Vicerrectoría de Investigación y Postgrado

Dirección de Postgrados y Postítulos

Facultad de Ciencias

Departamento de Astronomía

Identificación y determinación de metalicidades de estrellas gigantes rojas por medio de técnicas de Machine Learning aplicado a la fotometría de banda ancha y angosta del relevamiento S-PLUS

Autora:
Francisca Molina Jorquera

Profesor Patrocinante:
Guillermo Damke

Tesis presentada para optar al Grado Académico de Magíster en Astronomía

La Serena, Chile, 16 de mayo de 2022

CONSTANCIA

Don .....

HACE CONSTAR:

Que el trabajo correspondiente a la presente Tesis de Magíster, titulada "Identificación y determinación de metalicidades de estrellas gigantes rojas por medio de técnicas de Machine Learning aplicado a la fotometría de banda ancha y angosta del relevamiento S-PLUS", ha sido realizada por Doña Francisca Molina Jorquera, bajo mi dirección.

Para que conste y en cumplimiento de las normativas vigentes de la Universidad de la Serena, Chile, firmo el presente documento en La Serena, Chile, 16 de mayo de 2022

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN ASTRONOMÍA

TÍTULO	:	IDENTIFICACIÓN Y DETERMINACIÓN DE METALICIDADES DE ESTRELLAS GIGANTES ROJAS POR MEDIO DE TÉCNICAS DE MACHINE LEARNING APLICADO A LA FOTOMETRÍA DE BANDA ANCHA Y ANGOSTA DEL RELEVAMIENTO S-PLUS
PRESENTADA POR	:	FRANCISCA MOLINA JORQUERA
DIRECTOR DE TESIS	:	GUILLERMO DAMKE

TRIBUNAL CALIFICADOR

El tribunal de tesis, conformado por:

PRESIDENTE	:
MIEMBROS DEL TRIBUNAL	:

ACUERDAN OTORGARLE LA CALIFICACIÓN DE:

La Serena, Chile, 16 de mayo de 2022

A mi hermano, Matías, a quien admiro mucho y quien fue mi inspiración en este proceso y en muchos más.

Agradecimientos

A mi mamá, Carmen, por su apoyo incondicional, por jugársela por mí y creer en mí, durante este período y durante toda mi vida.

A mi director, Guillermo, por darme la oportunidad de desarrollar este proyecto, por sus consejos, por su apoyo y por todo lo que he aprendido gracias a él.

A Diego, por estar ahí en todo momento, por darme ánimo para seguir adelante y sacarme una sonrisa cuando lo necesito. Este proceso fue difícil, pero sin duda gracias a ti puedo decir que *las risas no faltaron.*

A mi familia, Luis, Matías, Tomás y Carla, quienes me han apoyado incondicionalmente, me han aconsejado y acompañado en todo momento.

A mis amigos y amigas, por el apoyo, por los consejos, por las risas y por los inolvidables momentos.

A los integrantes de la comisión evaluadora, Marcelo, Kathy, Douglas y Antonela, cuyos comentarios ayudaron a mejorar este trabajo.

A DIDULS, por el financiamiento de gran parte del arancel del Programa.

A *Alyria* por demostrarme que soy capaz de lograr mucho más de lo que pienso.

Resumen

El estudio de subestructuras en la Vía Láctea entrega indicios importantes para entender la formación de la Galaxia en un contexto cosmológico. Dichos estudios generalmente se realizan por medio de la identificación de las propiedades dinámicas y/o químicas de las estrellas individuales que las componen. Las propiedades químicas de las estrellas generalmente son determinadas a un primer orden por medio de la metalicidad, la cual generalmente requiere el uso de espectrógrafos. Sin embargo, la aplicación de métodos fotométricos para dichos fines presenta una oportunidad por la gran cantidad de estrellas que pueden observarse de manera simultánea, aunque generalmente las precisiones son menores a su contra parte espectroscópica.

