	1.28989
5	5818044445289178624	251.14777	0.04602	-62.6206	0.03309	0.17712	0.05373	3.29618	1.053955E6	10.6313	3.591284E5	11.4633	1.125041E6	9.634	1.82926
2	4684550589578636928	14.80106	0.02653	-75.08823	0.01756	-0.07054	0.02135	-3.30483	24017.47287	14.737	7946.56131	15.6009	24448.05604	13.7913	1.80963
3	4757458655865241984	81.25463	0.04145	-62.48022	0.04519	-0.06109	0.0413	-1.47913	8414.82455	15.8758	2476.19051	16.8669	9415.89965	14.8273	2.03966
4	4085984602713782784	281.98255	0.04415	-20.09752	0.04623	0.19453	0.05811	3.3478	5.481781E5	11.3411	1.678978E5	12.2888	5.750900E5	10.3626	1.9262
1	1829139027066224640	305.30545	0.06231	21.57183	0.0457	0.52015	0.08848	5.8788	7.531204E5	10.9962	2.019096E5	12.0885	9.589083E5	9.80748	2.28102
3	1824593371126624768	296.45645	0.03597	18.61322	0.04386	0.21164	0.05978	3.5403	2.381745E5	12.2461	36188.64478	13.955	3.489085E5	10.9051	3.04982
3	6448785024330499456	301.07727	0.18788	-55.72589	0.15358	-3.91637	0.25099	-15.604	5.046769E5	11.4308	3.073346E5	11.6324	4.753534E5	10.5694	1.06298
3	4174878674679897344	267.55484	0.03789	-6.70794	0.04191	0.4419	0.0527	8.38587	1.187114E6	10.5021	3.251214E5	11.5713	1.293303E6	9.48267	2.08861
3	5861308338109134208	188.7238	0.0437	-64.56559	0.04511	0.60957	0.05825	10.4647	1.400996E6	10.3223	1.556250E5	12.3712	1.968726E6	9.02646	3.34473
3	6191401378175266048	203.57545	0.05478	-25.38019	0.04215	0.79457	0.0589	13.4902	1.070565E7	8.11433	3.353014E6	9.0378	1.099041E7	7.15939	1.87841
3	5533427774732584192	123.55126	0.11809	-41.70807	0.13036	0.6467	0.1347	4.80091	2.472068E5	12.2057	29383.72426	14.1811	4.024027E5	10.7503	3.43085

Figura 2.10 Herramienta de visualizacion de tablas de TopCat (Captura del programa TopCat)

2.1.2.3 Bases de datos

La European Space Agency (ESA), es una organización internacional dedicada a la exploración espacial, con 22 Estados miembros (ESA,2018). Dentro de sus proyectos se encuentra el proyecto GAIA que es una misión para trazar un mapa tridimensional de la Vía Láctea, en el proceso se revelará la composición, la formación y la evolución de la Galaxia. GAIA proporcionará mediciones de velocidad transversal y radial con una precisión sin precedentes con la idea de producir un censo estereoscópico y cinemático de aproximadamente mil millones de estrellas de la Vía Láctea y en todo el grupo local.

El plan de liberación de las Bases de Datos del proyecto GAIA es:

primera entrega :14 de septiembre de 2016

segunda entrega: 25 de abril de 2018

tercer entrega: año 2020

cuarta entrega: año 2022

Como parte de la más reciente liberación de datos de abril del 2018, se proporcionan mediciones fotométricas y astrométricas, *paralajes* y movimientos propios.

Esta base de datos permite realizar consultas, a través de distintas formas, ya sea por medio de coordenadas en su interfaz o cargando un archivo de texto, o por medio del lenguaje ADQL (Astronomical Data Query Language). Además permite refinar búsquedas añadiendo distintos filtros sobre los parámetros disponibles. La interfaz principal de la base de datos se muestra a continuación en la Figura 2.11.

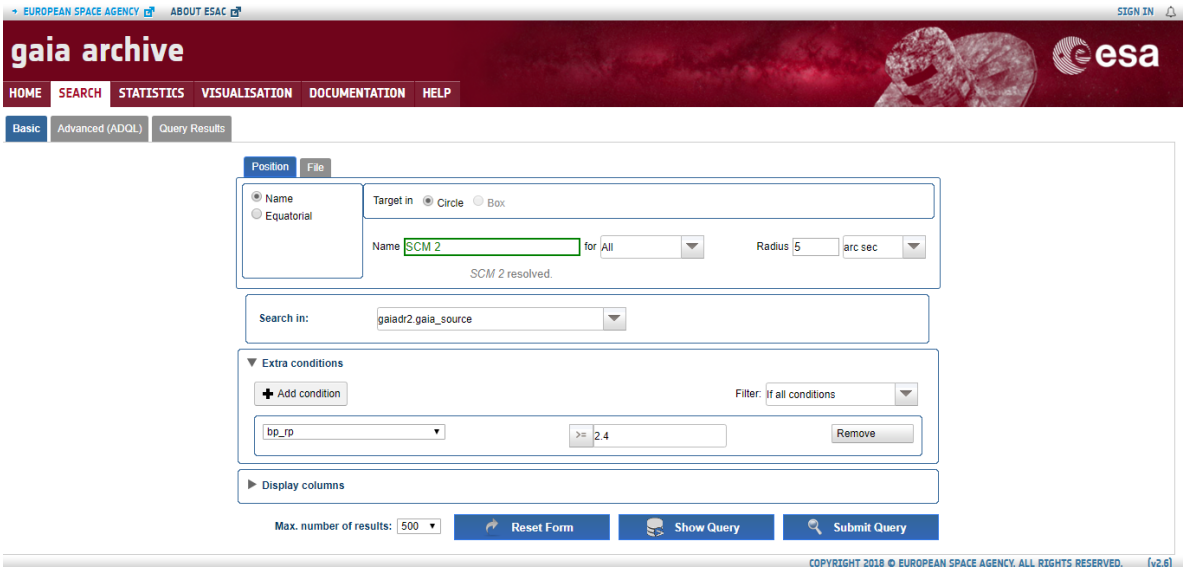


Figura 2.11 Interfaz principal de la base de datos GAIA (Captura de la interfaz de GAIA)

Dentro de las búsquedas que permite realizar GAIA, es posible seleccionar qué valores se desea recuperar. Los posibles valores se aprecian a continuación en la Figura 2.12. Aunque dichos valores aparezcan disponibles para todos los objetos, es importante recalcar que para un objeto dado, la base de datos podría no contar con un valor determinado.

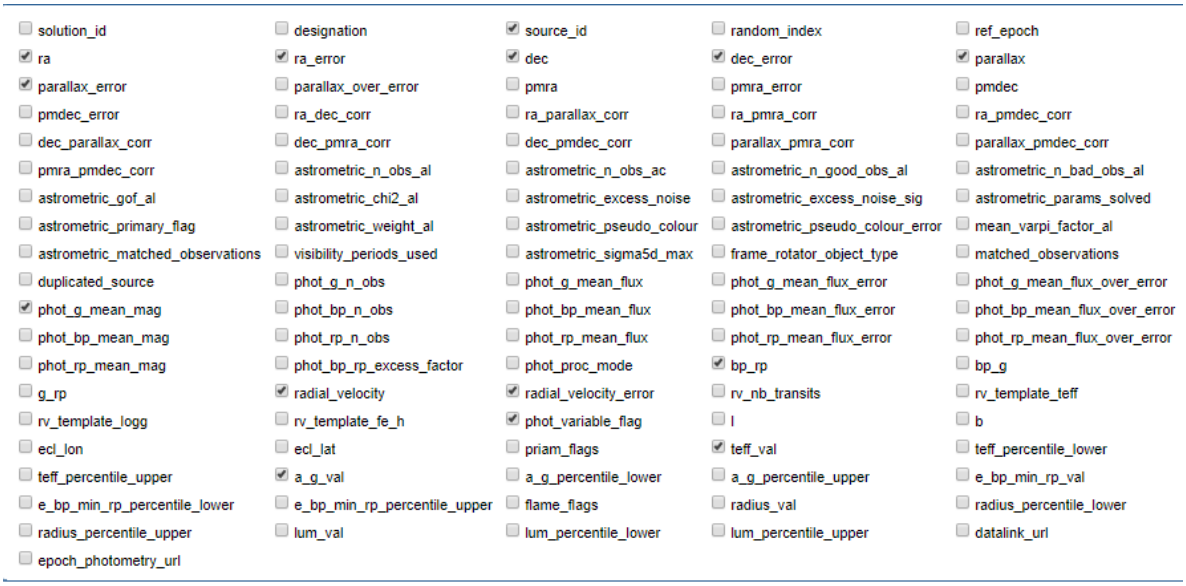


Figura 2.12 Valores disponibles para consulta en la base de datos de GAIA (Captura de la interfaz de GAIA)

SIMBAD (por sus siglas del inglés Set of Indications, Measurements, and Bibliography for Astronomical Data), es una base de datos astronómica, que fue creada por la fusión del *Catálogo* de Identificaciones Estelares (Catalog of Stellar Identifications, CSI) y el Índice Bibliográfico de Estrellas (Bibliographic Star Index) que existían en el Centro Informático de Meudon en 1979. Apartir de esa fecha, se

han ido añadiendo otros *catálogos* y se mantiene actualizada la base de datos (SIMBAD, 2019). La información es pública y contiene valores fotométricos.

SIMBAD recopila información de otras bases de datos, y los despliega a través de una sola interfaz. La interfaz principal de SIMBAD se aprecia en la Figura 2.13.

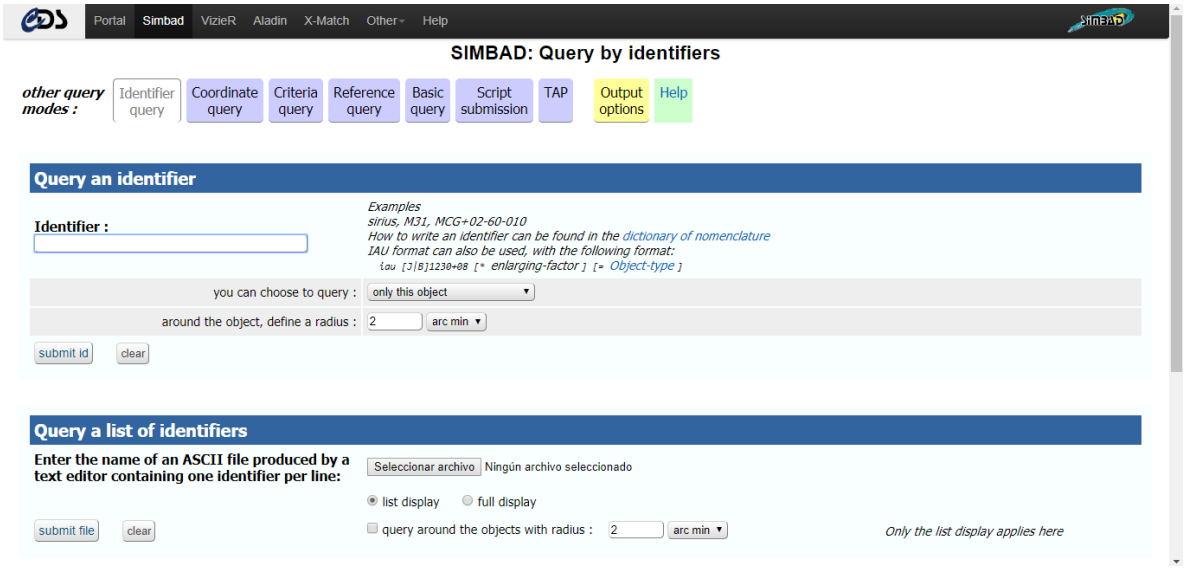


Figura 2.13 Interfaz principal de SIMBAD (Captura de la interfaz de SIMBAD)

SIMBAD proporciona una serie de valores pertenecientes a distintas bases de datos. Dentro de la información disponible para consulta se cuenta con una recopilación de los distintos nombres del objeto consultado, y un identificador correspondiente a las bases de datos de las que SIMBAD obtiene la información. La interfaz que ofrece para un objeto en concreto se aprecia en la Figura 2.14.

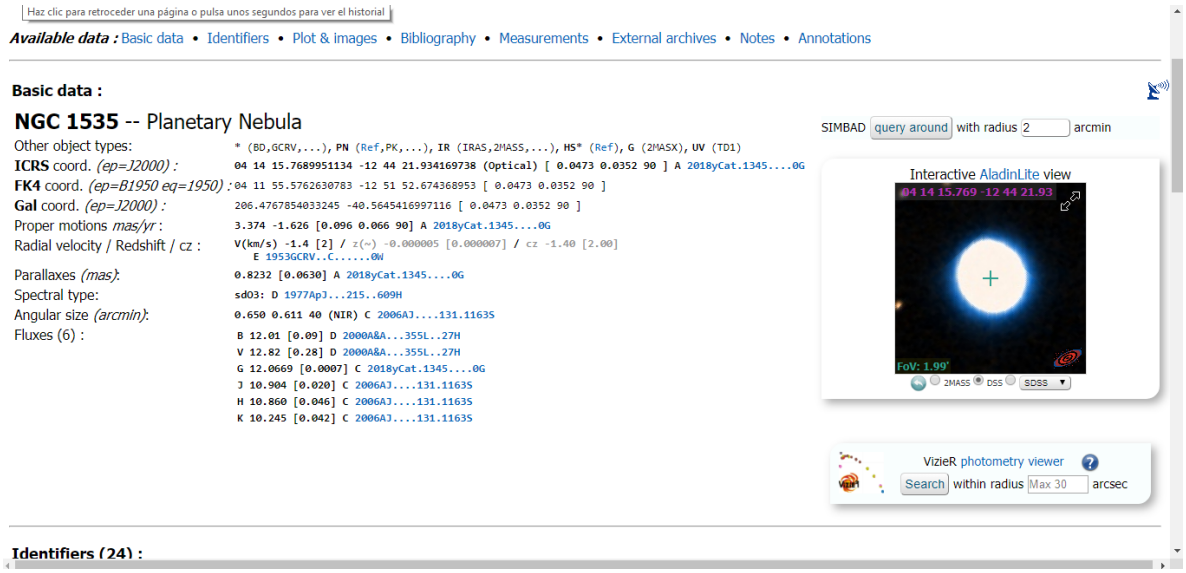


Figura 2.14 Interfaz de SIMBAD perteneciente a un objeto (Captura de la interfaz de SIMBAD)

2.1.3 Big Data

Big Data es un término evolutivo que describe cualquier cantidad voluminosa de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraídos para obtener información (Rouse, 2017).

Big Data se caracterizan a menudo por tres “Vs”: el Volumen extremo de datos, la gran Variedad de tipos de datos y la Velocidad a la que se deben procesar los datos. Aunque “Big Data” no equivale a ningún volumen específico de datos, el término se utiliza a menudo para describir terabytes, petabytes e incluso exabytes de datos reunidos a lo largo del tiempo.

Tales datos pueden provenir de innumerables fuentes, como registros de ventas comerciales, los resultados recogidos de experimentos científicos o sensores en tiempo real utilizados en el internet de las cosas (IoT). Los datos pueden estar en bruto o ser pre-procesados utilizando herramientas de software independientes antes de que se apliquen los análisis.

Los datos también pueden existir en una amplia variedad de tipos de archivo, incluyendo datos estructurados, como datos almacenados en bases de datos, y datos no estructurados, como archivos de documentos; o transmisión de datos desde sensores. Además, Big Data puede incluir múltiples fuentes de datos simultáneas, que de otro modo no podrían ser integradas.

La velocidad se refiere al lapso de tiempo en el que se deben analizar grandes volúmenes de datos. Cada gran proyecto de análisis de datos va a registrar, correlacionar y analizar las fuentes de datos, y luego proveer una respuesta o resultado basado en una consulta general. Esto significa que los analistas humanos deben tener una comprensión detallada de los datos disponibles y tener cierto conocimiento de qué respuesta se está buscando.

Con la evolución de Big Data y el trabajo de los analistas se han agregado otras dos “Vs” a las descritas previamente, dando un total de 5 “Vs”. Las nuevas características son Veracidad y Valor (Barranco, 2012).

La veracidad se refiere al grado de confianza que se establece sobre los datos a utilizar. Dentro de la caracterización del Big Data la veracidad determina su cuarta dimensión, y es de gran importancia para un analista de datos, ya que la veracidad de los mismos determinará la calidad de los resultados y la confianza en los mismos. Por lo tanto un alto volumen de información que crece a velocidad muy rápida y basada en datos estructurados y desestructurados y provenientes de una gran variedad fuentes, hacen inevitable dudar del grado de veracidad de los mismos. Por ello, dependiendo de la aplicación que se les dé, su veracidad puede ser imprescindible o convertirse en un acto de confianza sin llegar a ser vital.

El valor es un concepto que es muy parecido a la veracidad. La diferencia está en que su estudio se realiza antes de la captación de información. Tan importante es la veracidad de los datos, como que los datos tomados aporten

valor. Por lo tanto, antes de la toma de datos, habrá que investigar cuales son los que van a aportar ese valor necesario. Posteriormente es cuando se analiza su veracidad.

2.1.3.1 Analítica de datos

De acuerdo con Víctor Barrera (Barrera, V., 2016) se conoce como Analítica de Datos, o Data Analytics a todas aquellas tareas orientadas a la exploración de los datos, con la intención de encontrar patrones o conocimiento útil, que permita optimizar o rentabilizar un proceso de negocio.

De entre los proyectos que se pueden mencionar donde se ha llevado a cabo el uso de una solución de Big Data se encuentran:

- El Language, Interaction and Computation Laboratory (CLIC) en conjunto con la Universidad de Trento en Italia, son un grupo de investigadores cuyo interés es el estudio de la comunicación verbal y no verbal tanto con métodos computacionales como cognitivos.
- Lineberger Comprehensive Cancer Center - Bioinformatics Group utiliza Hadoop y HBase para analizar datos producidos por los investigadores de The Cancer Genome Atlas(TCGA) para soportar las investigaciones relacionadas con el cáncer.
- El PSG College of Technology, India, analiza múltiples secuencias de proteínas para determinar los enlaces evolutivos y predecir estructuras moleculares. La naturaleza del algoritmo y el paralelismo computacional de Hadoop mejora la velocidad y exactitud de estas secuencias.
- La Universidad Distrital Francisco José de Caldas utiliza Hadoop para apoyar su proyecto de investigación relacionado con el sistema de inteligencia territorial de la ciudad de Bogotá.
- La Universidad de Maryland es una de las seis universidades que colaboran en la iniciativa académica de cómputo en la nube de IBM/Google. Sus investigaciones incluyen proyectos en la lingüística computacional (machine translation), modelado del lenguaje, bioinformática, análisis de correo electrónico y procesamiento de imágenes.

Barranco realiza un análisis (Barranco, 2012) donde dice que cada uno de los anteriores proyectos son representativos de lo que él considera la clasificación básica para los proyectos de Big Data, misma clasificación que fue publicada por Soares en dataversity (Soares, 2012) la cual consiste en proyectos Web y social media, sistemas de máquina-a-máquina, Big Transaction Data y biométrica.

2.1.3.1.1 Enfoques

Dentro de Analítica de Datos existen distintos enfoques, para tratar los datos, y darles un verdadero valor. Dependiendo de cuál enfoque se seleccione, será las técnicas y herramientas que se usarán. Los principales enfoques son la analítica descriptiva y analítica predictiva.

De acuerdo con el análisis expuesto por Mikel Niño (Niño, 2016) se procederá a explicar en qué consiste la analítica descriptiva y la analítica predictiva.

La analítica descriptiva permite conocer las características de diversos fenómenos de interés y ayuda a descubrir tendencias y patrones de comportamientos a partir del análisis de datos históricos que, de otra manera, habrían pasado inadvertidos para los tomadores de decisiones.

Por otro lado la analítica predictiva está basada en métodos matemáticos avanzados como la minería de datos y el machine learning. Esto hace posible la creación de modelos que pronostican la ocurrencia de algún evento y guían la toma de decisiones.

Así mismo la analítica prescriptiva a través de técnicas de simulación y optimización, entre otras, permite detectar las alternativas óptimas dentro una gama de posibilidades y señala los caminos que más conviene seguir.

2.1.3.1.2 Técnicas y aplicaciones

De acuerdo a IntelDig (s.f) los enfoques de la analítica de datos utilizan diversas técnicas para lograr sus cometidos. Tienen una estrecha relación con la inteligencia artificial, específicamente con detección de patrones y machine learning. Otra de las áreas en que se apoya la Analítica de Datos es la estadística. A continuación se detallarán algunas de las técnicas más comunes, que forman parte de la mayoría de los proyectos.

Análisis de conglomerados (Cluster Analysis): de acuerdo con Borracci y Arribalzaga (2005) esta técnica agrupa en un mismo clúster a todas aquellas observaciones con características similares. En otras palabras, dos observaciones en distintos grupos en cierto sentido no son similares. Ejemplo: Segmentación de clientes en donde cada cliente corresponde a una observación.

Árboles de decisión (Decision Trees): según Breiman (1984) los árboles de decisión son una herramienta basada en reglas de decisión, la cual utiliza un diagrama lógico (árbol de decisión en forma de diagrama de flujo) para la realización de un proceso de clasificación. Ejemplo: Clasificación de transacciones con tarjetas de crédito como “Fraudulenta” o “No fraudulenta” a partir de atributos característicos de las transacciones.

Regresión lineal (Linear Regression): Peláez dice que la regresión lineal es un proceso estadístico para identificar la relación entre una variable dependiente y una o más variables independientes. Ejemplo: valorar la influencia de variables atmosféricas sobre las ventas de un producto y/o servicio.

Series de tiempo (Time Series): Villavicencio explica que las series de tiempo son un modelo estadístico para la descomposición y pronóstico de valores

futuros de una serie de datos cronológicamente ordenados. Además los datos son dependientes entre si ya que se encuentran ordenados en espacio y tiempo de forma uniforme. Ejemplo: el pronóstico del volumen de clientes en una empresa de servicios de televisión por cable.

Investigación de operaciones (Operations Research): Hillier(1997) expone en su libro que la investigación de operaciones permite conseguir la solución óptima o casi óptima de un problema complejo de optimización, por ejemplo: evaluar si es óptima la cantidad de tiendas de una empresa de bienes de consumo; este análisis puede realizarse teniendo presente el volumen de ventas de cada tienda, el volumen e ingreso per cápita de la población alrededor, entre otros indicadores sociales.

Redes neuronales artificiales (RNA): Borracci y Arribalzaga (2005) explican que las RNA son modelos matemáticos que emulan el proceso cerebral de aprendizaje mediante el entrenamiento, validación y decisión. Son utilizadas principalmente para resolver problemas de clasificación y pronóstico, por ejemplo: identificación de rostros a partir de atributos físicos y predicción de indicadores económicos.

Máquinas de Soporte Vectorial (MSV): según lo expuesto por Auria (2008) MSV son más potentes que las redes neuronales. Para efectos prácticos, los problemas de clasificación resueltos con redes neuronales también pueden ser resueltos con MSV. Las MSV realizan el aprendizaje de los datos llevando los datos a una dimensión mayor que su dimensión original.

2.2 Antecedentes

A continuación se presentan distintos artículos publicados que se relacionan con el tema de tesis que se está tratando. En estos se exponen diversos experimentos o propuestas, que representan información de interés, ya que muestran los errores y contratiempos que sufrieron, y sientan una base al explicar cómo resolvieron sus problemas.

En primer lugar se tiene que Carlos Allende Prieto (2003) desarrolló un sistema de clasificación espectroscópica que es independiente de los parámetros físicos. Basado en el sistema MKK (Morgan, Keenan & Kellman 1943), el cual establece una serie de reglas para asignar categorías espectrales a espectros de resolución media a baja. El modelo propuesto por Allende tiene la ventaja de proveer una referencia estándar independientemente de los modelos. Sin embargo el modelo es artificial y forzado, ya que las categorías espectrales definidas están correlacionadas con los parámetros atmosféricos estelares.

Por otra parte el sistema MK no proporciona estrellas pobres en metales, por lo que una estrella de este tipo como la estrella gigante pobre en metales HD 122563, ha sido comúnmente clasificada como una late-F o una estrella tipo early-

G. Por otra parte los métodos de clasificación basados en parámetros físicos son más naturales, pero dependientes de modelos. Allende sienta una base para la definición, creación e implementación de nuevos modelos, para casos específicos de investigación donde los modelos estándar no son totalmente útiles o resultarían más complicados. Este artículo aporta a la presente investigación, la metodología para la creación de nuevos modelos de acercamiento a problemas específicos, en especial para objetos peculiares.

Por otro lado se tiene a Kheirdastan (2016) que presenta en su artículo un estudio que hizo sobre las distintas técnicas para clasificación. Entre los algoritmos que utilizó para su estudio se encuentran las redes neuronales, las máquinas de soporte vectorial, los sistemas expertos y *k-means*. El ámbito en el que realizó su investigación y aplicó los algoritmos fue en la clasificación de espectros de estrellas. El conjunto de datos usados corresponde a un set de espectros estelares tomados del Sloan Digital Sky Survey SEGUE-1 y SEGUE-2. Para la investigación se usaron 100, 000 muestras de espectros correspondientes a las estrellas localizadas en el set de datos descargado para entrenamiento, y un total de 300, 000 para pruebas. Los datos fueron seleccionados para cubrir todos los tipos de espectros disponibles en el conjunto de datos descargado. Cada algoritmo fue probado 3 veces con el algoritmo de análisis de componentes principales (o por sus siglas en ingles *PCA)* con componentes principales de 280, 400 y 700. Como resultado de las pruebas se concluyó que la mejor opción para la clasificación de espectros, son las redes neuronales, las cuales con los distintos *PCAs* tuvieron un margen menor de error y un tiempo menor de cálculo. Como segunda opción las máquinas de soporte vectorial, y en tercer lugar *K-means* el cual durante las pruebas ocupó menos tiempo que las máquinas de soporte vectorial, sin embargo presenta un error mayor. De este artículo, se toma como referencia el estudio acerca de las técnicas de clasificación y reconocimiento de patrones, pero aplicadas al análisis de espectros de estrellas.

Por su parte Zhong-bao (2016) complementa la información ya que en su artículo realiza una comparativa entre 2 técnicas de reducción de dimensionalidad. Las 2 técnicas comparadas son *PCA* y el algoritmo de proyecciones de preservación local o *LPP (Locality Preserving Projections)*. *PCA* es ampliamente utilizado en la clasificación de espectros estelares. Por otra parte *LPP* aún no ha sido ampliamente utilizado en astronomía. La ventaja de *LPP* es que puede preservar la estructura local de los datos después de la reducción de dimensionalidad. Ambos algoritmos en la investigación de Zhong-bao fueron comparados utilizándose en conjunto con máquinas de soporte vectorial. El objetivo era realizar la clasificación de estrellas tipo K y tipo F. Los datos fueron tomados del proyecto Sloan Digital Sky Survey. Después de la comparativa, el autor encontró que el desempeño del algoritmo que implementaba *LPP* era mejor que en el que se implementó *PCA*. Sin embargo también se encontró que *LPP* conlleva un costo computacional mayor que *PCA*.

Liu (2017) menciona en su trabajo que aunque las máquinas de soporte vectorial, aplicadas a la clasificación de espectros estelares, se desempeñan bien en la práctica, no toman la distribución de las clases de espectros en consideración. Además se menciona que la efectividad de la clasificación esta extremadamente influenciada por el ruido. En base a esto Liu propone el uso de "fuzzy minumum within-class support vector machine" (FMWSVM). FMWSVM toma en cuenta tanto la distribución de las clases y la resistencia al ruido y provee una solución robusta. Con el fin de sustentar su propuesta Liu realizó un análisis experimental para comparar la efectividad de ambas técnicas. Para dicho experimento se tomaron los datos del Sloan Digital Sky Survey. Los datos consisten en 4 subclases del espectro tipo K: K1, K3, K5, K7, cuya relación señal-ruido (SNR) era 30<SNR<40. Se incluyeron también datos de 3 subclases de tipo F: F2, F5, y F9 cuyo SNR estaba 60<SNR<70. Del primer conjunto de datos se tomaron en total 19,453 muestras, mientras que del segundo 12,534 dando un total de 31,987 muestras. Para el tratamiento de los datos se dividieron las muestras en dos conjuntos, uno para entrenamiento y el segundo para pruebas. En el experimento se usó la técnica de reducción de dimensionalidad de datos *PCA* tanto para SVM como para FMWSVM. Como resultado para el primer conjunto de datos se promediaron los resultados de los índices de precisión obtenidos de los experimentos realizados utilizando el set de prueba, variando su tamaño de 30% a 70% con incrementos de 10% dando una precisión de 0.7894 para SVM y de 0.8625 para FMWSVM. Para el segundo conjunto de datos se obtuvo 0.6828 para SVM y 0.7658 para FMWSVM. Se concluyó que de acuerdo a los resultados obtenidos por ambas técnicas, se pueden usar para la clasificación de espectros, pero que comparado con SVM, FMWSVM se desempeña mejor con conjuntos de datos de diferente tamaño. En base a su análisis Liu termina por recalcar que el desempeño en clasificación de FMWSVM es superior a SVM. En base a esto se tiene que variando la dimensionalidad de los datos se puede alcanzar un distinto desempeño de una determinada técnica. Por lo tanto realiza una aportación sobre la metodología para la presente investigación.

Por su parte Díaz (2014) propone en su artículo la utilización de *sparse-representations*. Para la realización de su trabajo de clasificación de espectros, se tomaron 529 espectros estelares calibrados de clases B a K pertenecientes al *catálogo* Pulkovo Spectrophotometric. El conjunto de datos se utilizó para entrenamiento y prueba de la metodología. Las técnicas de *sparse-representations* han sido utilizadas para tareas de clasificación y reconocimiento en procesamiento de imágenes, dichas técnicas han reportado notables resultados. El objetivo de usar este método es identificar las características principales de cada tipo de espectro y adaptar un diccionario a ellos para usarlo para reconocer una señal del mismo tipo cuando se procese. Por esta razón, esta técnica involucra un proceso de clasificación basado en un conjunto de señales comunes con características similares en vez de un solo template. Para el trabajo de clasificación de espectros presentado la técnica propuesta es aplicada por medio del algoritmo voraz OMP (ortogonal matching pursuit). Después de la realización de 10 experimentos, se

realizaron cambios en los criterios, y utilizando técnicas de *sparse-representations* se reportó una precisión de clasificación de 84.3% con un error de +-3%, para la obtención de este resultado, realizaron el cálculo del índice de precisión promedio con base a la matriz de confusión generada por el modelo. Por lo que el artículo concluye que se ha probado obtener un mayor porcentaje de clasificación que el método *MCC* (máximum correlation coeffcients) en la mayoría de clases estelares.

La metodología presentada alcanza altos niveles de precisión en la clasificación, tanto como una red neuronal artificial. Pero los autores insisten en que representa una ventaja pues la construcción de diccionarios de aprendizaje se hace de forma automática, mientras que con redes neuronales es necesario un paso de verificación. Este trabajo refleja parte de la importancia de la presente investigación, aplicando nuevas técnicas al campo de la física y demostrando que aunque su técnica no es tan usada en ese ámbito, es muy competitiva, y podría ser superior con el crecimiento del proyecto. Además aporta una nueva técnica que probar para la clasificación de las estrellas simbióticas, sentando un precedente de comparación contra *MCC* y redes neuronales.

A través de su publicación Luna (2012) expone que algunas de las características que se le atribuyen a los espectros de estrellas simbióticas y por las cuales son identificadas son: las bandas de TiO que caracteriza la fotosfera de la gigante roja, líneas de emisión de H I, He II, [O III]. Sin embargo en su trabajo él ha identificado estrellas simbióticas en las cuales algunas de estas características no son detectadas, ya sea por el alto grado de variabilidad de la estrellas simbióticas y por distintos parámetros del sistema simbiótico como pueden ser el grado de ionización del gas o la densidad nebular. El trabajo de Luna (2012) representa un gran aporte a esta investigación, ya que proporciona un esquema de clasificación extendido, además de compartir una metodología para detectar estrellas simbióticas, con base a sus líneas de emisión y absorción.

Delchambre (2017) realizó una detección y clasificación de cuásares mediante la determinación de los parámetros astrofísicos de estos. El trabajo que realizó lo hizo con los datos proporcionados por el proyecto SDSS y recurrio a GAIA para la simulacion. Para trabajar con el espectro de los objetos utilizó el algoritmo de análisis de componentes principales ponderadas *WPCA*, y el algoritmo de fase ponderada. Delchambre utilizó estos algoritmos ya que reportan una fácil e intuitiva implementación y brindan un buen soporte frente a errores. El trabajo de Delchambre (2017) se encuentra cercano a lo que el presente trabajo de investigación pretende, ya que además de realizar búsqueda en la misma base de datos, la cual se ha considerado para la detección de simbióticas, realiza una búsqueda a través de parámetros astrofísicos. Lo que los diferencia es que Delchambre (2017) realizó una búsqueda sobre cuásares en vez de simbióticas, las cuales pueden llegar a ser más difíciles de identificar debido a diversos factores. Además el sólo realizó una fase de filtrado en base a parámetros

astrofísicos, mientras que para la búsqueda de estrellas simbióticas se planeó una fase de filtrado por parámetros astrofísicos, y una final en base a su espectro.

Uno de los más recientes resultados de la colaboración GAIA (collaboration 2018) ha sido la reproducción del diagrama H-R utilizando los datos de la base de datos GAIA. Para poder reproducir correctamente el diagrama, los investigadores no tomaron todas las estrellas disponibles, si no que establecieron una serie de filtros para poder obtener el gráfico de una forma más sencilla. De este algoritmo se consideraron las recomendaciones que exponen para seleccionar las estrellas. Además presentan algunas correcciones sobre los objetos que recuperaron, como la corrección del desplazamiento hacia el rojo.

En Akras (2018) se realizó un estudio para utilizar técnicas de inteligencia artificial para diferenciar las estrellas simbióticas de estrellas similares emisoras de Hα. Para el estudio se usaron los filtros del infrarrojo JHK y se recurrió a las base de datos que proporcionan los valores antes opuestos. Las bases de datos que cumplieron los requisitos que plantearon fueron 2MASS y WISE. Los algoritmos que emplearon fueron árboles de decisión y el algoritmo KNN. Realizan la separación entre las simbióticas de tipos S y D, así mismo aquellas que se comportan parecidas a éstas. Además de este acercamiento en Akras (2019) se realizó una recopilación de distintas estrellas simbióticas recurriendo al *catálogo* publicado por Belczy´nski (2000) y las observaciones realizadas por Miszalski, Mikołajewska y Udalski (2013) en Hα (IPHAS).

Capítulo 3
Metodología

En este capítulo se presentan los distintos procedimientos que se han realizado para hacer uso de cada uno de los algoritmos, así como el tratamiento sobre los datos para poder ser usados.

3.1 Tipo de investigación

El enfoque de la investigación que se realizó es cuantitativo, debido a que se obtiene un número de objetos buscados y localizados a través del software diseñado, la hipótesis se puede probar y concluir si se cumple o no.

El tipo de estudio es experimental, ya que se trabajó directamente con una base de datos proporcionada por GAIA.

3.2 Universo, población, o unidades de análisis

La población son los objetos de la vía láctea los cuales han sido observados y medidos por la misión GAIA y de los que se tiene 1.3 billones (1,290,305,106,153) de registros, lo que equivale a 68,840 GB de información (ESA, s.f.). La muestra será de tipo aleatorio por conglomerados o áreas. Este muestreo se utiliza principalmente para una agrupación de elementos que presentan características similares a toda la población, también se utiliza para volúmenes de la población y como la búsqueda se realizará en una base de datos se considera el más idóneo.

3.3 Criterios de inclusión/exclusión

La confirmación de la clasificación de un objeto por parte de un *catálogo* fue el principal criterio de inclusión. El criterio para la exclusión de un objeto fue la carencia de valores críticos para su clasificación.

3.4 Muestra

La muestra fue tomada a partir de *catálogos* para estrellas simbióticas y nebulosas planetarias, mientras que para las estrellas de la secuencia principal la muestra consistió en estrellas con luminosidad de clase V tomando como tamaño de la muestra 20, 000 que es el máximo número de objetos que es posible descargar.

A partir de estas consideraciones, se estructuró la Ecuación 1 para ilustrar la integración de la muestra de las estrellas simbióticas:

$$M=C-EI \quad (1)$$

Ecuación 1 Ecuación descriptiva para selección de la muestra de estrellas simbióticas.

Donde:

M: muestra final.

C: estrellas pertenecientes al *catálogo* consultado.

EI: estrellas pertenecientes al *catálogo* consultado y de los cuales no se cuenta con toda la información.

3.5 Instrumentos

Como instrumentos para la realización de la investigación fueron utilizados distintos software de propósito científico. Uno de estos es **Topcat**, en su versión 4.6-1 para Windows. Este se utilizó para la realizar el análisis de los datos, su graficación y concatenación de los mismos. Otro de los instrumentos fue **Python** en su versión 3.6.4., el cual fue utilizado para la codificación de los distintos algoritmos desarrollados a lo largo de la investigación. Para programar también se utilizó **Java JD**K versión 1.8.0_191 para Windows x64. Java se utilizó como complemento de Python para codificar parte de la red neuronal construida.

En la sección de las librerías utilizadas se ecuentra **beautiful soup4** en su versión 4.6.3 que fue utilizada para automatizar la descarga de archivos por medio de *WebScrapping*. **Selenium** para realizar la automatización del proceso que un usuario realizaría con el mouse al momento de descargar un objeto de la base de datos GAIA. **Astroquery** en su versión 0.3.8 para utilizar la api que conecta los algoritmos con las base de datos GAIA y SIMBAD. **Astropy** en su versión 3.0.4 para poder abrir los archivos descargados de las bases de datos y además realizar distintas operaciones sobre estos archivos como crear, leer, eliminar o añadir columnas o filas, o realizar búsquedas sobre la información contenida. **Numpy** en su versión 1.15.2 para realizar distintas operaciones sobre matrices. **Pandas** en su versión 0.24.2 para facilitar la separación automática de los conjuntos de datos. **Pydot** en su versión 1.4.1 fue utilizada para la creación de los modelos en archivo de los algoritmos random forest (RF) y árboles de decisión (DT). **Graphviz** en su versión 0.10.1 fue utilizada como complemento de Pydot para convertir los modelos guardados en imágenes con el fin de ilustrar el presente documento. **Scikit-learn** en su versión 0.20.3 para facilitar la codificación de los algoritmos random forest, arboles de decisión y las máquinas de soporte vectorial, al proveer herramientas que facilita su implementación.

3.6 Aparatos

Dentro de los aparatos utilizados para la realización de la investigación se encuentra una computadora con las características que se muestran en la Tabla 3.1.

Tabla 3.1 Características de la computadora utilizada para la investigación (Fuente: elaboración propia)

Característica	Valor
Sistema operativo	Windows 10

Procesador	Intel Celeron N3060
Frecuencia del procesador	1.60 GHZ
Arquitectura	X64
Memoria Ram	4 GB

3.7 Procedimiento

La metodología implementada consiste en distintas etapas las cuales son: recolección de datos, pre procesamiento de datos, generación y entrenamiento de un modelo clasificador y por último una evaluación del modelo generado. La metodología contempla un ciclo en las últimas dos fases, el cual es usado para generar modelos de clasificación cada vez mejores. En la Figura 3.1. se puede apreciar una representación gráfica. Así mismo, se anexan los pasos realizados en el proceso de investigación y a que fase de la metodología corresponde cada uno.

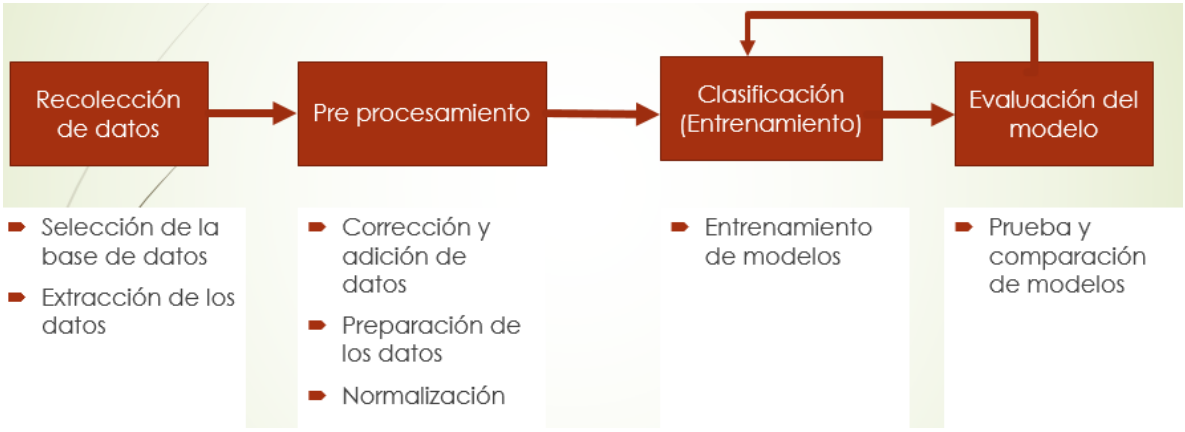


Figura 3.1 Imagen representativa de la metodología aplicada (fuente: elaboración propia)

3.7.1 Selección de la base de datos

Para la selección de la base de datos se llevó a cabo un análisis en el cual se tomaron en cuenta varios criterios. Dentro de las características que tuvieron más importancia fue la precisión o exactitud de las mediciones y como segundo punto decisivo, fue la novedad; que no existieran estudios previos relacionados con la base de datos, en este caso GAIA (DR2). La Agencia Espacial Europea (ESA por sus siglas en inglés), lanzó la misión al espacio con la inteción inicial de determinar el *paralaje* de más de 1.3 billones ($10^{9)}$ objetos en la vía láctea (Gaia Collaboration et al., 2016b). Dada la precisión de sus mediciones, es posible determinar distancias de hasta varios cientos de parsecs. Adicionalmente, GAIA realiza mediciones fotométricas en tres bandas (G, BP y RP), determinaciones de

movimientos propios y velocidades radiales de las estrellas así como espectros de baja y alta resolución.