En este trabajo se desarrollaron múltiples redes neuronales capaces de discriminar estrellas entre gigantes y enanas y estimar sus metalicidades fotométricas tomando diferentes combinaciones de colores de Southern-Photometric Local Universe Survey (S-PLUS) como características de entrada. Estas redes están enfocadas a estimar la metalicidad de estrellas de la rama de las gigantes rojas dada su relevancia como trazadores de subestructuras.

Los algoritmos fueron entrenados y probados utilizando un set de datos con 18,063 estrellas con información espectroscópica de APOGEE SDSS DR17. Este set comprende estrellas con 3,500 $<$ T_{eff} (K) $<$ 7,000, 0 $\leq$ log g (cgs) $\leq$ 5 y -2.5 $\leq$ [M/H] (dex) $\leq$ 0.5. Las estrellas de este set fueron seleccionadas como gigantes o enanas por medio de un algoritmo de agrupamiento jerárquico según su posición en el diagrama de Kiel.

La red neuronal de clasificación logró identificar correctamente el 97.1 % y el 97.9 % de las estrellas gigantes y enanas del set de pruebas, respectivamente. Entre las estrellas que este modelo clasificó como gigantes el 97.0 % estaban correctamente clasificadas, mientras que entre las enanas la clasificación correcta alcanzó el 98.0 %. La red para determinar las metalicidades fotométricas de las estrellas gigantes estimó las metalicidades de las estrellas del conjunto de pruebas con desviación estándar $\sigma_{giants} \sim 0.15$ dex, y el algoritmo para las enanas con $\sigma_{dwarf} \sim 0.13$ dex, con respecto a los valores espectroscópicos.

Las redes neuronales desarrolladas en este trabajo se utilizaron para clasificar y determinar las metalicidades de 812,378 estrellas de S-PLUS iDR3. Los algoritmos identificaron 130,172 estrellas gigantes con metalicidades entre -2.2 y 0.3 dex con un peak en $\sim$ -1.35 dex y un doble peak en $\sim$ -0.4 y $\sim$ -0.1 dex, y 682,206 estrellas enanas con metalicidades entre -0.9 y 0.4 dex, con un doble peak en $\sim$ -0.3 y $\sim$ -0.15 dex.

Finalmente, con las distancias de Gaia EDR3, se construyeron gráficas de la distribución espacial de metalicidad de estas estrellas en la Galaxia. Se encontró que las estrellas más ricas en metales se encuentran en la zona del disco Galáctico, y estas a su vez son las más numerosas. En las zonas más lejanas del Sol se empieza a notar la contribución de las estrellas gigantes del halo, representadas como un peak de estrellas pobres en metales pobre en metales. Dicha contribución incluso se nota a la altura del plano Galáctico pero más lejos del Sol, ya que en la vecindad solar la cantidad de estrellas ricas en metales es mucho mayor. Con respecto a las enanas, no se encontró mayor variación de metalicidades dentro de la Galaxia, puesto que la red sólo es capaz de determinar metalicidades en un rango entre -0.9 y 0.4 dex.

Summary

The study of substructures in the Milky Way provides important clues to understand the formation of the Galaxy in a cosmological context. Such studies are generally carried out by identifying the dynamic and/or chemical properties of the individual stars that compose them. The chemical properties of stars are generally determined to a first order by means of metallicity, which generally requires the use of spectrographs. However, the application of photometric methods for these purposes presents an opportunity due to the large number of stars that can be observed simultaneously, although the accuracy is generally lower than its spectroscopic counterpart.

In this work, multiple neural networks capable of discriminating between giants and dwarfs and estimating their photometric metallicities were developed by taking different color combinations from the Southern-Photometric Local Universe Survey (S-PLUS) as input features. These networks are focused on estimating the metallicity of stars of the red giant branch given their relevance as tracers of substructures.