3.7.2 Extracción de los datos

Para la extracción de datos de la base de datos de GAIA, es necesario completar un formulario con distintos datos sobre el objeto que se desea recuperar. Después de haber completado dicho proceso, se obtiene una serie de objetos que cumplen con las características solicitadas. A partir de este punto es necesario la refinación de la búsqueda, para recibir como único resultado el objeto deseado.

El proceso de extracción de datos se realizó con base al *catálogo* Belczyński, et al. (2000), que contiene un total de 188 objetos bajo la clasificación de estrella simbiótica. De dicho *catálogo* se anexa una porción en la Tabla 3.2. Las características solicitadas a la base de datos, se recibieron como una tabla en una página Web. Los resultados además de ser presentados como página Web, están disponibles para su descarga a través de un enlace generado automáticamente por GAIA. La descarga consiste en un archivo .vot. El archivo resultante de la búsqueda contiene la información de una sola estrella, lo cual es una desventaja, ya que de necesita descarga una a una cada estrella consultada.

Tabla 3.2 Catálogo Belczyński (2000) de estrellas simbióticas

No.	Name	α(2000) h m s	δ(2000) ° ′ ″	l^{II} °	b^{II} °	V [mag]	K [mag]	IR	*IUE*	X	IP_{max} [eV]
001	SMC1	00 29 10.9	−74 57 38.9			16.2		S:	+		114
002	SMC2	00 42 48.1	−74 42 00.0			16.2		S:	+	−	114
003	EG And	00 44 37.1	+40 40 45.7	121.54	−22.17	7.1	2.6	S	+	+	100
004*	SMC3	00 48 19.9	−73 31 54.9			15.5:		S:	+	+	235
005*	SMC N60	00 57 12.0	−74 13 00.0			16.8	13.0	S,D	+	−	114
006	LIN 358	00 59 24.0	−75 04 59.9			15.2	11.4	S	+	+	114
007	SMC N73	01 04 42.0	−75 48 00.0			15.5	11.6	S	+	−	114
008*	AX Per	01 36 22.7	+54 15 02.5	129.53	−8.04	10.9	5.5	S	+	−	109.3
009	V471 Per	01 58 49.6	+52 53 48.9	133.12	−8.64	13.0	9.8	D’	+	−	77.5
010*	o Ceti	02 19 20.7	−02 58 39.5	167.76	−57.98	6.0	−2.7		+	+	54.4
011*	BD Cam	03 42 09.3	+63 13 00.5	140.84	+6.44	5.1	0.2		+		77.5
012	S32	04 37 45.0	−01 19 05.9	197.48	−30.04	13.5		S	+	+	114
013	LMC S154	04 51 50.2	−75 03 36.0			15.7	10.1	D	+	−	114
014	LMC S147	04 54 04.6	−70 59 34.0			16.0	11.9	S	+		114
015	LMC N19	05 03 24.0	−67 56 35.0			16.4				−	114
016*	UV Aur	05 21 48.8	+32 30 43.1	174.22	−2.35	8.5	2.1	S	+	−	41.0
017*	V1261 Ori	05 22 18.6	−08 39 58.0	210.63	−23.72	6.8	2.1		+	+	77.5
018*	LMC1	05 25 01.0	−62 28 46.9			15.9	9.9	D	+		97.1
019	LMC N67	05 36 02.8	−64 43 23.9			15.9	11.4	S	+	−	77.5
020*	Sanduleak’s star	05 45 19.6	−71 16 09.9			16.9	13.0	D:	+		114

En la Tabla 3.2 se añade las coordenadas de la estrella, α hace referencia a la ascensión recta y δ a la declinación. Además de la información antes mencionada sobre el *catálogo* Belczyński (2000), el mismo *catálogo* proporciona información adicional, como el espectro de la componente fría, su radio, entre otros (ver Tabla 3.3). En la Tabla 3.4. se muestra una lista de nombres con los cuales se pueden referir en otros *catálogos* a la misma estrella.

Tabla 3.3 Catálogo Belczyński (2000) información adicional

No.	Cool-star spectrum	Radio [mJy]	*IRAS* F_{12}[Jy]	*IRAS* F_{25}[Jy]	References
001	C3.2:,C(232, 210)				210(fc,spc,class) 232(parm)
002	K,G-K(232, 210)				210(fc,spc,class) 232(parm)
003	M3(234)	0.54(3.6cm)	4.5	1.25	18(fc,class) 139(spc) 310(parm)
004	M0,K-M(210, 232)				210(fc,spc,class) 125(parm)
005	C3.3(232)				18(fc,class) 232(spc,parm)
006	mid K(232)				18(fc,class) 232(spc,parm)
007	K7(232)				18(fc,class) 232(spc,parm)
008	M6(234)	0.58(3.6cm)	0.32	0.10	18(fc,class) 200, 112(spc,parm)
009	G5(234)	0.45(3.6cm)	1.53	2.70	18(fc,class) 35(spc,parm)
010	Mira,M2-7(306, 134)		4881	2261	312(spc) 134(parm) 1(class)
011	M3,S5.3(5, 137)	0.18(3.6cm)	40.95	10.82	137(spc) 5(parm) 1(class)
012	C1.1CH(257)				98(fc) 64(class) 257(spc) 260(parm)
013	C2.2(232)				248(fc) 232(class,spc,parm)
014	M1,K5(232, 212)				212(fc,spc,class) 232(parm)
015	M4(211)	96(6cm)			36(fc) 211(class,spc,parm)
016	Mira C8.1Je(238)	<0.34(3.6cm)	69.41	20.64	18(fc,spc,class) 247(parm)
017	S4.1,M3(129, 7)	<0.1(3.6cm)	7.98	2.01	1(class) 129(parm)
018	C4.3,C(232, 210)				210(fc,spc,class) 232(parm)
019	C3.2,C(232, 52)				99(fc) 52(spc,class) 232(parm)
020					18(fc) 1(class) 232(spc,parm)

Tabla 3.4 Catálogo Belczyński (2000) Nombres alternativos para cada estrella a simbiótica.

001=SMC 1=NAME SMC1=[MH95] 183

002=SMC 2=NAME SMC2

003=EG And=HD 4174=BD+39 167=SAO 36618=GCRV 403=HIC 3494=GEN# +1.00004174= AG+40 66=GC 880=DO 8473=GPM1 20=SKY# 1157=AGKR 609=IRC +40014=JP11 413= PPM 43262=HIP 3494=*IRAS* 00415+4024

004=SMC 3=NAME SMC3=RX J0048.4−7332

005=SMC N60=LHA 115−N 60=LIN 323=HV 1707(???)

007=SMC N73=LHA 115−N 73=LIN 445 a

008=AX Per=MWC 411=HV 5488=CSI+54−01331=GCRV 896=JP11 5465=*IRAS* 01331+5359

009=V471 Per=PN M 1−2=PK 133−08 1=PN VV 8=LS V +52 1=PN G133.1−08.6=PN ARO 116=CSI+52−01555=PN VV' 11=*IRAS* 01555+5239

010=o Cet=MIRA=HD 14386=RAFGL 318=SKY# 3428=GC 2796=omi Cet=ADS 1778 AP= IRC +00030=YZ 93 562=MWC 35=BD−03 353=GEN# +1.00014386J= CCDM J02194−0258AP=PLX 477=GCRV 1301=JP11 625=CSI−03 353 1=DO 430= HIC 10826=SAO 129825=68 Cet=HR 681=UBV 21604=LTT 1179=TD1 1361= PPM 184482=JOY 1AP=*IRAS* 02168−0312

011=BD Cam=HD 22649=BD+62 597=HR 1105=SAO 12874=GC 4383=IRC +60125=FK4 129= UBV 3468=CSS 79=[HFE83] 244=GEN# +1.00022649=AG+63 277=N30 751= GCRV 2027=RAFGL 506=PPM 14446=HIP 17296=SKY# 5606=UBV M 9615=PLX 758= JP11 803=CSV 328=HIC 17296=S1* 60=*IRAS* 03377+6303

012=S32=StHA 32

El proceso de descarga se llevó acabo al entrar a la página Web oficial del proyecto GAIA (https://gea.esac.esa.int/archive/) y se seleccionó el apartado **search**. Después de que se desplegó la interfaz para realizar la búsqueda, se introdujo el nombre del objeto a recuperar. En caso de que GAIA no encuentre el nombre, es necesario seleccionar la opción **ecuatorial**, he introducir los valores de la ascensión recta en el campo marcado como **RA** en la interfaz y la declinación en el campo **Dec**. El siguiente paso para realizar la descarga es dar click en **extra conditions** para desplegar el menú y agregar por medio del botón **add condition** los filtros deseados, en el caso de la recuperación de estrellas simbióticas del

catálogo de Belczyński, et al. (2000) este paso se omitió. A continuación se seleccionan los campos que se van a recuperar. Esto se hace haciendo click en **Display columns** y marcar aquellos en los que se tiene interés. Por último para iniciar la búsqueda se debe dar click en **Submit Query**. Después de esto GAIA redirecciona a una nueva interfaz donde genera un nuevo job.

Después de haberse completado la búsqueda aparecerá una lista de los objetos que cumplen con las características especificadas en la interfaz de búsqueda. Puede devolver 0, 1 o varios resultados con el nombre de una estrella. Si se devuelve 0 es necesario regresar a la interfaz de búsqueda y aumentar el valor del campo **Radius** hasta que se obtenga 1 o varios resultados. En caso de obtener un solo resultado es necesario corroborar que el objeto encontrado es el objeto de interés que se esta buscando. Este paso se puede realizar recurriendo al *catálogo* o a SIMBAD y ver que parámetros arroja, como el valor en la banda G, que sea exactamente el mismos. En caso de que la búsqueda encuentre varias coincidencias es necesario regresar a la interfaz de búsqueda y disminuir el valor del campo **Radius** y volver a realizar la búsqueda. Este proceso se sigue hasta obtener un sólo resultado o una cantidad razonablemente pequeña para realizar el proceso de corroborar el objeto explicado anteriormente.

Una vez que sólo se tiene el objeto de interés, se procede a su descarga. GAIA proporciona distintas opciones de formato para descargar el objeto; sin embargo se seleccionó VOTable por sus ventajas, dentro de las que se encuentran de que es un formato utilizado por el software TopCat y funciona perfectamente con estos archivos, permite el etiquetado de información y la creación de valores por defecto, además de agregar información descriptiva como las unidades utilizadas.

El proceso de recuperación para un objeto es un trabajo repetitivo y requiere una gran inversión de tiempo, por lo que el proceso es inviable conforme aumenta el número de objetos de interés para recuperar. Por lo que se optó por automatizar este proceso.

Para hacer la automatización de la descarga de estrellas, se recurrió a una técnica de *WebScrapping*. Se utilizó la librería BeautifulSoup de python, para realizar la extración y la librería Selenium para la automatización. Sin embargo no se obtuvieron resultados satisfactorios. Los principales problemas que se enfrentaron, fue la incapacidad de cambiar parámetros por medio de las librerías escogidas, debido al diseño de la interfaz de GAIA. Por estos motivos se optó por buscar una mejor solución.

La solución seleccionada fue hacer uso de la librería astropy disponible para python. Se diseñó, he implementó un software para lograr recuperar las distintas características de una estrella. El software desarrollado se diseñó para trabajar sobre un conjunto de coordenadas J2000 previamente proporcionadas mediante un documento de texto. A continuación en la Figura 3.2 se incluye un fragmento del código mencionado anteriormente.

```
file=open("estrellas.txt","r")
custom_search=Simbad()
custom_search.TIMEOUT = 60
custom_search.add_votable_fields("id(Gaia)", "flux(J)", "flux(H)", "flux(K)", "flux(V)", "flux(B)", "ids")
objectList=None
for line in file:
 result=custom_search.query_object(line)
 if type(result)==type(None):
  print("No encontrado", line)
 else:
  print (line)
  for star in result:
   id=str(star['ID_Gaia'],  "utf-8")
   if id.find('Gaia DR2')>=0:
    id=id[9:len(id)]
    objectList=find_all_info(objectList, id, line)
   else:
    id=str(star['IDS'],  "utf-8")
    index=id.find('Gaia DR2')
    if index>=0:
     id=id[index+9:len(id)]
     index=id.find('|')
     if id.find('|')>=0:
      id=id[0:index]
     index=id.find(' ')
     if id.find('|')>=0:
      id=id[0:index]
     objectList=find_all_info(objectList, id, line)
if(os.path.exists('salida.vot')):
 os.remove('salida.vot')
objectList.write('salida.vot', format='votable')
```

Figura 3.2 Fragmento del algoritmo en python para realizar la descarga de estrellas de GAIA (Fuente: elaboración propia)

Otra base de datos importante es 2MASS (2 Micron All-Sky Survey) la cual cuenta con valores en bandas del infrarrojo. Este proyecto escanea el cielo por hemisferios, uno de los fines que persigue es la catalogación de las estrellas y galaxias detectadas (2MASS, 2006). Los datos almacenados en esta base sin embargo no tienen relación con los almacenados en GAIA.

Con el objetivo de complementar la información recuperada de GAIA se decidió realizar una búsqueda cruzada de GAIA y 2MASS, utilizando SIMBAD, la cual es una base de datos astronómicos que es actualizada diariamente. Esta base de datos reúne la información publicada de los objetos e incluye la identificación cruzada entre las diversas base de datos de los grandes proyectos de investigación, esto de acuerdo a lo expuesto por Wenger (2000). De la información obtenida, se recuperaron los valores de las mediciones fotométricas en distintos filtros (G, Bp, Rp, B, V, J, H, K, estos últimos de 2MASS). Los parámetros recuperados de GAIA fueron: source_id, ra, ra_error, dec, dec_error, parallax, parallax_error, parallax_over_error, phot_g_mean_flux, phot_g_mean_mag, phot_bp_mean_flux, phot_bp_mean_mag, phot_rp_mean_flux, phot_rp_mean_mag, bp_rp, bp_g, g_rp, radial_velocity, radial_velocity_error, phot_variable_flag, teff_val, a_g_val. Estos valores se describen a continuación en la Tabla 3.5.

Tabla 3.5 Descripción de los campos de la base de datos de GAIA (Fuente: elaboración propia).

Campo recuperado	Descripción
source_id	Identificador único perteneciente a la base de datos de la misión GAIA. En este se encuentra codificado la posición del objeto.
RA	Valor perteneciente a la misión GAIA. Ascensión recta del objeto, el valor se encuentra expresado en formato ICRS
ra_error	Valor perteneciente a la misión GAIA. Error estándar de la ascensión directa.
Dec	Valor perteneciente a la misión GAIA. Declinación del objeto, el valor se encuentra expresado en formato ICRS
dec_error	Valor perteneciente a la misión GAIA. Error estándar de la declinación del objeto.
Parallax	Valor perteneciente a la misión GAIA. *Paralaje* estelar del objeto
parallax_error	Valor perteneciente a la misión GAIA. Error estándar del *paralaje*.
parallax_over_error	Valor perteneciente a la misión GAIA. Valor obtenido de la división del *paralaje* sobre su error.
phot_g_mean_flux	Valor perteneciente a la misión GAIA. Flujo medio integrado en

	la banda G.
phot_g_mean_mag	Valor perteneciente a la misión GAIA. Magnitud media en la banda G calculada en base a la magnitud de Vega.
phot_bp_mean_flux	Valor perteneciente a la misión GAIA. Flujo medio integrado en la banda BP.
phot_bp_mean_mag	Valor perteneciente a la misión GAIA. Magnitud media en la banda BP.
phot_rp_mean_flux	Valor perteneciente a la misión GAIA. Flujo medio integrado en la banda RP.
phot_rp_mean_mag	Valor perteneciente a la misión GAIA. Magnitud media en la banda de RP .
bp_rp	Valor perteneciente a la misión GAIA. Color resultante de la diferencia de las magnitudes de las bandas BP y RP
bp_g	Valor perteneciente a la misión GAIA. Color resultante de la diferencia de las magnitudes de las bandas BP y G
g_rp	Valor perteneciente a la misión GAIA. Color resultante de la diferencia de las magnitudes de las bandas G y RP
radial_velocity	Valor perteneciente a la misión GAIA.

	Velocidad radial espectroscópica en el marco de referencia baricéntrico solar
radial_velocity_error	Error de la velocidad radial.
phot_variable_flag	Valor perteneciente a la misión GAIA.
	Bandera de variabilidad fotométrica. Los posibles valores para este campo son *variable* que indica que el objeto se identificó y procesó como variable, *constant* lo cual hace referencia a que no se encontró variación, y *not_available* que es usado cuando no ha sido procesado y/o exportado al *catálogo*.
teff_val	Valor perteneciente a la misión GAIA.
	Temperatura estelar efectiva.
a_g_val	Valor perteneciente a la misión GAIA.
	Valor de extinción en la banda G.
B	Valor recuperado de SIMBAD.
	Magnitud en la banda B.
V	Valor recuperado de SIMBAD.
	Magnitud en la banda V.
J	Valor de 2MASS, recuperado con SIMBAD.
	Magnitud en la banda J.
H	Valor de 2MASS, recuperado con SIMBAD.
	Magnitud en la banda H.

K	Valor de 2MASS, recuperado de SIMBAD. Magnitud en la banda K.

Del *catálogo* Belczyński (2000) que contiene 188 estrellas simbióticas, sólo se lograron recuperar 102 objetos de la base de datos de GAIA, ya que algunos de ellos presentan irregularidades en sus campos, entre estas, se encuentran estrellas con parallax de 0 o falta de mediciones en las bandas G, BP o RP y por lo tanto la falta en los colores bp_rp, bp_g y g_rp. Estos valores son importantes ya que son los que permiten que se caracterizen las estrellas simbióticas. Otra de las razones por las que es importante contar con valores como el *paralaje* es que, como se explicará en la sección de corrección y adición de datos, se utiliza para calcular la distancia, como el inverso del *paralaje*, por lo que al contar con un *paralaje* de 0 se produciría una división entre 0. La distancia es utilizada para calcular la magnitud absoluta de la banda G, la cual es necesaria para ubicar una estrella en el diagrama H-R.

Además de la recuperación de las estrellas simbióticas, también se procedió a descargar 20,000 estrellas la secuencia principal descrita en el artículo Gaia Data Release 2-Observational Hertzsprung-Russell diagrams (Gaia Collaboration et al., 2018) donde se construye y describe el diagrama de Hertzprung-Russell, el cual se muestra a continuación en la Figura 3.3. También se incluyeron otros conjuntos de objetos evolucionados como lo son las nebulosas planetarias. Es importante señalar que tanto las estrellas de la secuencia principal como las nebulosas planetarias se descargaron con los mismos campos que las estrellas simbióticas y por lo tanto, se les aplicó los mismos criterios de exclusión que a éstas. Esto se hizo con el fin de poder crear un conjunto de datos compatible, para ser usado por los algoritmos con los que se hará la clasificación.

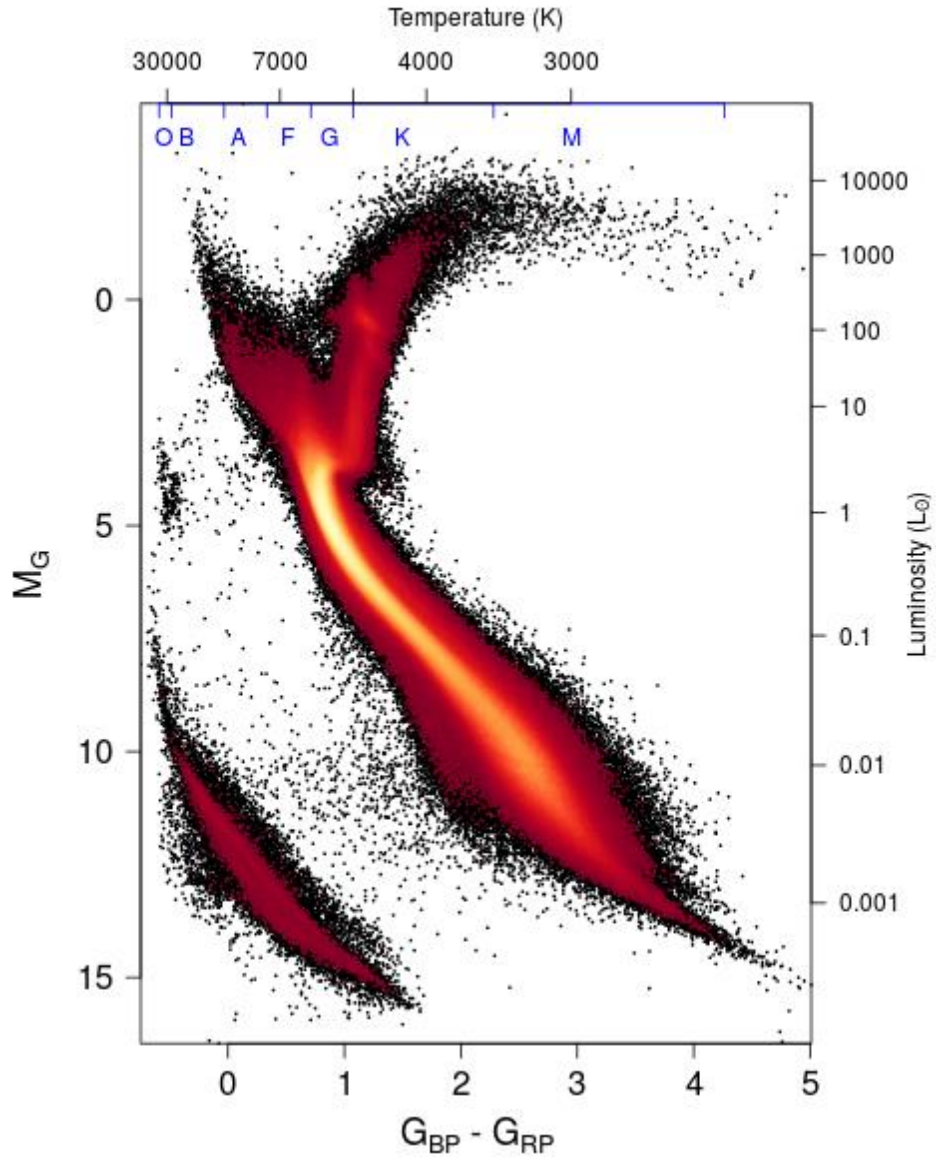


Figura 3.3 Diagrama H-R recuperado de Gaia Data Release 2-Observational Hertzsprung-Russell diagrams (Gaia Collaboration et al., 2018)

3.7.3 Corrección y adición de datos

Los objetos fueron recuperados en formato .vot en el caso de la base de datos de GAIA, y en formato Web para SIMBAD. A causa de la generación de distintos archivos y en distinto formato, fue menester la unión de todos los valores en un sólo formato, además de realizar correcciones y la generación de otros campos necesarios para la clasificación. Se agregó una columna que contiene la magnitud absoluta de cada uno de los objetos de estudio, debido a que GAIA sólo

contiene una magnitud observada. Es importante realizar el cálculo de la magnitud absoluta para poder conocer la verdadera luminosidad de un objeto sin que ésta se vea afectada por la distancia real del objeto y así tener todos los objetos con la magnitud aparente que tendrían si estuvieran a una distancia de 10 parsecs.

La magnitud absoluta en G se determinó utilizando la Ecuación 2, con la que se realiza también la corrección por extinción. Además ya que GAIA no proporciona distancias, fue necesario usar la Ecuación 3 para poder calcularla, a partir del *paralaje* del objeto.

$$Gabs=5+G-5\log(d)+Ag \quad (2)$$

Ecuación 2 Magnitud absoluta en G

$$d=(1/parallax) \quad (3)$$

Ecuación 3 Distancia a partir del paralaje

Gabs: magnitud absoluta en la banda G

G: magnitud observada en la banda G

Ag: valor de extinción en G

d: distancia expresada en parsecs

parallax: valor de *paralaje* expresado en segundos de arco

3.7.4 Preparación de los datos

Los datos descargados desde GAIA, pertenecientes al *catálogo* de estrellas simbióticas, fueron descargados de forma individual. Para la visualización de los datos se recurrió a TopCat. Sin embargo TopCat no permite concatenar la información de los archivos en grupos mayores a dos por lo que el primer paso para la preparación de los datos, fue crear un software para fusionar los distintos archivos generados.

El software encargado de realizar la fusión de los archivos descargados consiste en ejecutar una búsqueda de todos los archivos en una ruta especificada. Después de recuperar la lista de todos los archivos disponibles, solo se toma en cuenta aquellos que tienen el formato de los archivos descargados de GAIA (.vot), pero excluye al archivo out.vot, ya que este es el nombre que asigna al archivo resultante de la fusión, y si existe este archivo en la ruta especificada significa que este archivo contiene varias estrellas en su interior y tomarlo en cuenta significaría duplicar información, además de que ese archivo tiene una estructura diferente, por estos motivos se excluye de la búsqueda. Para cada nombre de archivo válido, se eliminan los últimos 4 caracteres pertenecientes a la extensión .vot, acto seguido, se abre el archivo y se le agrega una nueva columna que guardará el nombre de la estrella, mismo nombre que es con el que se guardó originalmente el

archivo descargado de GAIA. Para finalizar se envía la información de la estrella con su nombre incluido a una nueva tabla. Después de haber realizado este proceso con todos los archivos válidos disponibles, se guarda esta tabla con el nombre out.vot; en caso de que el archivo ya exista, se sobrescribe. El código encargado de realizar este proceso se muestra a continuación en la Figura 3.4.

```
def getFiles():
 return[arch.name for arch in scandir(getcwd()) if arch.is_file()]


def addNewRow(original, temp):
 for st in range(len(temp)):
  original.add_row(temp[st])

warnings.filterwarnings('ignore', category=UserWarning, append=True)
warnings.simplefilter('ignore', category=AstropyWarning)
mainTable=None
for names in getFiles():
 if(len(names)>=5):
  if(names[len(names)-4:]=='.vot'):
   star=names[:len(names)-4]
   if(star!='out'):
    if(type(mainTable)==type(None)):
     mainTable=Table.read(names, format='votable')
     #a=Column( data=['          '],name='Name' )
     #mainTable.add_column(a)
     #addName(mainTable,star)
     mainTable['g_abs_mag']=calculateAbsoluteGaiaMagnitude(mainTable[0]['phot_g_mean_mag'], mainTable[0]['parallax'], mainTable[0]['a_g_val'])
     addAbsoluteGaiaMagnitude(mainTable)
     print(names,"columnas: ",len(mainTable[0]))
    else:
     temp=Table.read(names, format='votable')
     #addName(temp,star)
     addAbsoluteGaiaMagnitude(temp)
     print(names,"columnas: ",len(temp[0]))
     addNewRow(mainTable, temp)
if(os.path.exists('out.vot')):
 os.remove('out.vot')
mainTable.write('out.vot', format='votable')
```

Figura 3.4 Algoritmo en python encargado de realizar la fusión de los archivos de GAIA (Fuente: elaboración propia)

El siguiente paso en el tratamiento de los datos fue la graficación. Para la visualización de los datos se usó TopCat. Esto se realizó a través de un diagrama magnitud-color emulando el diagrama H-R que se muestra en la Figura 3.5. Los valores que se tomaron en la abscisa son la magnitud bp_rp resultante de la resta de los filtros bp y rp. Por el lado de las ordenadas para el valor referente a la magnitud, se utilizó el filtro de la banda G de GAIA.

La visualización de los datos se realizó para comprobar si había alguna distribución de los datos. El gráfico resultante demostró que el filtro G representaba un valor aparente, por lo que se procedió a la transformación de este valor, consistente en convertir las magnitudes aparentes en magnitudes absolutas. La ecuación realiza el cálculo de la magnitud que tendría la estrella si se encontrara a 10 parsecs.

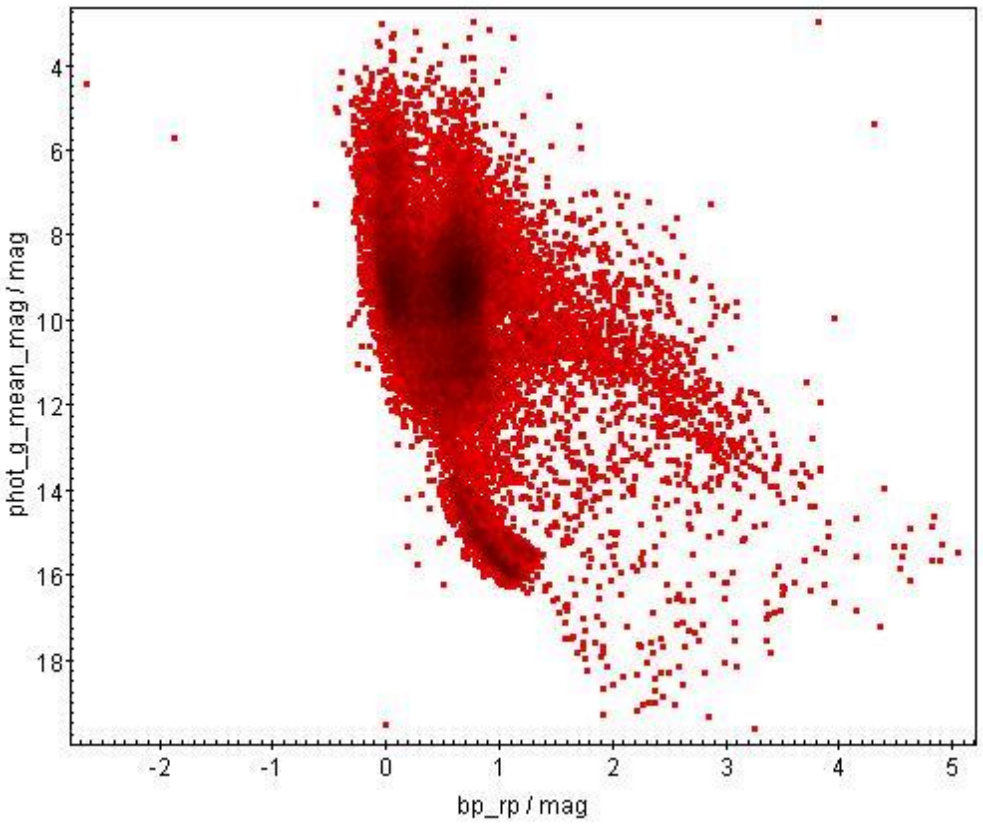


Figura 3.5 Diagrama H-R generado con una magnitud en la banda G aparente y sin corrección por extinción (Fuente: elaboración propia).

Además de la corrección de la magnitud aparente, la luminosidad de las estrellas sufre de un corrimiento hacia el rojo, el cual afecta el color observado de las estrellas. Dicho enrojecimiento es causado (entre otras cosas) por el gas y polvo presente entre las estrellas, el gas está compuesto principalmente por hidrógeno; y el polvo en su mayoría contiene grafitos y silicatos, por lo que la corrección de este enrojecimiento también es necesario para la mayor parte de las estrellas.

El resultado de la corrección de la magnitud aparente es desplegada en la Figura 3.6, donde en las abscisas se utilizó el color bp_rp como en la Figura 3.5 por otro lado, en la ordenada se utilizó el valor corregido de la magnitud de la banda G de GAIA, la cual es nombrada como g_abs_mag.

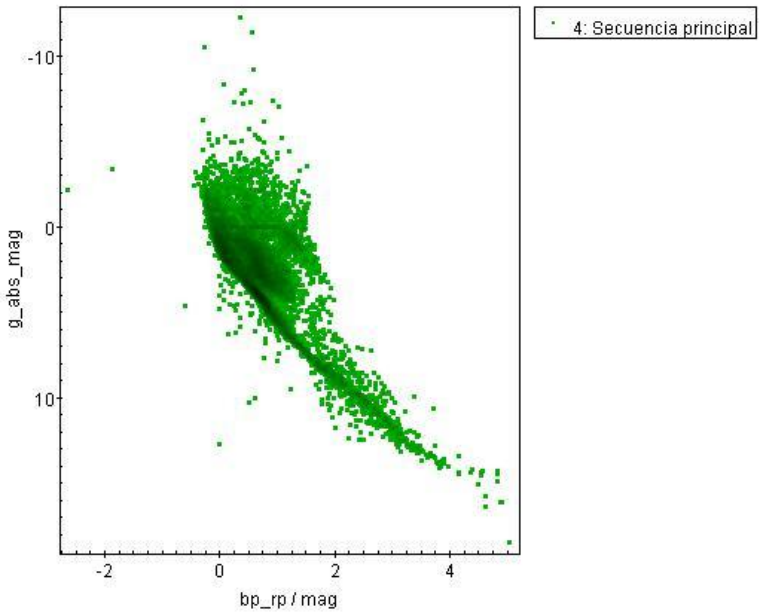


Figura 3.6 Diagrama H-R generado con una magnitud en la banda G absoluta y corregido por extinción (Fuente: elaboración propia).

3.7.5 Normalización de los datos

Los distintos conjuntos de datos se etiquetaron de acuerdo a un criterio binario: si el objeto es una estrella simbiótica o no. Posterior al etiquetado se realizó una fusión de los datos generados para cada tipo de objeto con el fin de crear un único set de datos. Sobre el conjunto de datos resultante se procedió a realizar una normalización. La cual se realizó de acuerdo a la Ecuación 4. Los valores utilizados para los distintos parámetros del vector de características se recuperaron a través de consultas sobre las bases de datos GAIA (DR2) y 2MASS. El proceso consistió en encontrar la combinación de valores frontera para cada parámetro de los objetos. Los valores se muestran a continuación en la Tabla 3.6. La normalización generada por la ecuación es necesaria para poder comparar los valores entre ellos y así definir el rango de valores máximos y mínimos que podrá adoptar cada parámetro. Para la investigación se seleccionó como valor máximo, identificado en la ecuación como d2, el valor de 1 y para el valor mínimo, identificado como d1, el valor de 0. Por lo tanto todos los parámetros del vector de características estarán comprendidos en el rango [0,1].

$$y=((x-xmin)(d2-d1))/(xmax-xmin)+d1 \quad (4)$$

Ecuación 4 Normalización de datos

Donde:

y: valor de x normalizado

x: valor a normalizar

xmin: valor mínimo posible para x

xmax: valor máximo posible para x

d1: valor mínimo que podrá tomar y

d2: valor máximo que podrá tomar y

Tabla 3.6 Valores frontera usados para la normalización de los parámetros utilizados (Fuente: elaboración propia).

Parámetro	Valor mínimo	Valor máximo
bp_rp	-5.489193	9.800095
g_rp	-4.157661	3.650905
bp_g	-2.073900	7.795892
g_abs_mag	-39.51653	26.91888
teff_val	3229	9803

3.7.6 Separación del conjunto de datos

Con el conjunto de datos normalizados se procedió a hacer una separación del 16.5% del total, con el fin de ser usados como validación para los 4 algoritmos que se programaron, 16.5% para las pruebas y el 66.6% para el entrenamiento. El conjunto de prueba se usó para verificar el nivel de aprendizaje de los algoritmos, mientras que el set de validación se utilizó para comprobar que el algoritmo fuera capaz de generalizar la clasificación y detectar de esta manera una especialización sobre los datos de entrenamiento y prueba.

De la información recuperada de ambas base de datos, se procedió a seleccionar los campos con los que trabajarían los algoritmos seleccionados. El criterio de selección utilizado, fue tomar aquellos datos que ayudan a una mejor separación de las clases a identificar. Con los parámetros elegidos, el vector de características final se compone con los campos que se muestran a continuación en la Tabla 3.7.

Tabla 3.7 Descripción de los campos del vector de características usado para los algoritmos de clasificación

Parámetro	Valor máximo
Gabs	Magnitud absoluta en la banda G
teff_val	Temperatura efectiva
bp_rp	Color resultante de la diferencia de las magnitudes de la banda

	Bp y Rp
b_rp	Color resultante de la diferencia de las magnitudes de la banda G y Rp
bp_g	Color resultante de la diferencia de las magnitudes de la banda Bp y G
j_h	Color resultante de la diferencia de las magnitudes de la banda J y H
h_k	Color resultante de la diferencia de las magnitudes de la banda H y K
b_v	Color resultante de la diferencia de las magnitudes de la banda B y V
v_k	Color resultante de la diferencia de las magnitudes de la banda V y K

3.7.7 Selección de técnicas de clasificación

Dada la naturaleza de los datos, y que ya están identificadas las estrellas simbióticas dentro del conjunto de entrenamiento, se decidió utilizar técnicas de aprendizaje supervisado.

Para la selección de la técnica de clasificación a usar se tomaron en cuenta una serie de factores, según el tipo de problema son:

- Cantidad de información disponible
- Separabilidad de los datos
- Antecedentes en problemas similares

Los algoritmos que se analizaron como posibles candidatos para usarse son: redes neuronales artificiales (ANN), árboles de decisiones, máquinas de soporte vectorial (SVM), k-ésimo vecino más próximo (KNN).

Las redes neuronales como se describen en Olabe (1998), son una herramienta muy útil ya que poseen una capacidad genérica de mapeo de patrones. Las redes son capaces de aprender una gran variedad de relaciones. Además que no requieren un conocimiento previo de la función de relación. Ofrecen una gran flexibilidad, al poder cambiar el diseño de la red fácilmente, cambiar el número de capas, interconexiones, neuronas y representación de la información. Sin embargo también presentan algunos inconvenientes como podría ser que comparados con otros algoritmos presentan una menor robustez frente a datos con ruido, la probabilidad de llegar a una no convergencia, la desventaja de no poder llegar a un modelo utilizable cuando la ANN funciona como una caja negra.

El algoritmo del K-ésimo vecino más próximo es un algoritmo que presenta ventajas interesantes, como la robustez al ruido en los datos de entrenamiento, funciones objetivo complejas y que su efectividad aumenta cuando el conjunto de datos de entrenamiento es grande. Sin embargo presenta desventajas entre las cuales resalta el coste computacional, ya que los cálculos del k-ésimo vecino aumentan con el tamaño de los datos. Dependiendo de la dimensionalidad de los datos, se requerirá un conjunto mayor de datos desventaja que comparte con la ANN.

Los árboles de decisiones según el análisis realizado en Minguillón (2002) ofrecen una gran ventaja, ya que representan un método entendible aún cuando no se tengan conocimientos avanzados de estadística. Son una opción de fácil implementación. Otra de sus ventajas es que ofrece una alta precisión y confiabilidad al clasificar, estando a la par de otras técnicas, pero con un coste computacional menor. A pesar de esto, entre sus desventajas se encuentra que no es posible cuantificar la magnitud con la que una variable aporta para la predicción de una clasificación. Y la desventaja más importarte es que pueden volverse demasiado complejos si no se tienen objetivos claros.

Las máquinas de soporte vectorial (SVM) según lo expuesto en Auria (2008) tienen la ventaja de presentar un buen rendimiento cuando los datos son no linealmente separables. Además de poder manejar una alta dimensionalidad de datos. Y la principal desventaja es que son ineficientes para su entrenamiento.

Tomando en cuenta las características de cada técnica, las ventajas y desventajas que conllevaría su implementación en el proyecto, se realizó un filtrado haciendo primero un descarte por medio de una comparativa. La decisión fue tomada, además de lo antes mencionando, poniendo atención a las técnicas usadas en otros proyectos de investigación similares, sus resultados y sus recomendaciones.

La primera comparativa fue entre ANN y KNN. Después de la revisión de la comparación entre estos algoritmos realizados en distintos proyectos y como lo expone Kheirdastan (2016) y Colas (2006), con sus respectivos experimentos encontraron que las redes neuronales tienen un buen nivel de precisión y con gran cantidad de datos en el conjunto de entrenamiento, pueden competir a la par del algoritmo KNN, sin embargo como demostró el experimento de Kherdastan (2016), las redes neuronales bajo ciertas condiciones presentaron una superioridad frente a KNN tanto en precisión como en costo computacional. Y dado que la dimensionalidad de los datos puede variar, y no se dispone de un gran set de datos para el entrenamiento, dado el limitado número de estrellas simbióticas que se conocen, se decidió que era más conveniente el uso de ANN sobre KNN.

Kheirdastan (2016) dentro de sus experimentos también comparó los algoritmos ANN y SVM y encontró que ANN puede llegar a ser más conveniente, ya que sus niveles de confiabilidad pueden llegar a ser altos, SVM supone un mayor costo computacional, y a pesar de lo mostrado en Zhong-bao (2016), Collazo (2016), donde dada la naturaleza de los datos SVM mostró estar en igualdad e inclusive, en una etapa del experimento se tuvo un mejor resultado de la SVM sobre la ANN, concluye que al ser demasiado cercano su nivel de confiabilidad no es posible realizar una comparativa real ya que influyen demasiados factores a la hora de la creación de la topología de ANN. Por lo tanto con base a lo encontrado en la literatura, y a la naturaleza binaria de los datos, su ruido, y su dimensionalidad, se decidió que sería más conveniente el uso de ANN sobre SVM.