The algorithms were trained and tested using a dataset of 18,063 stars with spectroscopic information from APOGEE SDSS DR17. This set comprises stars with 3,500 $< \mathrm{T}_{eff}$ (K) $<$ 7,000, $0 \leq \log g$ (cgs) ≤ 5 and -2.5 $\leq$ [M /H] (dex) $\leq$ 0.5. The stars in this set were selected as giants or dwarfs by means of a hierarchical clustering algorithm according to their position in the Kiel diagram.

The classification neural network was able to correctly identify 97.1 % and 97.9 % of the giant and dwarf stars in the test set, respectively. Among the stars classified

by the model as giants, 97.0 % were correctly classified, while among the dwarfs the correct classification reached 98.0 %. The network for determining the photometric metallicities of giant stars estimated the metallicities of stars in the test set with standar deviation $\sigma_{giants} \sim 0.15$ dex, and the algorithm for dwarfs with $\sigma_{dwarf} \sim 0.13$ dex, with respect to the spectroscopic values.

The neural networks developed in this work were used to classify and determine the metallicities of 812,378 stars from S-PLUS iDR3. The algorithms identified 130,172 giant stars with metallicities between -2.2 and 0.3 dex with a peak at $\sim$ -1.35 dex and a double peak at $\sim$ -0.4 and $\sim$ -0.1 dex, and 682,206 dwarf stars with metallicities between -0.9 and 0.4 dex, with a double peak at $\sim$ -0.3 and $\sim$ -0.15 dex.

With the distances from Gaia EDR3, graphs of the spatial distribution of metallicity of these stars in the Galaxy were constructed.

Finally, with the distances from Gaia EDR3, graphs of the spatial distribution of metallicity of these stars in the Galaxy were constructed. It was found that the most metal-rich stars are found in the Galactic disk zone, and these in turn are the most numerous. In the most distant zones of the Sun, the contribution of the giant stars of the halo begins to be noticed, represented as a peak of metal-poor stars poor in metals. This contribution is even noticeable at the height of the Galactic plane but further from the Sun, since in the solar neighborhood the number of stars rich in metals is much greater. Regarding dwarfs, no greater variation of metallicities was found within the Galaxy, since the network is only capable of determining metallicities in a range between -0.9 and 0.4 dex.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En esta tesis se desarrolló un método para clasificar estrellas entre gigantes y enanas y luego determinar sus metalicidades, a partir de la fotometría del *survey* S-PLUS. En este capítulo se presentan los antecedentes teóricos de las distintas temáticas que se abordan en este trabajo, empezando por la estructura de nuestra Galaxia y la importancia de las abundancias químicas y de las estrellas gigantes rojas en la materia. También se describe el Southern Phtometric Local Universe Survey (S-PLUS), haciendo particular énfasis en el sistema fotométrico que utiliza. Finalmente se realiza una descripción de las técnicas y se definen algunos conceptos relativos a las áreas de *machine learning* y *deeep learning* que han tenido un auge durante los últimos años, destacando su utilidad en la resolución de diversos problemas, incluyendo trabajos que abordan el problema de la determinación de metalicidades estelares.

1.1. Estructura Galáctica

Las galaxias han sido consideradas como los bloques fundamentales que componen el Universo; para entenderlas es necesario comprender sus elementos individuales y cómo se relacionan entre ellos (Binney et al., 1998). En este sentido, la estructura y fenómenos a gran escala suelen ser observados y estudiados con una perspectiva completa en galaxias externas. Pero en el caso de los fenómenos a pequeña escala, la Vía Láctea otorga una oportunidad única para estudiarlos en gran detalle (Binney et al., 1998; Ivezić et al., 2008).