Una comparativa entre ANN y árboles de decisión fue realizada en Muhammad (2015). La comparación se realizó entre una red neuronal y distintos algoritmos de árboles de decisión. Los resultados mostraron que todos los algoritmos de árboles fueron superiores que el algoritmo de redes neuronales. Sin embargo el estudio no aporta información sobre la velocidad o coste que presentó cada implementación de los algoritmos.

Como conclusión de la comparativa entre estos distintos algoritmos se tomó la decisión de realizar pruebas con los algoritmos de ANN y árboles de decisión, ya que en varios de los artículos consultados, estos algoritmos mostraron un mayor grado de confiabilidad, y los casos en los que fueron probados presentan similitud en el formato de los datos. En los casos donde estos algoritmos mostraron menor precisión al clasificar, el algoritmo ganador presentó una mayor complejidad, tiempo de respuesta, y coste computacional.

3.7.8 Redes neuronales

El primer acercamiento para la clasificación de las estrellas simbióticas se realizó a través de las redes neuronales supervisadas. El modelo de red neuronal que se usó fue el perceptrón multicapa. El algoritmo de entrenamiento seleccionado fue el de retro-propagación con gradiente descendente y una función

sigmoidal para activación. Una porción de la codificación de este algoritmo en java se muestra a continuación en la Figura 3.7. Cabe mencionar que este algoritmo funciona en dos etapas, en la primera etapa de propagación, dada una entrada, para cada neurona de la primera capa se realiza la suma ponderada de los valores de entrada con los valores de peso (inicializado previamente con valores aleatorios entre -1 y 1). Después de calcular la suma, el valor se pasa a una función de activación de la neurona y esta calcula un valor de salida. Este proceso se repite hasta llegar a la última capa, tomando como entrada la salida de las neuronas de la capa anterior. Después de tener la salida de la última capa, se compara con el valor esperado y se calcula el error, si el error está por arriba del mínimo error permitido, se pasa a la segunda fase que es la retro propagación. Esta etapa consiste en calcular cuánto contribuyen al error cada valor a través de derivadas parciales del error con respecto a los pesos de cada neurona. Después de realizar el cálculo se realiza un pequeño reajuste de los valores. Esto proceso se repite de la última capa hasta llegar a la primera. La principal ventaja de este algoritmo es su capacidad de mapeo de patrones.

```
class Neural_Network(object, RandomNumberOfNeurons):
  def __init__(self):
    #parameters
    self.inputSize = 8
    self.outputSize = 1
    self.hiddenSize = RandomNumberOfNeurons

    #weights
    self.W1 = np.random.randn(self.inputSize, self.hiddenSize) # (8x?) weight matrix from input to hidden layer
    self.W2 = np.random.randn(self.hiddenSize, self.outputSize) # (?x?) weight matrix from hidden to output layer
    self.W3 = np.random.randn(self.hiddenSize, self.outputSize) # (?x1) weight matrix from hidden to output layer
    self.B1 = np.random.randn(?) # (?x1) weight matrix from hidden to output layer
    self.B2 = np.random.randn(?) # (?x1) weight matrix from hidden to output layer
    self.B3 = np.random.randn(1) # (1x1) weight matrix from hidden to output layer

  def forward(self, X):
    #forward propagation through our network
    self.z = np.dot(X, self.W1) # dot product of X (input) and first set of 3x2 weights
    self.z2 = self.sigmoid(self.z) # activation function
    self.z3 = np.dot(self.z2, self.W2) # dot product of hidden layer (z2) and second set of 3x1 weights
    o = self.sigmoid(self.z3) # final activation function
    return o

  def sigmoid(self, s):
    # activation function
    return 1/(1+np.exp(-s))

  def sigmoidPrime(self, s):
    #derivative of sigmoid
    return s * (1 - s)

  #def dot(self, output, w):
```

Figura 3.7 Algoritmo BackPropagation en Java (Fuente: elaboración propia)

Con el fin de determinar la mejor topología para la red neuronal se automatizó el proceso de creación, entrenamiento y prueba del algoritmo mediante un programa en Java, el cual se puede apreciar en la Figura 3.8. Con el proceso automatizado se procedió a codificar una búsqueda aleatoria sobre espacios continuos de distintos parámetros de la red neuronal. Dichos parámetros se muestran a continuación en la Tabla 3.8.

```
CandidateGenerator candidateGenerator = new RandomSearchGenerator(hyperparameterSpace, null);


Class<? extends DataSource> dataSourceClass = ExampleDataSource.class;
Properties dataSourceProperties = new Properties();
dataSourceProperties.setProperty("minibatchSize", "402");

String baseSaveDirectory = "arbiterExamp/";
File f = new File(baseSaveDirectory);
if (f.exists()) f.delete();
f.mkdir();
ResultSaver modelSaver = new FileModelSaver(baseSaveDirectory);


ScoreFunction scoreFunction = new EvaluationScoreFunction(Evaluation.Metric.PRECISION);


//TerminationCondition[] terminationConditions = {
  //      new MaxTimeCondition(15, TimeUnit.MINUTES),
    //    new MaxCandidatesCondition(3)};

TerminationCondition[] terminationConditions = {
                new MaxTimeCondition(72, TimeUnit.HOURS),
            new MaxCandidatesCondition(100000)};


OptimizationConfiguration configuration = new OptimizationConfiguration.Builder()
        .candidateGenerator(candidateGenerator)
        .dataSource(dataSourceClass,dataSourceProperties)
        .modelSaver(modelSaver)
        .scoreFunction(scoreFunction)
        .terminationConditions(terminationConditions)
        .build();
```

Figura 3.8 Algoritmo en java para automatizar la creación de modelos de redes neuronales (Fuente: elaboración propia).

Tabla 3.8 Descripción de los parámetros utilizados para la creación automática de redes neuronales (Fuente: elaboración propia)

Parámetro de la Red Neuronal	Selección
Factor de aprendizaje	Búsqueda aleatoria en el espacio continuo [.3. .9]
Numero de Capas	Búsqueda Aleatoria en el espacio discreto 1-3
Numero de neuronas por capa	Búsqueda Aleatoria en el espacio discreto 3-255

El algoritmo de automatización se diseñó para trabajar limitado por tiempo. Dado un tiempo específico, crea la mayor cantidad de sujetos (redes neuronales) y se evalúan. Una vez finalizado el tiempo especificado, sólo se despliega la información correspondiente al sujeto con las mejores características registradas.

El proceso de evaluación que se realiza sobre cada sujeto es llevado a cabo separando el set de datos. La separación se llevó como se describe en la Tabla 3.9.

Tabla 3.9 Descripción de la división y uso del conjunto de datos (Fuente: elaboración propia).

Fracción del set de datos total (porcentaje)	Uso	Separación
2/3 (66%)	Entrenamiento	Realizada de forma aleatoria al comienzo del entrenamiento de cada red neuronal
1/6 (16.6%)	Pruebas	Realizada de forma aleatoria al comienzo del entrenamiento de cada red neuronal
1/6 (16.6%)	Validación	Realizada y separada antes del proceso de creación de redes neuronales.

3.7.9 Random Forest

El algoritmo Random Forest consta de una gran cantidad de árboles de decisión, que a diferencia de las redes neuronales, cuenta con una menor cantidad de parámetros para su construcción y entrenamiento (Keller, 2019). Los valores que deben ser cambiados son el número de clasificadores que serán generados, el número de variables a tomar y un parámetro opcional que es limitar la profundidad máxima que puede tener cada árbol generado.

La principal característica del algoritmo Random Forest es que aunque sea un algoritmo de aprendizaje supervisado como las redes neuronales, no es posible controlar la topología de los estimadores generados por el algoritmo. Otra de las características notables del algoritmo es que para la creación de cada uno de los árboles estimadores, los parámetros que usarán no son controlados, sino que son escogidos aleatoriamente. Este algoritmo también se distingue por el hecho de que el set de datos que requiere, puede o no estar normalizado.

Para el caso de estudio se realizaron pruebas variando el número de clasificadores entre [1,3000]. Para todas las pruebas realizadas se estableció como máximo número de variables a tomar por cada estimador la longitud del vector de características. Por lo que la configuración para este algoritmo quedó de la forma en que se describe en la Tabla 3.10.

Tabla 3.10 Descripción de los parámetros utilizados en la generación de estimadores para el algoritmo Random Forest (Fuente: elaboración propia).

Parámetro	Selección
Profundidad máxima	Sin límite
Árboles	[1, 3000]
Mínimo de muestras por hoja	1
Máximo de características a tomar	Número de características
Número máximo de hojas	Sin límite

El algoritmo Random Forest se escribió en Python como se muestra en la Figura 3.9. Este algoritmo lee un archivo donde se encuentra el set de datos con un total de ocho características. Al abrir el archivo el algoritmo separa de las características el etiquetado el cual se identifica con el nombre type y consiste en un valor numérico binario; 0 para estrellas no simbióticas y 1 para estrellas simbióticas. Después de separar las etiquetas en un arreglo diferente, elimina la columna del set de datos. Acto seguido, divide el conjunto de datos para los datos de entrenamiento y los de prueba utilizando la librería pandas. A continuación se configura el módulo de creación de Random Forest para clasificación con una variable la cual contiene un número aleatorio que representa el número de estimadores con el que se creará el Random Forest, también se envía el número máximo de características a tomar y el número máximo de hojas como se explica en la Tabla 3.10.

Posteriormente se entrena utilizando el conjunto train generado por la librería pandas, con su respectivo etiquetado. Una vez que se finaliza el proceso se prueba con el conjunto test. La matriz generada con las predicciones se compara con las etiquetas reales, y con base a su diferencia se calcula el error absoluto.

El algoritmo realiza el cálculo de valores como el porcentaje de error absoluto, la precisión y la matriz de confusión. Con el fin de poder entender el modelo se incluyeron instrucciones para recuperar en formato gráfico de uno de los estimadores, un ejemplo de esto se incluirá en el capítulo 4 junto con los resultados del algoritmo.

```
labels = np.array(features['type'])
features= features.drop('type', axis = 1)
feature_list = list(features.columns)
features = np.array(features)

train_features, test_features, train_labels, test_labels = train_test_split(features, labels, test_size = 0.33, random_state = 42)

rf = RandomForestClassifier(n_estimators=randomNumerOfTrees, max_features=n_features, max_leaf_nodes=None, random_state=42)
rf.fit(train_features, train_labels);

predictions = rf.predict(test_features)
# Calculate the absolute errors
errors = abs(predictions - test_labels)
# Print out the mean absolute error (mae)
print('Mean Absolute Error:', round(np.mean(errors), 2), 'degrees.')

# Calculate mean absolute percentage error (MAPE)
mape = 100 * (errors / test_labels)
# Calculate and display accuracy
accuracy = 100 - np.mean(mape)
print('Accuracy:', round(accuracy, 2), '%.')

# Pull out one tree from the forest
tree = rf.estimators_[5]

# Export the image to a dot file
export_graphviz(tree, out_file = 'tree.dot', feature_names = feature_list, rounded = True, precision = 1)
# Use dot file to create a graph
(graph, ) = pydot.graph_from_dot_file('tree.dot')
# Write graph to a png file
graph.write_png('tree.png')

print ("Test Accuracy  :: ", accuracy_score(test_labels, predictions))
print (" Confusion matrix ", confusion_matrix(test_labels, predictions))
```

Figura 3.9 Algoritmo Random Forest en python (Fuente: elaboración propia)

3.7.10 Árboles de decisión

Los árboles de decisión son un algoritmo que genera un árbol como un diagrama de flujo, donde cada nodo interno representa un valor del vector de características, las ramas representan decisiones, y los nodos hojas representan un resultado de la clasificación. Este es un algoritmo de clasificación de aprendizaje supervisado, no paramétrico, que no requiere supuestos distribucionales, permite modelar relaciones no lineales y no es sensible a la ausencia de datos como lo describe Breiman (1984). La forma básica de funcionamiento es la creación de particiones recursivas de acuerdo con reglas de asignación, partición y parada.

Una de las principales ventajas de este algoritmo es que crea un modelo de caja blanca, por lo que es fácil de comprender el resultado. Otra ventaja es que se solicita poca preparación del set de datos, ya que no requiere la eliminación de datos faltantes, o normalización.

El algoritmo de árboles de decisión se escribió en python como se muestra en la Figura 3.10. Este algoritmo lee un archivo donde se encuentra el set de datos con un total de ocho características. Al abrir el archivo el algoritmo separa de las características el etiquetado, el cual se identifica con el nombre type y consiste en un valor numérico binario; 2 para estrellas no simbióticas, y 4 para estrellas simbióticas. Después de separar las etiquetas en un arreglo diferente, se elimina la columna del set de datos. Acto seguido se divide el conjunto de datos para los datos de entrenamiento y los de prueba utilizando pandas.

A continuación se configura el módulo de creación DesicionTreeClassifier para crear un árbol de decisión con el propósito de clasificar. Posteriormente se

entrena utilizando el conjunto train generado por pandas, con su respectivo etiquetado. Una vez que se finalizó el proceso se prueba con el conjunto test. La matriz generada con las predicciones se compara con las etiquetas reales, y con base a su diferencia, se calcula el error absoluto. El algoritmo realiza el cálculo de valores la precisión y la matriz de confusión. Con el fin de poder entender el modelo se incluyeron instrucciones para recuperar en formato gráfico uno de los estimadores, un ejemplo de esto se incluirá en el capítulo 4 junto con los resultados del algoritmo.

```
labels = np.array(features['type'])
features= features.drop('type', axis = 1)
feature_list = list(features.columns)
features = np.array(features)

train_features, test_features, train_labels, test_labels = train_test_split(features, labels, test_size = 0.33, random_state = 42)


clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(train_features,train_labels)

#Predict the response for test dataset
y_pred = clf.predict(test_features)

print("Accuracy:",metrics.accuracy_score(test_labels, y_pred))
print (" Confusion matrix ", confusion_matrix(test_labels, y_pred))
```

Figura 3.10 Algoritmo de Árbol de decisión en python (Fuente: elaboración propia).

3.7.11 Máquinas de soporte vectorial

Las máquinas de soporte vectorial son un algoritmo de aprendizaje supervisado. Este algoritmo puede manejar fácilmente variables continuas y paramétricas. El funcionamiento básico del algoritmo es la creación de un hiperplano en un espacio multidimensional para la separación de las clases. Las máquinas de soporte vectorial según lo expuesto por Auria (2008) tienen la ventaja de presentar un buen rendimiento cuando los datos son no linealmente separables. Además de poder manejar una alta dimensionalidad de datos.

El algoritmo de máquinas de soporte vectorial se escribió en Python como se muestra en la Figura 3.11. Este algoritmo lee un archivo donde se encuentra el set de datos con un total de ocho características. Al abrir el archivo el algoritmo separa de las características el etiquetado el cual se identifica con el nombre type y consiste en un valor numérico binario; 2 para estrellas no simbióticas, y 4 para estrellas simbióticas. Después de separar las etiquetas en un arreglo diferente, elimina la columna del set de datos. Acto seguido, se divide el conjunto de datos para los datos de entrenamiento y los de prueba utilizando pandas. A continuación se configura el módulo de creación SVM para crear una máquina de soporte vectorial con un kernel de tipo linear. Posteriormente se entrena utilizando el conjunto train generado por pandas, y con su respectivo etiquetado. Una vez que se finalizó el proceso se prueba con el conjunto test. La matriz generada con las predicciones se compara con las etiquetas reales, y con base a su diferencia se

calcula el error absoluto. El algoritmo realiza el cálculo de valores, la precisión y la matriz de confusión.

```
labels = np.array(features['type'])
features= features.drop('type', axis = 1)
feature_list = list(features.columns)
features = np.array(features)

train_features, test_features, train_labels, test_labels = train_test_split(features, labels, test_size = 0.33, random_state = 42)



#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel

#Train the model using the training sets
clf.fit(train_features, train_labels)

#Predict the response for test dataset
y_pred = clf.predict(test_features)

print("Accuracy:",metrics.accuracy_score(test_labels, y_pred))
# Model Precision: what percentage of positive tuples are labeled as such?
#print("Precision:",metrics.precision_score(test_labels, y_pred))

# Model Recall: what percentage of positive tuples are labelled as such?
#print("Recall:",metrics.recall_score(test_labels, y_pred))

print (" Confusion matrix ", confusion_matrix(test_labels, y_pred))
```

Figura 3.11 Algoritmo de Maquina de Soporte Vectorial en python (Fuente: elaboración propia)

Capítulo 4
Resultados

Los productos generados por la investigación son los resultados de la aplicación de los algoritmos de Inteligencia Artificial, explicados en el capítulo anterior. Así mismo, se muestran las gráficas generadas a partir de la base de datos de GAIA DR2, información que primeramente se depuró y normalizó; todo este proceso fue explicado en en el capítulo 3.

La primera imagen generada fue un diagrama color-magnitud correspondiente a los valores bp_rp y g_abs respectivamente, que se muestra en la Figura 4.1. Este diagrama fue utilizado para realizar una primera caracterización de las estrellas simbióticas y saber la localización de este tipo de objetos en el diagrama Hertzsprung-Russell. Una vez que se obtuvó la gráfica se encontró que la gran mayoría de las estrellas simbióticas se encuentran en la parte superior derecha, esto es importante, porque la búsqueda de las estellas simbióticas candidatas, se realizará particularmente en esa región.

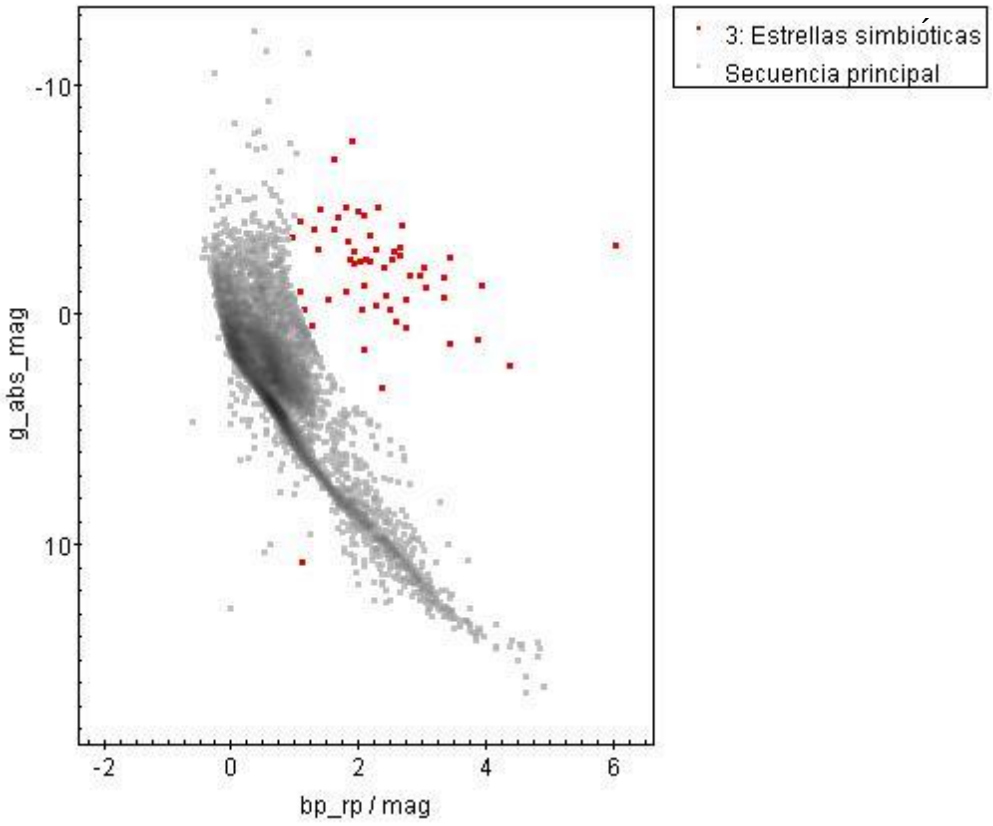


Figura 4.1 Diagrama H-R generado para caracterizacion (Fuente: elaboración propia).

Se generó el diagrama color-color para los valores b_v y j_h. El diagrama resultante se muestra a continuación en la Figura 4.2. Dicho diagrama sirve para comprobar si ambos colores son útiles para realizar una separación de clases. Por

lo que se aprecia en dicho diagrama las estrellas simbióticas si son separables de las nebulosas planetarias utilizando b_v y j_h.

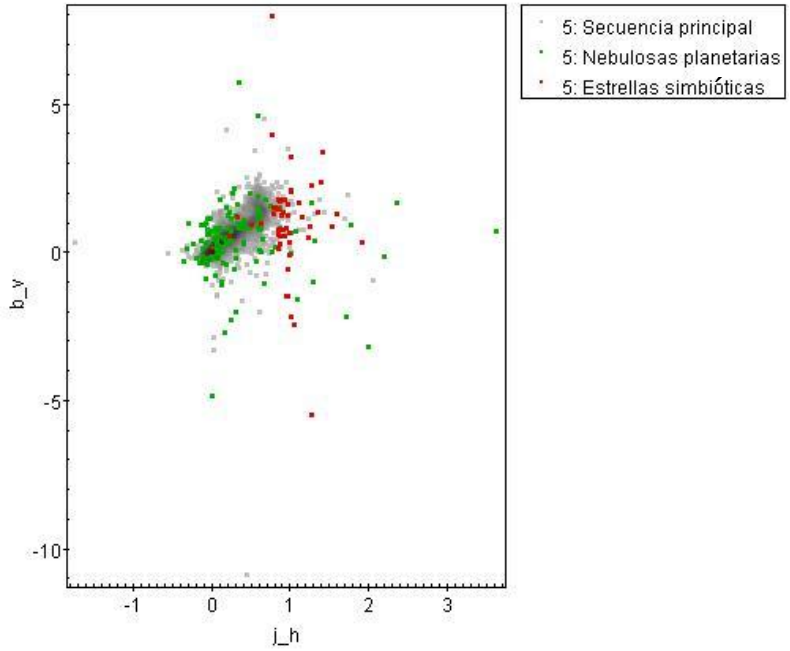


Figura 4.2 Diagrama color-color (j_h, b_v) (Fuente: elaboración propia)

Otro diagrama color-color generado fue para los valores de j_h en X y g_rp en Y. el diagrama se muestra a continuación en la Figura 4.3. Asi mismo, en este diagrama se aprecia la agrupación de objetos con los de la misma clase y que presentan una separabilidad.

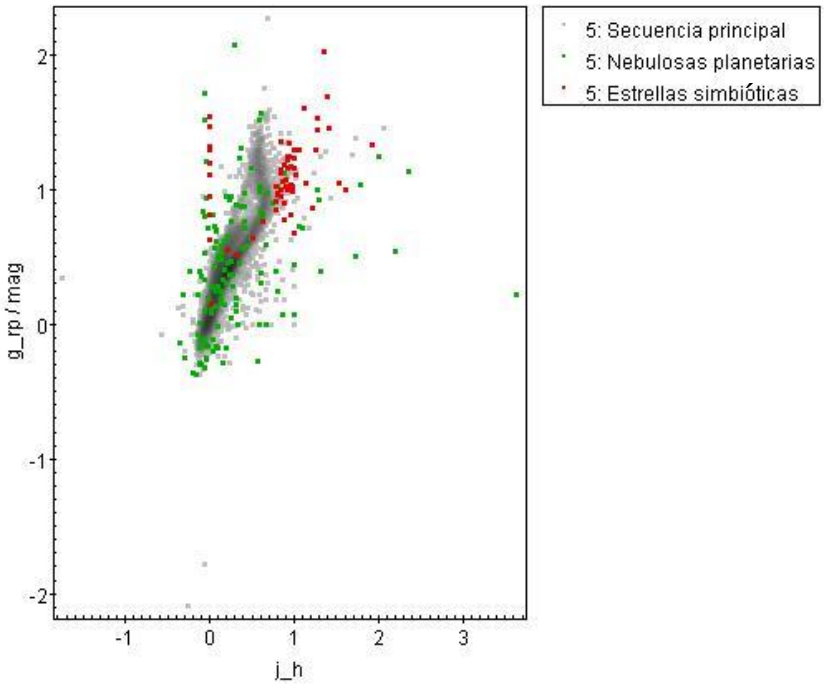


Figura 4.3 Diagrama color-color (j_h, g_rp) (Fuente: elaboración propia)

Otro de los de los diagramas generados para comprobar la separabilidad de las clases fue color-color-color. Los valores utilizados b_v para X, j_h para Y y g_rp en Z. El diagrama resultante se muestra a continuación en la Figura 4.4. Al analizar este diagrama generado que es una fusión de los diagramas de la Figura 4.3 y la Figura 4.2 se aprecia que aunque la separabilidad con los colores seleccionados no es buena para diagramas con dos dimensiones, al añadir una nueva dimensión la separabilidad mejora.

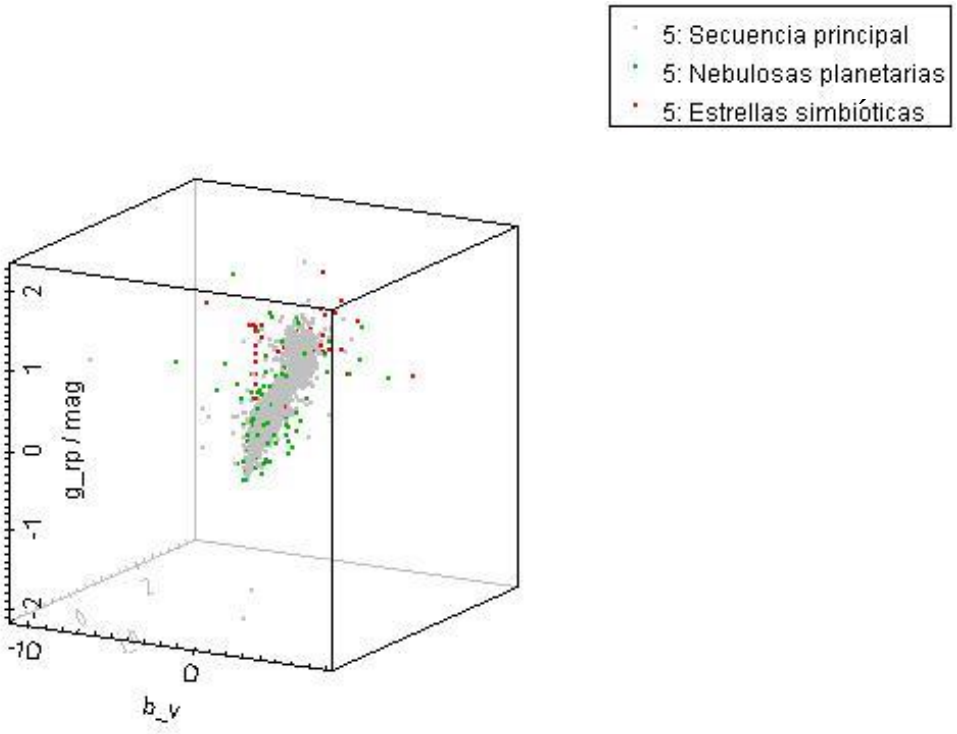


Figura 4.4 Diagrama color-color-color (b_v, j_h, g_rp) (Fuente: elaboración propia)

Con el fin de poder comparar la eficacia de los algoritmos utilizados para clasificar el set de datos de validación, se tomó en cuenta la precisión mostrada en la matriz de confusión. La matriz de confusión se explica a continuación en la Tabla 4.1. Dentro de esta tabla se usa VN como abreviatura de verdaderos negativos, FP de falsos positivos, FN de falsos negativos y VP de verdaderos positivos.

Tabla 4.1 Matriz de confusión explicada (fuente: elaboración propia)

Negativo Predicho	Positivo Predicho	
VN	FP	Negativo Real
FN	VP	Positivo Real

A partir de la matriz de confusión es posible obtener información acerca de un clasificador. La información que es posible obtener se explicada a continuación:

La exactitud, la cual se refiere al número de predicciones correctas del clasificador, se puede obtener con la Ecuación 5 que se muestra a continuación.

$$\text{Exactitud} = (VP+VN)/(VN+FP+FN+VP) \quad (5)$$

Ecuación 5 Ecuación para calcular la exactitud

La tasa de error la cual refiere al número de predicciones incorrectas realizadas por el clasificador, se puede obtener con la Ecuación 6 que se muestra a continuación.

$$\text{Tasa de error} = (FN+FP)/ (VN+FP+FN+VP) \quad (6)$$

Ecuación 6 Ecuación para calcular la tasa de error

La sensibilidad la cual refiere al número de predicciones positivas correctamente identificadas por el clasificador, se puede obtener con la Ecuación 7 que se muestra a continuación.

$$\text{Sensibilidad} = VP/(FN+VP) \quad (7)$$

Ecuación 7 Ecuación para calcular la sensibilidad

La especificidad la cual refiere al número de predicciones negativas correctamente identificadas por el clasificador, se puede obtener con la Ecuación 8 que se muestra a continuación.

$$\text{Especificidad} = VN/(VN+FP) \quad (8)$$

Ecuación 8 Ecuación para calcular la especificidad

La precisión la cual refiere al número de predicciones positivas clasificadas correctamente por el clasificador, se puede obtener con la Ecuación 9 que se muestra a continuación.

$$\text{Precisión} = VP / (VP+FP) \quad (9)$$

Ecuación 9 Ecuación para calcular la precisión

El valor de predicción negativa la cual refiere al número de predicciones negativas clasificadas correctamente por el clasificador, se puede obtener con la Ecuación 10 que se muestra a continuación.

$$\text{Valor de predicción negativa} = VN /(VN+FN) \quad (10)$$

Ecuación 10 Ecuación para calcular el valor de predicción negativa

Los resultados de evaluar los distintos algoritmos pueden considerarse, en forma general, como muy positivos, ya que las estrellas simbióticas son un tipo de objeto difícil de identificar. Además que de la manera tradicional de identificar este tipo de objetos es hacerlo sobre los espectros de los objetos, por lo tanto, el realizar la búsqueda con valores fotométricos y que se logre separar las estrellas no simbióticas de las estrellas simbióticas a través de los algoritmos de clasificación, se puede afirmar que implantación de estos algoritmos ha sido provechosa.

En cuanto la evaluación de los resultados arrojados por las diferentes implementaciones, el algoritmo Random Forest mostró una mejor clasificación sobre las redes neuronales. A continuación en la Tabla 4.2 se despliega la matriz de confusión del algoritmo Random Forest. Los resultados de las distintas técnicas fue una completa separación de las estrellas no simbióticas de las simbióticas, por lo que la matriz no refleja falsos positivos. Sin embargo las técnicas presentan falsos negativos, donde se clasificaron estrellas simbióticas como estrellas no simbióticas.

En la matriz de confusión se representa la predicción que hizo el algoritmo Random Forest, donde a las estrellas simbióticas se les etiqueta con un 1 y a las estrellas no simbióticas con un 0. La primera fila en la matriz corresponde a el número de estrellas no simbióticas y la segunda se refiere a el número de estrellas simbióticas reales. Así mismo, La primera columna representa las estrellas que fueron clasificadas como no simbióticas por el algoritmo mientras que la segunda columna corresponde a el número de estrellas que fueron clasificadas como simbióticas.

Lo que significa que el algoritmo clasificó 148 estrellas como no simbióticas, de un conjunto de 147 reales, y clasificó 36 como simbióticas de 37 estrellas simbióticas reales. A partir de la matriz de confusión se calculó que el algoritmo tiene un 99.45% de exactitud, una tasa de error de 0.54%, una sensibilidad de 97.29%, una especificidad de 100%, una precisión de 100% y un valor de predicción negativo de 99.32%.

Tabla 4.2 Matriz de confusión correspondiente al algoritmo Random Forest (Fuente: elaboración propia).

0	1	
147	0	0=0

1	36	1=1

En la Tabla 4.3 se muestra la matriz de confusión obtenida del algoritmo de automatización sobre redes neuronales. En ella se aprecia la predicción que hizo el algoritmo y se representa a las estrellas simbióticas con un 1 y las no simbióticas con un 0. En la primera fila se encuentra el número de estrellas no simbióticas y en la segunda, el número de estrellas simbióticas reales. En la primera columna representa las estrellas que fueron clasificadas como no simbióticas por el algoritmo, mientras que en la segunda columna se encuentra el número de estrellas que fueron clasificadas como simbióticas.

Lo que significa que el algoritmo clasificó 149 estrellas como no simbióticas, de un conjunto de 147 reales, y clasificó 35 como simbióticas de 37 estrellas simbióticas reales. A partir de la matriz de confusión se calculó que el algoritmo tiene un 98.91% de exactitud, tiene una tasa de error de 1.08%, una sensibilidad de 94.59%, una especificidad de 100%, una precisión de 100% y un valor de predicción negativo de 98.65%.

Tabla 4.3 Matriz de confusión correspondiente al algoritmo sobre redes neuronales (Fuente: elaboración propia).

0	1	
147	0	0=0
2	35	1=1

En la Tabla 4.4 se muestra la matriz de confusión resultante del algoritmo de árboles de decisiónl la cual representa la predicción que hizo el algoritmo de árboles de decisión. La interpretación de la tabla debe tomarse como en las anteriores, el número uno representa a las estrellas simbióticas y el número cero a las no simbióticas. También la primera fila representa el número de estrellas no simbióticas, mientras que la segunda fila corresponde a el número de estrellas simbióticas reales. La primera columna representa las estrellas que fueron clasificadas como no simbióticas por el algoritmo y la segunda columna el número de estrellas que fueron clasificadas como simbióticas.

La interpretación de estos resultados es que el algoritmo clasificó 145 estrellas como no simbióticas, de un conjunto de 139 reales, y clasificó 39 como simbióticas de 45 estrellas simbióticas reales. A partir de la matriz de confusión se calculó que el algoritmo tiene un 96.73% de exactitud, tiene una tasa de error de 3.26%, una sensibilidad de 86.66%, una especificidad de 100%, una precisión de 100% y un valor de predicción negativo de 95.86%.

Tabla 4.4 Matriz de confusión correspondiente al algoritmo de árboles de decisión (Fuente: elaboración propia).

0	1	
139	0	0=0
6	39	1=1

En la Tabla 4.5 se despliega la matriz de confusión obtenida a partir del algoritmo de máquinas de soporte vectorial. La lectura de la matriz debe hacerse igual que las anteriores. A diferencia de los otros algoritmos, éste clasificó 142 estrellas como no simbióticas, de un conjunto de 139 reales y clasificó 42 como simbióticas de 45 estrellas simbióticas reales. A partir de la matriz de confusión se calculó que el algoritmo tiene un 98.36% de exactitud, tiene una tasa de error de 1.63%, una sensibilidad de 93.33%, una especificidad de 100%, una precisión de 100% y un valor de predicción negativo de 97.88%.

Tabla 4.5 Matriz de confusión correspondiente al algoritmo de máquinas de soporte vectorial (Fuente: elaboración propia)

0	1	
139	0	0=0
3	42	1=1

Los resultados mostrados para los algoritmos representan al mejor sujeto generado durante las pruebas. Para las redes neuronales, se utilizó el algoritmo de automatización durante 72 horas. Los resultados obtenidos fueron la generación de 1128 modelos, por lo que se tiene que en promedio se han generado 15.6 modelos de redes neuronales por hora, lo que da un tiempo promedio de entrenamiento de 3.49 minutos en promedio para cada modelo. Por otra parte, para la generación de clasificadores donde se utilizó el algoritmo Random Forest, se procedió a la creación de 1128 sujetos, y el tiempo que se llevó la ejecución fue de 10 horas.

El criterio para la evaluación de cada sujeto fue la precisión obtenida al generar la matriz de confusión. La evaluación fue realizada dentro de la automatización y únicamente se tenía control sobre el mejor sujeto. Por este motivo, al momento de que se generaba un sujeto superior, se sustituía el mejor sujeto por el nuevo.

Con motivo de comparar mejor la matriz de confusión los distintos algoritmos se calculó el índice kappa para cada algoritmo con base al mejor sujeto presentado. El índice kappa consiste en un método estadístico para calcular la verdadera precisión a partir de una matriz de confusión, sin que este valor se vea afectado por la proporción de los datos por cada clase. El índice kappa asegura que el valor no se vea afectado cuando una clase es considerablemente mayor que otra, ya que los valores de precisión y exactitud se verían afectados. La comparación se muestra a continuación en la Tabla 4.6. De acuerdo a la escala propuesta por Landis y Koch (1977) el algoritmo Random Forest presenta un grado de concordancia casi perfecto.

Tabla 4.6 Valores de los índices kappa para los algoritmos (Fuente: elaboración propia)

	Redes neuronales	Random Forest	Árboles de decisión	Máquinas de soporte vectorial
Índice Kappa	0.9654	0.9829	.9075	.9548

Con el fin de ilustrar el funcionamiento del algoritmo de Random Forest (RF) se extrajó uno de los árboles estimadores perteneciente al mejor sujeto generado. Dicho árbol es desplegado en la Figura 4.5. Al observar el árbol se aprecia que la forma de clasificación que utilizó es similar a la creación de un diagrama color-magnitud, proceso usado en astrofísica para la caracterización. La magnitud usada fue Gabs y el color que usa el g_rp. La primera separación que realiza el árbol es sobre el color, acto seguido confirma las regiones a través de las magnitudes. Sólo en el par de nodos hoja más profundos se realizó una segunda comprobación por color. El presente árbol es uno de los distintos árboles generados, por lo que su configuración puede no ser correcta para todos los casos presentes en el set de validación.

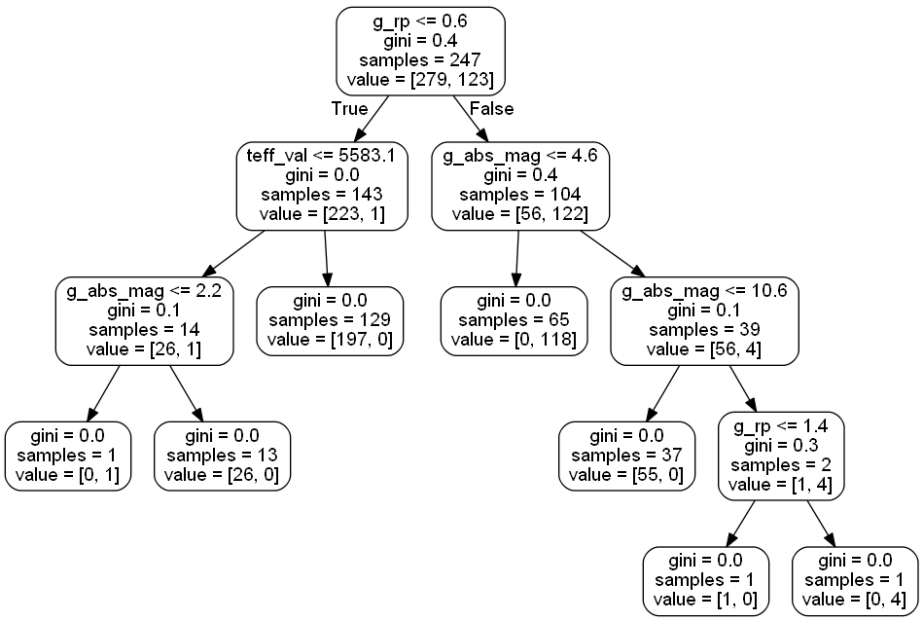


Figura 4.5 Árbol de decisión generado aleatoriamente por Random Forest (RF) (Fuente: elaboración propia).

El índice Gini permite seleccionar dos elemetos de una población, estos deben ser de la misma clase, donde la probabilidad de que esto suceda es uno, cuando la población es “pura”. Utiliza dos variables categóricas: “Success” o “Failure” y entre más grande sea el índice Gini, mayor es la homogeneidad de los datos. El cálculo de los subnodos usa la Ecuación 11 que consiste en suma de los cuadrados de probabilidad para success y failure según lo expuesto en Orellana Alvear (2018).

$$(p^2 + q^2) \tag{11}$$

Ecuación 11 Índice Gini.

Dónde

p: es success y

q: es failure.

Además del índice gini, en la Figura 4.5 se aprecian dos valores más que son: samples y value. Samples hace referencia al número total de muestras (objetos) con los que se realizó el árbol, mientras que la matriz que se encuentra en value representa como se distribuyen las muestras para cada clase de la clasificación. Por lo que los nodos hoja del árbol deberán tener solo un valor diferente a cero en algunas de las clases con las que se esté trabajando (Ceballos, 2019). Como en la presente investigación se trabajó con clasificación binaria el value de los nodos del árbol solo tiene [x,y].

Con el fin de ilustrar el mejor modelo generado de redes neuronales (ANN) se creó la representación perteneciente al mejor sujeto generado. Dicha red es desplegada en la Figura 4.6. Esta red fue generada automáticamente por el

algoritmo realizando variaciones aleatorias de los hiperparametros, el resultado fue que el mejor candidato tenía una topología de dos capas ocultas, cada una con 71 neuronas cada una.