En términos generales, las estrellas de la Vía Láctea están distribuidas en diferentes componentes: disco fino, disco grueso, un bulbo central y el halo Galáctico. El disco contiene a su vez una estructura de brazos espirales similares a los observados en otras galaxias (Schneider, 2006). El disco fino contiene una importante fracción de gas, polvo y estrellas jóvenes ($\leq$ 8 Gyr de edad), mientras que el disco grueso está conformado por una población de estrellas más viejas ($\sim$ 10 Gyr de edad) (Carroll & Ostlie, 2017). Mediante el estudio de la densidad numérica de estrellas en la Galaxia, se encontró que el disco fino estaría caracterizado por una escala de altura de $\sim$ 300 pc y una escala de longitud de $\sim$ 2600 pc, mientras que el disco grueso tendría una escala de altura de $\sim$ 900 pc y una escala de longitud de $\sim$ 3600 pc (Jurić et al., 2008). El bulbo es la estructura central de la Galaxia, en la cual se han detectado cúmulos globulares y estrellas viejas con edades de $\sim 10 \pm 2.5$ Gyr (Ortolani et al., 1995; Zoccali, 2005). En esta componente se detectó la presencia de una estructura barrada, observada en mapas en infrarrojo del experimento DIRBE (Stanek et al., 1994) y también en el estudio de estrellas del *red clump* (Dwek et al., 1995). El halo estelar está compuesto por un sistema de cúmulos globulares y estrellas de campo (no asociadas a cúmulos) de edades entre $\sim$ 10 y 12 Gyr (Unavane et al., 1996) distribuidas hasta más de 100 kpc desde el centro Galáctico (Sanderson et al., 2017; Deason et al., 2021; Belokurov et al., 2019) e incluso hasta 335 kpc (Stringer et al., 2021). Adicionalmente, existe evidencia de la presencia de un halo de materia oscura (Carroll & Ostlie, 2017). En la Figura 1.1 se muestra una representación de la estructura de la Vía Láctea con sus diferentes componentes desde una vista *edge-on*. En la Figura 1.2 se muestra una representación artística de la estructura espiral de la Galaxia desde una vista *face-on*.

1.1.1. Abundancias químicas

Una población estelar simple es un grupo de estrellas que comparten un origen e historia en común, y por lo tanto, además de la edad, comparten contenido químico y propiedades dinámicas. Es por esto que el estudio de las poblaciones estelares presentes en la Vía Láctea es de gran importancia para comprender la historia de formación y evolución de la Galaxia. Los estudios de las poblaciones estelares se llevan a cabo analizando la cinemática, distribución espacial, edades y abundancias químicas de las estrellas. Mientras que las distribuciones espaciales y la cinemática

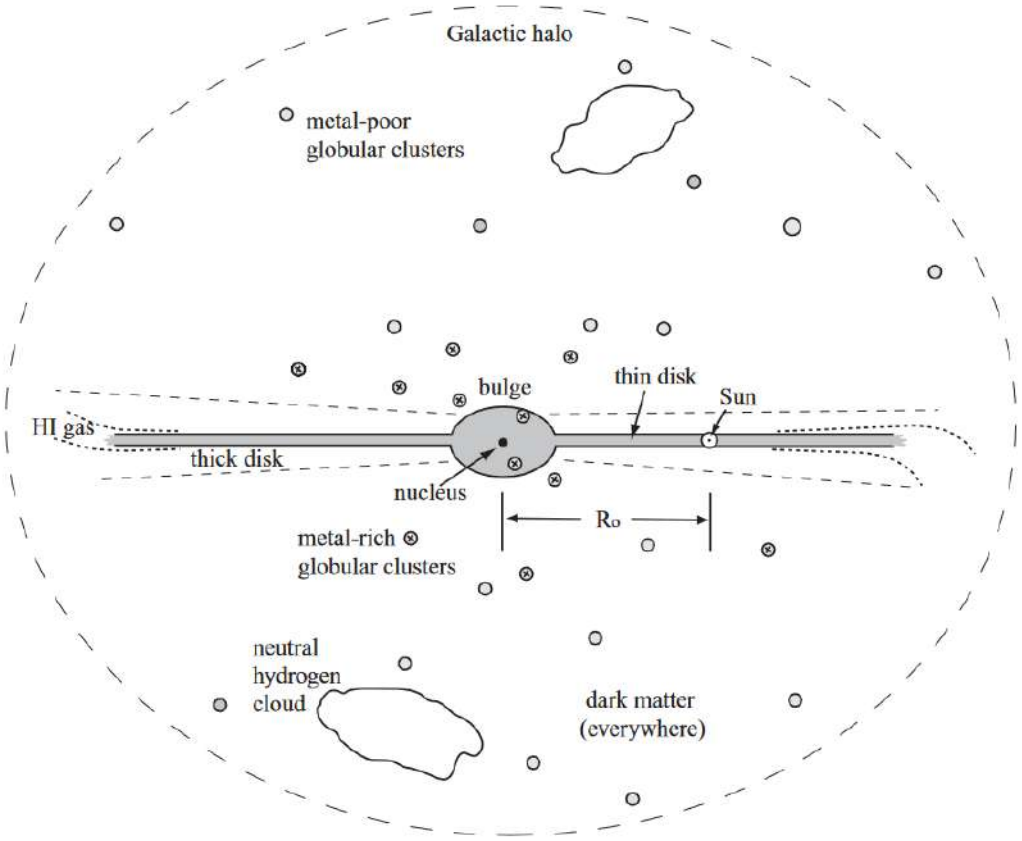


Figura 1.1: Diagrama *edge-on* de la Vía Láctea (fuente: Sparke & Gallagher, 2017).