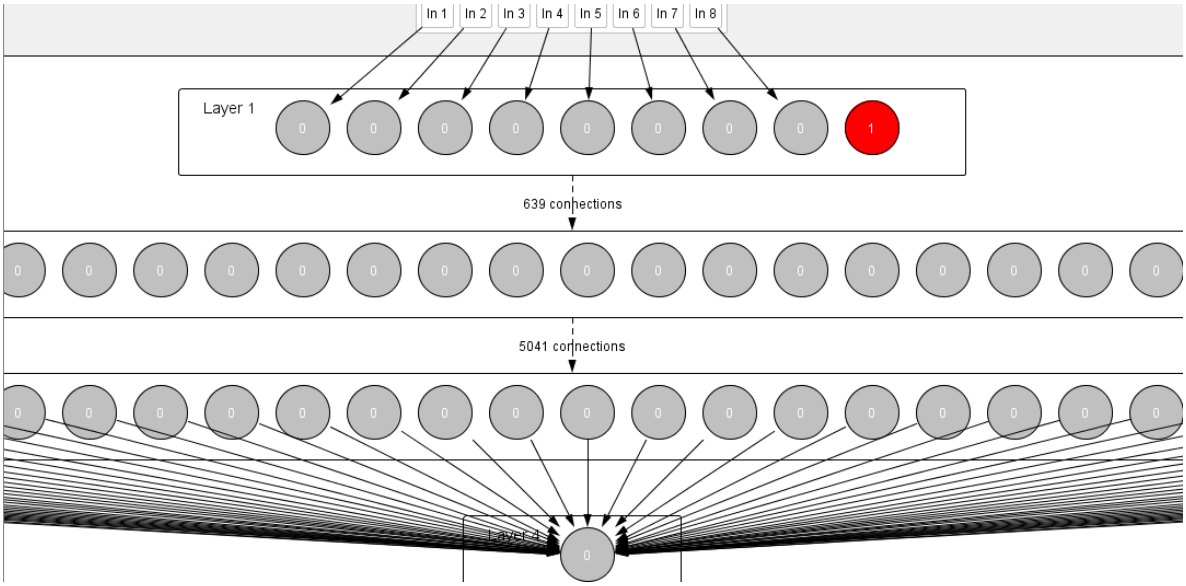


Figura 4.6 Red neuronal generada automáticamente (Fuente: elaboración propia)

Con el fin de ilustrar el funcionamiento del modelo generado por el algoritmo de árboles de decisión (DT) se extrajo el árbol estimador perteneciente al mejor sujeto generado. Dicho árbol es desplegado en la Figura 4.7. Al observar el árbol se aprecia que la forma de clasificación que utilizó es similar a la creación de un diagrama color-magnitud como primer paso de separación, proceso usado en astrofísica para la caracterización. El color que usa como primera comparación es h_k, y avanza a hacer la comparación con la magnitud g_abs_mag. La primera separación que realiza el árbol es sobre el color, acto seguido confirma las regiones a través de las magnitudes. Después continúa separando los objetos por el color g_rp, y termina por examinar la temperatura efectiva por medio del valor teff_val. El presente árbol es uno de los distintos árboles generados, por lo que su configuración puede no ser correcta para todos los casos presentes en el set de validación.

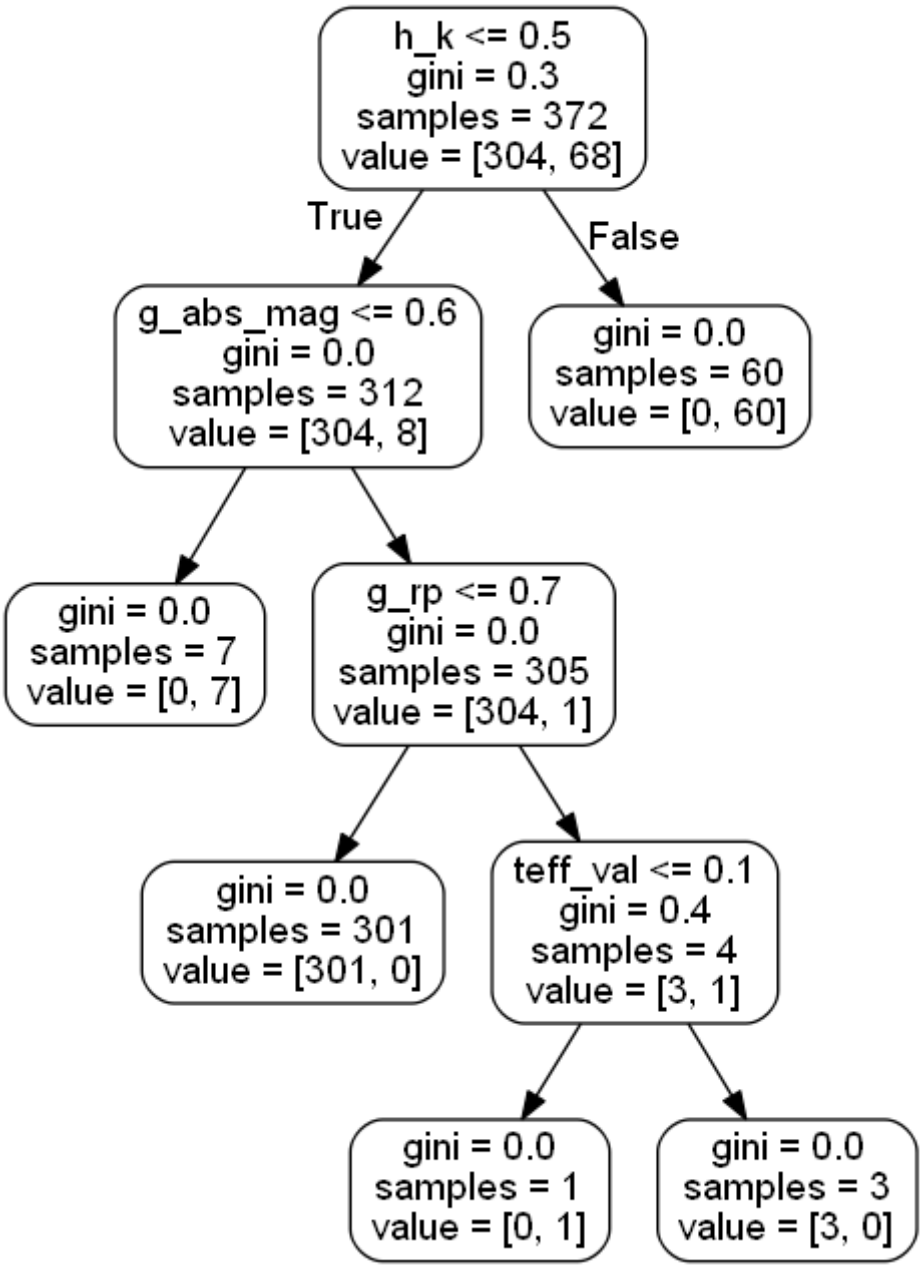


Figura 4.7 Modelo de árbol de decisión generado (Fuente: elaboración propia)

El modelo generado por el algoritmo de soporte vectorial, dado el conjunto de datos es de tipo caja negra, por lo que no se pudo extraer información para ilustrar su funcionamiento. Sin embargo el funcionamiento se puede representar de forma parcial como se aprecia en la Figura 4.8. Esta representación fue realizada con el vector de características completo, pero sólo se graficaron los primeros dos valores, los cuales son bp_rp y g_rp. El límite que encontró el algoritmo es el más óptimo y a pesar de que en la Figura 4.8 no se aprecia una buena separación, es importante recordar que es una imagen 2D perteneciente a un modelo con ocho dimensiones.

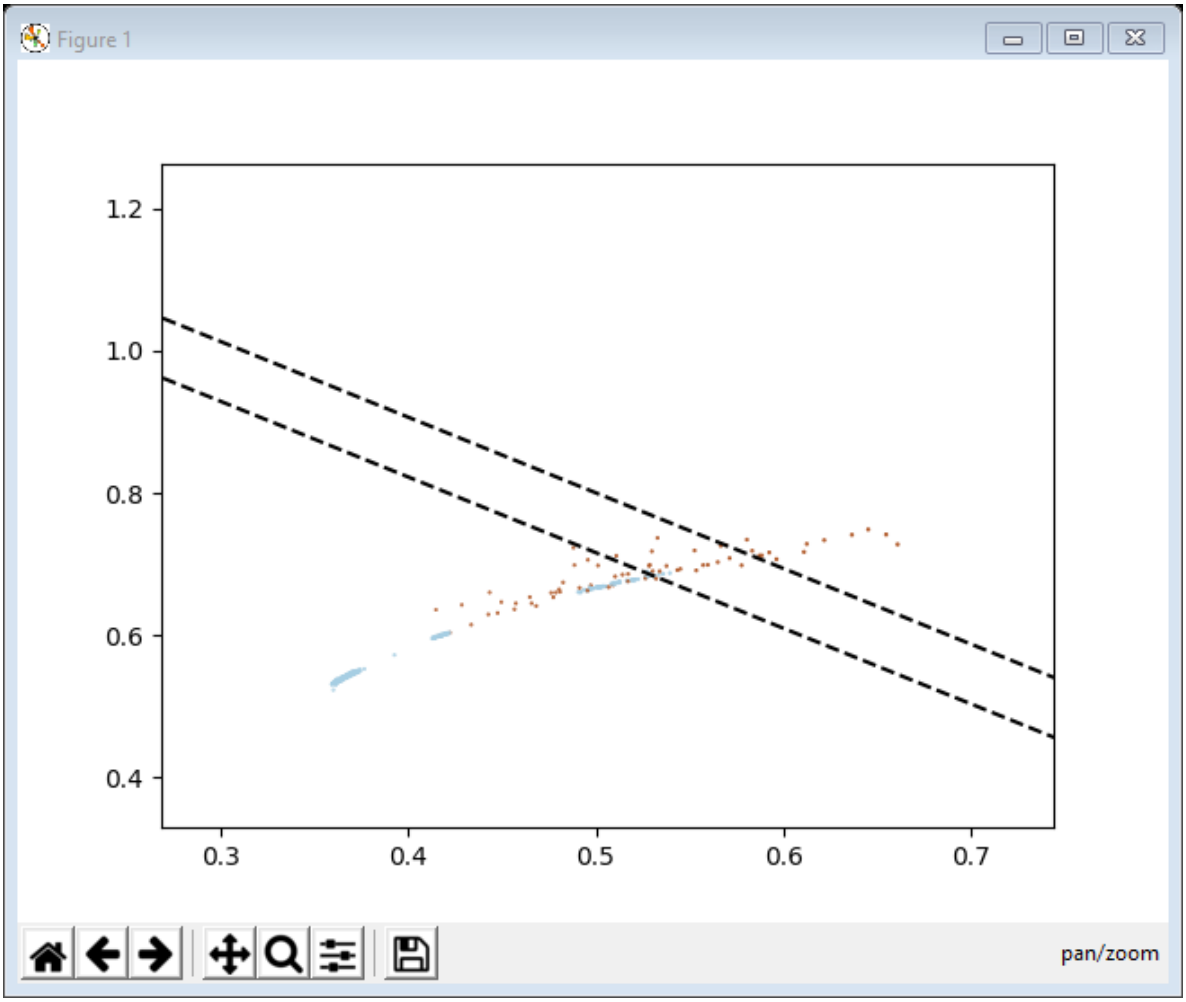


Figura 4.8 Modelo de máquinas de soporte vectorial (Fuente: elaboración propia)

Hasta este punto se ha explicado cómo se entrenaron, validaron y probaron los diferentes algoritmos de clasificación programados con la finalidad de realizar una búsqueda de estrellas simbióticas en la base de datos de la misión GAIA. Después de analizar y evaluar los resultados obtenidos, el siguiente paso fue alimentar las redes neuronales y los algoritmos de aprendizaje automatizado con una lista de objetos astronómicos considerados como candidatos para ser estrellas simbióticas (Tabla 4.7), esta información se obtuvo de el *catálogo* de Belczyński (2000).

Tabla 4.7 Listado de estrellas simbióticas candidatas (Fuente: Belczyński,2000)

Table 2. Suspected symbiotic stars

No.	Name	$\alpha(2000)$ $h\ m\ s$	$\delta(2000)$ ° ′ ″	l^{II} °	b^{II} °	V [mag]	K [mag]	IR	*IUE*	X	IP_{max} [eV]
s01⋆	RAW 1691	01 18 36.1	−72 42 24.0			15.3	12			–	13.6
s02⋆	[BE74] 583	05 26 54.0	−71 06 00.0			16.1					13.6
s03⋆	StHA 55	05 46 42.0	+06 43 48.0	199.34	−11.12	13.5					13.6
s04⋆	GH Gem	07 04 04.9	+12 02 12.0	203.57	+8.23	14.6	>9.7			–	
s05⋆	ZZ CMi	07 24 13.9	+08 53 51.7	208.64	+11.30	9.9	2.8	S			35.1
s06⋆	NQ Gem	07 31 54.5	+24 30 12.5	194.63	+19.35	7.9	3.0		+	–	54.4
s07⋆	WRAY 16−51	09 33 29.4	−46 34 49.0	271.35	+3.80		4.4				13.6
s08⋆	Hen 3−653	11 25 32.5	−59 56 31.9	292.36	+1.16	12.5	5.4	S		–	29.6
s09⋆	NSV 05572	12 21 52.5	−13 53 09.9	292.10	+48.36	15					13.6
s10⋆	AE Cir	14 44 52.0	−69 23 35.9	312.67	−8.69	14.1					54.4
s11⋆	V748 Cen	14 59 37.0	−33 25 23.9	331.51	+22.24	12.6	8.1			–	13.6
s12⋆	V345 Nor	16 06 44.3	−52 02 30.1	330.51	+0.05	11.4				–	13.6
s13⋆	V934 Her	17 06 34.5	+23 58 18.5	45.15	+32.99	7.8	3.3		+	+	77.5
s14⋆	Hen 3−1383	17 20 31.5	−33 09 55.7	353.53	+2.20	12.5				–	24.6
s15⋆	V503 Her	17 36 46.0	+23 18 18.0	47.00	+26.23	13.8				–	
s16⋆	WSTB 19W032	17 39 02.8	−28 56 35.0	359.24	+1.22	17.2					35.1
s17⋆	WRAY 16−294	17 39 13.9	−25 38 06.0	2.06	+2.94	15.5		S			35.1
s18⋆	AS 241	17 44 58.0	−38 18 12.9	351.92	−4.76	12.0	7.8	S		–	24.6
s19⋆	DT Ser	18 01 52.0	−01 26 06.0	25.92	+10.34	15.4					54.4
s20⋆	V618 Sgr	18 07 57.0	−36 29 35.9	355.77	−7.83	15.2					13.6
s21⋆	AS 280	18 09 52.9	−33 19 41.9	358.77	−6.69	13.2	>9.4	S			54.4
s22⋆	AS 288	18 12 48.0	−28 20 00.9	3.49	−4.87		8.4	D?			54.4
s23⋆	Hen 2−379	18 16 17.4	−27 04 32.9	4.97	−4.96	12.5	9.3				35.1
s24⋆	V335 Vul	19 23 14.2	+24 27 40.2	58.22	+4.40	11.8					13.6
s25⋆	V850 Aql	19 23 34.6	+00 38 03.0	37.18	−6.85		5.0	S			13.6
s26⋆	Hen 2−442	19 39 39.0	+26 30 42.0	61.80	+2.13	14	5.3				100
s27⋆	*IRAS* 19558+3333	19 57 48.4	+33 41 15.9	69.98	+2.38			D?			
s28⋆	V627 Cas	22 57 41.2	+58 49 14.9	108.66	−0.86	12.9	3.3	D			13.6

Al algoritmo RF generado y entrenado con los sets de entrenamiento y prueba, se ingresaron los 28 objetos candidatos de ser estrellas simbióticas pertenecientes al *catálogo* de Belczyński (2000). Los resultados fueron tabulados y posteriormente, se tomó el segundo mejor algoritmo que fue el de redes neuronales. A este se le introdujo el mismo set de objetos candidatos que se usó con la técnica de Random Forest, y dichos resultados se compararon con los resultados previamente obtenidos. Los resultado de ambos algoritmos coincidieron por lo que se procedió a dar un formato entendible para una mejor comprensión.

Para poder representar los resultados obtenidos, se utilizan tres diferentes etiquetas, las cuales son identificación positiva, identificación negativa, y N/A. La etiqueta de identificación positiva es usada para aquellas estrellas que después de pasar por el modelo de aprendizaje automático, que obtuvo el mejor puntaje en el índice kappa, fueron señaladas como posibles estrellas simbióticas. La etiqueta de identificación negativa, son aquellas que de acuerdo a los datos recuperados de GAIA y SIMBAD, fueron catalogadas por el software como no simbióticas. Por último aquellas estrellas con la etiqueta N/A son las que no fueron capaz de analizarse por falta de información. La lista es presentada a continuación en la Tabla 4.8.

Tabla 4.8 Lista resultante de estrellas simbióticas (Fuente: elaboración propia)

Estrella	Identificación
[BE74] 583	Positiva
StHA 55	Positiva
GH Gem	Positiva
ZZ CMi	Positiva
NQ Gem	Positiva
WRAY 16-51	Positiva
Hen 3-653	Positiva
NSV 05572	Positiva
AE Cir	Positiva
V748 Cen	Positiva
V345 Nor	Positiva
V934 Her	Positiva
Hen 3-1383	Positiva
V503 Her	Positiva
WSTB 19W032	Positiva
WRAY 16-294	Positiva
AS 241	Positiva
V618 Sgr	Positiva
AS 288	Positiva
Hen 2-379	Positiva
V335 Vul	Positiva
V850 Aql	Positiva
V627 Cas	Positiva
RAW 1691	N/A
DT Ser	N/A
IRAS 19558+3333	N/A
Hen 2-442	Negativa

Con la evaluación realizada a través de técnicas de inteligencia artificial, se concluye que los algoritmos implementados arrojan resultados fiables, con un porcentaje alto de exactitud, donde el peor de los casos arroja un 96%, cómo lo es el de árboles de decisión, mientras que el algoritmo de random forest, da un porcentaje de exactitud mayor al 99%. Con el desarrollo de estos programas, se cumple en objetivo de realizar la búsqueda de estrellas simbióticas en la misión de GAIA (DR2) a través de técnicas de aprendizaje automatizado.

Adicionalmente, se realizó otra descarga de 20, 000 espectros estelares, (número máximo permitido por GAIA), para ser revisadas a través de el software. Los resultados de esta última prueba fueron negativos. La razón de no haber obtenido ninguna identificación positiva se atribuye a que las estrellas simbióticas son objetos difíciles de encontrar y además de que su población es pequeña en comparación con las estrellas normales. Para darnos una idea, se han descubierto cerca de 200 objetos de este tipo en nuestra galaxia, la cual contiene alrededor de 10^{11} estrellas, sin embargo se estima que su número es muy superior. El hecho de que la mayoría se encuentren cerca del plano de la galaxia hace que la extinción interestelar sea muy considerable. Esto, aunado a los extensos periodos orbitales determinados en estos sistemas (de hasta varias décadas), complica más su detección y confirmación.

Capítulo 5
5.1 Conclusiones

La presente tesis tuvo como objetivo demostrar que es posible realizar una búsqueda y clasificación, por medio de alguna técnica de aprendizaje automático, de estrellas simbióticas en la base de datos de la misión GAIA. Como producto final, lograr obtener una lista de estrellas que presenten coincidencias con el perfil de una estrella candidata a estrella simbiótica.

Para realizar esto, primeramente se estudió el contenido y funcionamiento de la base de datos de GAIA. Se observó que los datos contenidos en la misión eran valiosos para realizar una clasificación, pero también fue notoria la necesidad de ampliar dicha información con otra base de datos, para lo cual, se utilizó la información de SIMBAD.

Una vez que se fusionó la información de ambas bases de datos, se observó como las estrellas simbióticas se agrupaban en un área sobre la secuencia principal del diagrama H-R, por lo tanto, se procedió a la elaboración de otros diagramas en los que también aparecieran las estrellas separadas de otros objetos, pero agrupadas entre sí. Ante el hecho de que las estrellas simbióticas presentan una buena separabilidad con los valores comparados, se concluyó que dichos valores podrían constituir un perfil de estrella candidata a ser simbiótica y se agregaron a un vector de características.

Identificado el perfil de los objetos a clasificar, se inició una búsqueda para encontrar que técnica era las más adecuada para obtener la lista de objetos deseados. El proceso que se siguió fue identificar en artículos científicos las técnicas más utilizadas para la clasificación y localizar aquellas que sobresalían en problemas similares al que se presenta en la investigación. Se elaboró una lista de cuatro técnicas, las cuales cumplían los requerimientos que se necesitaban y habían probado ser de utilidad para otros autores. Sin embargo, no fue posible identificar cual era la más adecuada, por lo que se decidió, tomar las cuatro y compararlas. El siguiente reto fue elegir algún lenguaje de programación para implementar estas técnicas e iniciar a hacer pruebas. El lenguaje seleccionado fue python ya que presentaba un mayor apoyo de librerías que facilitaran la codificación las técnicas seleccionadas.