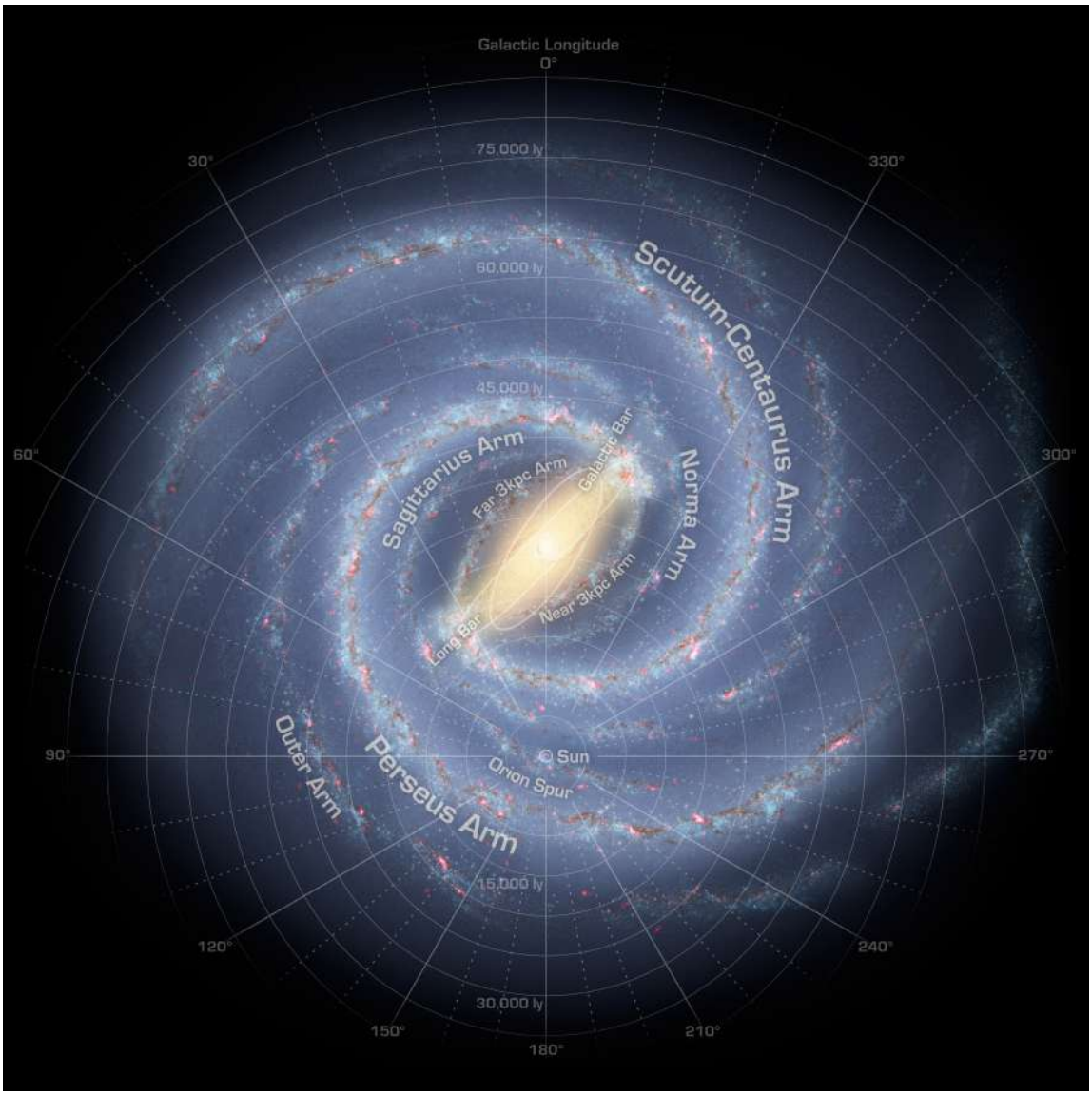


Figura 1.2: Concepto artístico de la Vía Láctea en vista *face-on*, basada en la evidencia observacional disponible al momento de su publicación. Se aprecia la estructura espiral identificada mediante imágenes del Telescopio Espacial Spitzer de la NASA. La estructura espiral de la Galaxia está dominada por dos brazos principales (Scutum-Centaurus y Perseus) unidos a los extremos de una barra central, dos brazos menores degradados (Norma y Sagitario) y menos distinguibles ubicados entre los brazos principales. También incluye los brazos espirales llamados en inglés "Far 3 kiloparsec" y "Near 3 kipolarsec", denominados de esta manera por encontrarse cada uno a 3 kpc del centro Galáctico alrededor del bulbo por el lado lejano y cercano, respectivamente, desde nuestra perspectiva. El Sol se encuentra cerca de un brazo parcial llamado Brazo de Orión, ubicado entre los brazos de Sagitario y Perseo (fuente: NASA/JPL-Caltech/R. Hurt (SSC/Caltech), 2017).

de las estrellas pueden variar a lo largo del tiempo, la composición química en las atmósferas estelares se considera un registro "fósil" que puede llevar a trazar el lugar y momento de formación de la estrella (Oswalt & Gilmore, 2013).

Para cuantificar la composición química de una estrella se utiliza la **metalicidad**. Es importante notar que, en astronomía, todos los elementos más pesados que el helio son denominados como metales. Para un elemento X, la metalicidad de una estrella está definida como:

$$[X/H] \equiv log_{10}\left(\frac{N_X}{N_H}\right)_* - log_{10}\left(\frac{N_X}{N_H}\right)_\odot \tag{1.1}$$

donde N_X y N_H son el número de átomos + iones del elemento X y de hidrógeno, respectivamente. Esta cantidad se mide en "dex" —unidad que proviene de *decimal exponent*. La metalicidad estelar es comúnmente considerada como el contenido de hierro [Fe/H], ya que las líneas de hierro suelen ser muy numerosas y fáciles de identificar en los espectros estelares (Carroll & Ostlie, 2017).

Los metales no fueron producidos en el Universo temprano, sino que fueron generados posteriormente al interior de las estrellas de generaciones anteriores a partir de hidrógeno y helio primordiales. En particular, los elementos alfa (O, Ne, Mg, Si, S, Ar, Ca, Ti, creados mediante adiciones sucesivas de núcleos de helio) son sintetizados en estrellas masivas y expulsados al medio interestelar durante el colapso del núcleo (supernovas tipo II), mientras que el hierro es principalmente liberado al medio interestelar por supernovas de tipo Ia (Oswalt & Gilmore, 2013). Las estrellas que se formen posteriormente en este medio enriquecido, serán por lo tanto, más ricas en metales (Schneider, 2006; Carroll & Ostlie, 2017; Matteucci, 2001).