Una vez que se programaron las distintas técnicas de aprendizaje automático, se obtuvieron porcentajes de exactitud, tasa de error y precisión entre otros; estos resultados se presentaron a través de una matriz de confusión. Los diferentes algoritmos al trabajar con estos datos, obtuvieron un porcentaje de exactitud sobre el 96%, donde random forest tiene un 99.45%, las redes neuronales un 98.91%, el árbol de desición un 96.73% y la máquina de soporte vectorial obtuvo un 98.36%.

En la búsqueda de una forma de poder comparar los resultados, se encontró el índice kappa, el cual consiste en eliminar falsos resultados por la mala proporción de alguna clase con respecto a otra en algún set de datos. Incluso el índice kappa proporcionaba una escala para saber qué tan buena, o mala había sido una clasificación. Por lo que se incorporó al proyecto, y sorpresivamente se obtuvo que el algoritmo con mayor índice kappa es el Random Forest con un 0.9829, seguido de las redes neuronales con un 0.9654 de índice.

Una vez se obtuvo un algoritmo ganador de la comparativa, se le enfrentó a datos no etiquetados cómo objetos simbióticos, para comprobar la eficacia de este. Se alimentó con otras 20,000 estrellas, y 28 objetos sospechosos de ser estrellas simbióticas.

Para verificar que tan consistentes eran los resultados del algoritmo de random forest, se utilizó también la red neuronal con los mismos datos descargados. ANN presentó resultados iguales a los que se obtuvieron con RF. Los dos algoritmos con una confiabilidad de 99% coincidieron en los resultados, por lo que se realizó un etiquetado para su mayor comprensión y así presentar una lista de objetos candidatos a ser estrellas simbióticas.

La hipótesis de la que se partió al inicio de este proyecto, plantea que a través de técnicas de aprendizaje automatizado, se pueden encontrar estrellas simbióticas en las bases de datos de la misión GAIA (DR2). Después de los resultados obtenidos se concluye que la hipótesis planteada en un inicio se cumple.

5.2 Trabajo futuro

En un trabajo tan ambicioso como el presente, es normal que durante su desarrollo surjan nuevas líneas de investigación, o existan algunas las cuales no es posible abordar. Esta tesis no es la excepción por lo que se presentan las siguientes fases para continuar ampliando la presente, algunas de las cuales han surgido con base a preguntas sobre qué pasaría si se cambian algunos valores, como afecta a la confiabilidad si se introduce un objeto en vez de otro a los sets de datos o se varían las clases a identificar.

Como primer fase se encuentra realizar búsquedas más extensas y utilizar los algoritmos RF y ANN para el análisis de los objetos, con el fin de detectar nuevos candidatos a estrellas simbióticas. Como parte del trabajo futuro también se encuentra aumentar los sets de datos con nuevas clases de objetos para aumentar la fiabilidad de la clasificación binaria. Además puede ser interesante analizar el comportamiento de los algoritmos clasificadores así como su desempeño al cambiar la clasificación binaria por una multiclasificación al aumentar el número de objetos a clasificar como podrían ser nebulosas planetarias, regiones de formación estelar, novas, supernovas, entre otras.

5.3 Recomendaciones

El objetivo de realizar investigación consiste en realizar una mejora continua, y no repetir trabajo ya realizado, así como utilizar el estado del arte para avanzar más rápido, recogiendo las recomendaciones realizadas por otros investigadores, así como saber que se está haciendo en algún área, o saber a qué problemas se enfrentaron los autores. Por tanto, se recomienda a cualquier persona interesada en el proyecto, agregar nuevos objetos a clasificar y modificar el vector de características en caso de ser necesario. Realizar el análisis de los nuevos objetos a incluir, tanto a través de información obtenida de GAIA, como de publicaciones en el área de astrofísica. Incluir nuevos clasificadores para continuar ampliando la comparativa, y verificar si RF sigue obteniendo la precisión mas alta. Por último se recomienda que al momento de buscar los valores frontera para un nuevo set de datos, siempre se realice combinando los valores a evaluar, con el fin de tener límites reales.

Otras recomendaciones que no están ligadas al tema de investigación, es que el algoritmo RF representa una forma más sencilla de llegar a una clasificación, tanto en el tiempo que se ocupa invertir entrenando, como al momento de realizar el análisis del modelo generado. Se recomienda RF para casos en los que el tiempo sea limitado y se necesiten resultados de forma rápida. Así mismo se invita a utilizar ANN que aunque requieren un mayor esfuerzo, con el adecuado conjunto de hiperparámetros y entrenamiento puede superar a random forest.

Glosario

Catálogo: también conocido como catálogo de estrellas o catálogo astronómico consiste en una lista o tabulación de objetos, los cuales fueron agrupados ya que comparten el mismo tipo, morfología, origen o método de descubrimiento.

Fuzzy minimun whitin-class support vector machine ó FMWSVM: es un algoritmo derivado de las máquinas de soporte vectorial pero en este caso se ha añadido un miembro difuso con el objetivo de que distintas entradas puedan hacer distintas contribuciones al aprendizaje, además de que reduce el ruido.

K-means: es un algoritmo de clasificación no supervisada que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo. Se suele usar la distancia cuadrática (Palma, 2008).

LPP: es un algoritmo lineal para la reducción de datos (Nigoyi y he, 2004). Su principal característica es que preserva la estructura local de los datos. Por lo que la relación entre dos datos de manera global y local se mantendrá, siempre y cuando pertenezcan a la misma clase, pero estarán alejados si no.

MCC: es un algoritmo utilizado para aprendizaje, básicamente es una correlación de coeficientes entre las clasificaciones binarias observadas y pronosticadas; devuelve un valor entre −1 y +1. Un coeficiente de +1 representa una predicción perfecta, 0 no mejor que la predicción aleatoria y −1 indica un desacuerdo total entre la predicción y la observación.

Paralaje: es la diferencia entre la aparente posición de un objeto debido a la variación del punto de observación. Suele expresarse en ángulos, y su utilidad es determinar la posición de una estrella con respecto de la tierra.

PCA: principal component analisis o por su traducción al español análisis de componentes principales, es un algoritmo de reducción de dimensionalidad de datos(Alquicira, 2016) reteniendo la mayoría de la variación de ellos en vectores que son llamados componentes principales. Los componentes principales son combinaciones no lineales que no guardan relación entre sí. Además los componentes principales maximizan la varianza de las observaciones. Este algoritmo es ampliamente utilizado para identificar patrones en conjuntos de datos con un número dimensiones considerables.

Sparse representations: es un algoritmo ampliamente utilizado para el reconocimiento de objetos en imágenes.

WebScrapping: técnica que consiste en extraer la información del código HTML de un documento Web.

White dwarf ó WD: son las siglas para White dwarf o por su traducción al español enana blanca (EcuRed, s.f.). Las enanas blancas son los remanentes de

una estrella de masa menor a 9-10 masas solares, la cual ha agotado su combustible nuclear.

WPCA: es una variación del algoritmo base de PCA donde los datos se ponderan, dando como resultado una mejora significativa al algoritmo.

Referencias

Akras, S., Guzman-Ramirez, L., Leal-Ferreira, M. L., & Ramos-Larios, G. (2019). A Census of Symbiotic Stars in the 2MASS, WISE, and Gaia Surveys. The Astrophysical Journal Supplement Series, 240(2), 21

Akras, S., Leal-Ferreira, M. L., Guzman-Ramirez, L., & Ramos-Larios, G. (2018). A machine learning approach for identification and classification of symbiotic stars using 2MASS and WISE. Monthly Notices of the Royal Astronomical Society, 483(4), 5077-5104.

Aladin (s.f.). V* CI Cyg. [Figura]. Recuperado de http://aladin.u-strasbg.fr/AladinLite/?target=V*%20CI%20Cyg&fov=0.034508835357033356&survey=P%2fDSS2%2fcolor

Alquicira, J. (2016). Análisis de componentes principales (PCA). Recuperado el 9 de diciembre de 2017, de http://conogasi.org/articulos/analisis-de-componentes-principales-pca/

Allende Prieto, C. (2003). An automated system to classify stellar spectra - I. Monthly Notices of the Royal Astronomical Society, 339(4), 1111-1116. doi:10.1046/j.1365-8711.2003.06260.x

Astromia (s.f.). Binarias Eclipsantes. Recuperado el 27 de Septiembre de 2018, de https://www.astromia.com/fotouniverso/binariaseclipsantes.htm

Astromia (s.f.). La espectroscopia en la astronomia. Recuperado el 30 de Septiembre de 2018, de https://www.astromia.com/historia/espectrohistoria.htm

Auria, L., Moro, R. A., Support Vector Machines (SVM) as a Technique for Solvency Analysis (August 1, 2008). DIW Berlin Discussion Paper No. 811.

Barranco Fragoso, R. (2012). ¿Qué es Big Data? Recuperado el 13 de Noviembre de 2017, de IBM DeveloperWorks: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html

Barrera, V. (2016). ¿Qué es Analítica de Datos, data & analytics o Data Analytics o simplemente Analytics? Recuperado el 14 de Noviembre de 2017, de Instituto Internacional de Ciencia de Datos: http://www.i2ds.org/algunas-consideraciones-sobre-la-analitica-de-datos-o-data-analytics

Belczyński, K., Mikołajewska, J., Munari, U., Ivison, R. J., & Friedjung, M. (2000). A catalogue of symbiotic stars. Astronomy and Astrophysics Supplement Series, 146(3), 407-435.

Borracci, R. A., & Arribalzaga, E. B. (2005). Aplicación de análisis de conglomerados y redes neuronales artificiales para la clasificación y

selección de candidatos a residencias médicas. Educación Médica, 8(1), 22-30.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. International Group, 432.

Ceballos, F., (2019). Scikit-Learn Decision Trees Explained Training, Visualizing, and Making Predictions with Decision Trees, Towards Data Science, Recuperado el 12 de Abril de 2019, de https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d

Colas F., Brazdil P. (2006) Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In: Bramer M. (eds) Artificial Intelligence in Theory and Practice. IFIP AI 2006. IFIP International Federation for Information Processing, vol 217. Springer, Boston, MA.

Collazo, Rodrigo Abrunhosa, Pessôa, Leonardo Antonio Monteiro, Bahiense, Laura, Pereira, Basílio de Bragança, Reis, Amália Faria dos, & Silva, Nelson Souza e. (2016). A COMPARATIVE STUDY BETWEEN ARTIFICIAL NEURAL NETWORK AND SUPPORT VECTOR MACHINE FOR ACUTE CORONARY SYNDROME PROGNOSIS. Pesquisa Operacional, 36(2), 321-343. https://dx.doi.org/10.1590/0101-7438.2016.036.02.0321

Colomer Sanmaríın, F. (1998). EL DESCUBRIMIENTO DE OBJETOS. Anuario Astronómico del Observatorio de Madrid, 289-303

Díaz-Hernández, R., Peregrina-Barreto, H., Altamirano-Robles, L., González-Bernal, J. A., & Ortiz-Esquivel, A. E. (2014). Automatic stellar spectral classification via sparse representations and dictionary learning. *Experimental Astronomy, 38*(1), 193-211. 10.1007/s10686-014-9413-2

Delchambre, L. (2017). Determination of astrophysical parameters of quasars within the Gaia mission. *Monthly Notices of the Royal Astronomical Society*, *473*(2), 1785-1800.

EcuRed (s.f.). Enana blanca. Astrometría. Recuperado el 2 de Diciembre de 2017, de https://www.ecured.cu/Enana_blanca

EcuRed (s.f.). Proyecto de Constitución de la República de Cuba. Astrometría. Recuperado el 29 de Noviembre de 2017, de https://www.ecured.cu/Astrometr%C3%ADa

Echevarría Román, J. (2009). Estrellas binarias. Revista ciencia (60), pp 34.

ESA (s.f.) MISSION STATUS NUMBERS. Recuperado el 01 de junio de 2019, de https://www.cosmos.esa.int/web/gaia/mission-numbers

Gaia Collaboration, Babusiaux, C., van Leeuwen, F., Barstow, M. A., Jordi, C., Vallenari, A., Bossini, D., ... & Prusti, T. (2018). Gaia Data Release 2-Observational Hertzsprung-Russell diagrams. Astronomy & astrophysics, 616, A10.

Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, A. Vallenari, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans and et al. (2016b) c. A&A 595, p A1.

González Marmol, Juan Miguel (2007). Astro y Ciencia. Formación y evolución de una estrella, Recuperado el 25 de Septiembre de 2018, de http://www.astroyciencia.com/2007/12/21/formacion-y-evolucion-de-una-estrella/

Hillier, F. S., Lieberman, G. J., & Osuna, M. A. G. (1997). Introducción a la Investigación de Operaciones (Vol. 1). McGraw-Hill.

ILCE (s.f.) CÓMO FUNCIONA EL TELESCOPIO, Instituto Latinoamericano de la Comunicación Educativa. Recuperado el 10 de junio de 2019 de http://bibliotecadigital.ilce.edu.mx/sites/ciencia/volumen2/ciencia3/057/htm/sec_6.htm

Instituto de Astronomía. (s.f.). Estrellas binarias, Universidad Nacional Autonoma de México. Recuperado el 14 de Noviembre de 2017, de http://www.astroscu.unam.mx/~wlee/OC/SSAAE/AAE/Sistemas%20Binarios/Estrellas%20binarias.html

Instituto Geografico Nacional (s.f.). Formación de estrellas. Recuperado el 20 de Agosto de 2018, de http://astronomia.ign.es/formacion-de-estrellas

IntelDig (s.f.). Técnicas de Análisis Big Data. Recuperado el 15 de Diciembre de 2017, de https://www.tecnologias-informacion.com/tecnicasbigdata.html#

Karttunen, H., Pountanen, M., Donner, KJ., (1996) Fundamental Astronomy, Tercera Edición, Springer- Verlag Berlin.

Kheirdastan, S., & Bazarghan, M. (2016). SDSS-DR12 bulk stellar spectral classification: Artificial neural networks approach. Astrophysics and Space Science, 361(9), 1-8. doi:10.1007/s10509-016-2880-3

Keller, C. A., & Evans, M. J. (2019). Application of Random Forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. Geoscientific Model Development, 12(3), 1209-1225.

Kenyon, S. J., & Fernandez-Castro, T. (Abril de 1987). The cool components of symbiotic stars. I. Optical spectral types, 93, 938-949. Astronomical Journal. Recuperado el 15 de Noviembre de 2017, de http://adsbit.harvard.edu//full/1987AJ.....93..938K/0000938.000.html

Lacy, M., Riley, J. M., Waldram, E. M., McMahon, R. G., & Warner, P. J. (1995). A radio-optical survey of the North Ecliptic CAP. Monthly Notices of the Royal Astronomical Society, 276, 614-626.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

Liu, Z., Song, W., Zhang, J., & Zhao, W. (2017). Classification of stellar spectra with fuzzy minimum within-class support vector machine. *Journal of Astrophysics and Astronomy, 38*(2), 1. doi:10.1007/s12036-017-9441-1

Luna, G. J. M., Sokoloski, J. L., Mukai, K., & Nelson, T. (2012). Symbiotic stars in X rays. *ArXiv Preprint ArXiv*, 1–17. https://doi.org/10.1051/0004-6361/201220792

Malatesta, K. (2010). CI Cygni, Recuperado el 15 de Junio de 2019, de https://www.aavso.org/vsots_cicyg

Martinez Troya, Daniel (2008). La evolución estelar. Ed. Libros en red.

Miszalski, B., Mikołajewska, J., & Udalski, A. (2013). Symbiotic stars and other Hα emission-line stars towards the Galactic bulge. Monthly Notices of the Royal Astronomical Society, 432(4), 3186-3217.

Minguillón i Alfonso, J., & Pujol Capdevila, J. (2002). Árboles de decisión.

NASA (s.f.) Líneas de Absorción y de Emisión. Recuperado el 4 de Diciembre de 2017, de https://www.mdscc.nasa.gov/index.php?Section=Lineas_de_Absorcion_y_de_Emision

Niño, M. (7 de Marzo de 2016). Entendiendo la diferencia entre analítica de datos descriptiva, predictiva y prescriptiva. Recuperado el 12 de Noviembre de 2017, de http://www.mikelnino.com/2016/03/diferencia-analitica-datos-descriptiva-predictiva-prescriptiva.html

Niyogi, X., & He, X. (2004). Locality preserving projections. In Neural information processing systems (Vol. 16, No. 2004).

Obiols, A. (20 de Mayo de 2015). ¿Qué es un Data Scientist? Recuperado el 15 de Noviembre de 2017, de InLab FIB: https://inlab.fib.upc.edu/es/blog/que-es-un-data-scientist

Olabe, X. B. (1998). Redes neuronales artificiales y sus aplicaciones. Publicaciones de la Escuela de Ingenieros 101pp.

Orellana Alvear, J. (2018) Árboles de decisión y Random Forest. Recuperado el 01 de Abril de 2019, de Bookdown: https://bookdown.org/content/2031/

Palma M. J.T. (2008), Inteligencia Artificial: Tecnicas, métodos y aplicaciones, McGrawHill, pp: 703-706.

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. Revista SEDEN.

PennState (s.f.). Energy levels of electrons in Bohr model and how those correspond to the wavelengths of an absorption or emission line in an object's spectrum. [Figura]. Recuperado de: https://astro.psu.edu/public-outreach/fireworks-masks-1/absorption-and-emission-spectra

Portalastronomico (2018). R Aquarii Simbiótica. [Figura]. Recuperado de https://www.portalastronomico.com/r-aquarii-simbiotica/

Powell, R. (2007). The Hertzsprung Russell Diagram. [Figura]. Recuperado de http://www.atlasoftheuniverse.com/

Simbad (s.f). The SIMBAD astronomical database. [Figura]. Recuperado de http://simbad.u-strasbg.fr/simbad/sim-id?Ident=Beta+Persei&submit=submit+id

Soares, S. (2012). *Not Your Type? Big Data Matchmaker On Five Data Types You Need To Explore Today*. Recuperado el 13 de Noviembre de 2017, de Dataversity: http://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/

Rouse, M. (2017). Big Data, SearchDataCenter TechTarget. Recuperado el 15 de Noviembre de 2017, de: http://searchdatacenter.techtarget.com/es/definicion/Big-data

Russo, A. E. (2018). Estrellas binarias. Recuperado el 2 de Junio de 2018, de http://estrellasbinarias.com.ar/

Schmidt, J. (2018). Symbiotic R Aquarii [Figura]. Recuperado el 2 de Junio de 2018, de https://apod.nasa.gov/apod/ap180711.html.

Suárez, Liyuan (2013). ASTROFISICA: Introducción, Historia, Teorías físicas implicadas. Recuperado el 5 de noviembre de 2017, de https://www.canaldeciencias.com/2013/02/08/astrofisica-introducci%C3%B3n-historia-y-teor%C3%ADas/

Tohmé J., Tomas (2002). Astronomía Online. La evolución estelar y el diagrama Hertzsprung-Russell. Recuperado el 27 de Septiembre de 2018, de https://www.astronomiaonline.com/2002/12/evolucion-estelar-diagrama-hertzsprung-russell/

Tohmé Lopez, C. (2014). Cátedra de Cultura Científica. De la paralaje. Recuperado el 28 de Septiembre de 2018, de https://culturacientifica.com/2014/11/18/de-la-paralaje/

Villavicencio, J. (2010). Introducción a series de tiempo. Obtenido de Sitio Web del Instituto de Estadısticas de Puerto Rico: http://www. estadisticas. gobierno. pr/iepr/LinkClick. aspx.

Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., ... & Monier, R. (2000). The SIMBAD astronomical database-The CDS reference database for astronomical objects. Astronomy and Astrophysics Supplement Series, 143(1), 9-22.

Zamorano, J., (s.f.). Técnicas experimentales en Astrofísica, Universidad Complutence Madrid. Recuperado el 02 de Febrero de 2019 de https://webs.ucm.es/info/Astrof/users/jaz/TEA/tea_04.pdf

Zhong-bao, L. (2016). Stellar spectral classification with locality preserving projections and support vector machine. Journal of Astrophysics and Astronomy, 37(2), 1-7. doi:10.1007/s12036-016-9387-8

The Two Micron All Sky Survey (2MASS), M.F. Skrutskie, R.M. Cutri, R. Stiening, M.D. Weinberg, S. Schneider, J.M. Carpenter, C. Beichman, R. Capps, T. Chester, J. Elias, J. Huchra, J. Liebert, C. Lonsdale, D.G. Monet, S. Price, P. Seitzer, T. Jarrett, J.D. Kirkpatrick, J. Gizis, E. Howard, T. Evans, J. Fowler, L. Fullmer, R. Hurt, R. Light, E.L. Kopan, K.A. Marsh, H.L. McCallon, R. Tam, S. Van Dyk, and S. Wheelock (2006), AJ, 131, 1163.