El concepto de poblaciones estelares, introducido inicialmente por Baade (1944), fue entendido como un conjunto de estrellas bien definido en el diagrama H-R, lo que implica distribuciones bien definidas de edad y metalicidad (Oswalt & Gilmore, 2013). Esto dio la base para la separación de las estrellas en una Población I, formada por las estrellas azules del disco, jóvenes y ricas en metales, y una Población II de estrellas rojas de los cúmulos globulares del halo, viejas y pobres en metales (Matteucci, 2001; Oswalt & Gilmore, 2013). Sin embargo, el panorama resulta ser mucho más complejo que uno definido simplemente por dos poblaciones estelares. De hecho, hoy se aplica el término población estelar simple para una sola generación de estrellas, con una

misma función de masa inicial, abundancia química, cinemática y edad (Oswalt & Gilmore, 2013). Cada componente de la Vía Láctea tiene poblaciones químicamente diferentes. Se describen al menos cuatro tipos de poblaciones estelares principales en la Vía Láctea: la población del halo estelar, de estrellas con metalicidades [Fe/H] entre -4.5 a -0.5 dex; la población del bulbo con [Fe/H] entre -2 y 0.5; la población del disco grueso con [Fe/H] entre -2.2 y -0.5; y la población del disco fino con [Fe/H] entre -0.5 y 0.3 (Schneider, 2006; Carroll & Ostlie, 2017).

1.1.2. Subestructuras en la Vía Láctea

En los últimos años, desde la salida de SDSS (York et al., 2000), se ha encontrado evidencia sobre la complejidad de la estructura de la Vía Láctea, sobre todo en el halo. En la Figura 1.3 se observa que la distribución de las estrellas en el halo está lejos de ser uniforme, sino que muestra una serie de subestructuras. Entre las características más dominantes se encuentran los restos de marea de la galaxia enana esferoidal de Sagitario, (Sgr dSph; Ibata et al., 1995; Yanny et al., 2000; Ibata et al., 2001; Majewski et al., 2003) y la corriente estelar Monoceros (Newberg et al., 2002; Rocha-Pinto et al., 2003). Se observan también la corriente Orphan (Ibata et al., 1997; Majewski et al., 2003; Belokurov et al., 2007a), la sobredensidad de Virgo (Jurić et al., 2008) y galaxias satélites (Belokurov et al., 2006a, 2007b). Otros componentes en el halo de la Galaxia son las Nubes de Magallanes (también galaxias satélites), la corriente de Magallanes (Wannier & Wrixon, 1972; Nidever et al., 2008) o el cúmulo globular ω Centauri, el cual presenta evidencias de tratarse del remanente de una galaxia enana (Villard et al., 2008).

En los años más recientes, los surveys a gran escala han aportado toda una nueva visión en el área. Por ejemplo, Shipp et al. (2018) a partir de los datos de DES (Dark Energy Survey Collaboration et al., 2016) encontraron 11 corrientes estelares adicionales en la Galaxia y Bechtol et al. (2015) identificaron 8 candidatos a satélites Galácticos en el halo. Por otro lado, a partir de la información provista por Gaia (Gaia Collaboration et al., 2016) se han extendido los estudios en el área. Por ejemplo, Ibata et al. (2021) encontraron 9 corrientes estelares adicionales (Hríd, Gunnthrá y Gaia-6 a Gaia-12) con datos de Gaia DR2 y Gaia EDR3, además de identificar otros candidatos a corrientes estelares y colas de marea asociadas a cúmulos globulares de la Galaxia, o la identificación de Gaia-Enceladus por Belokurov et al. (2018).

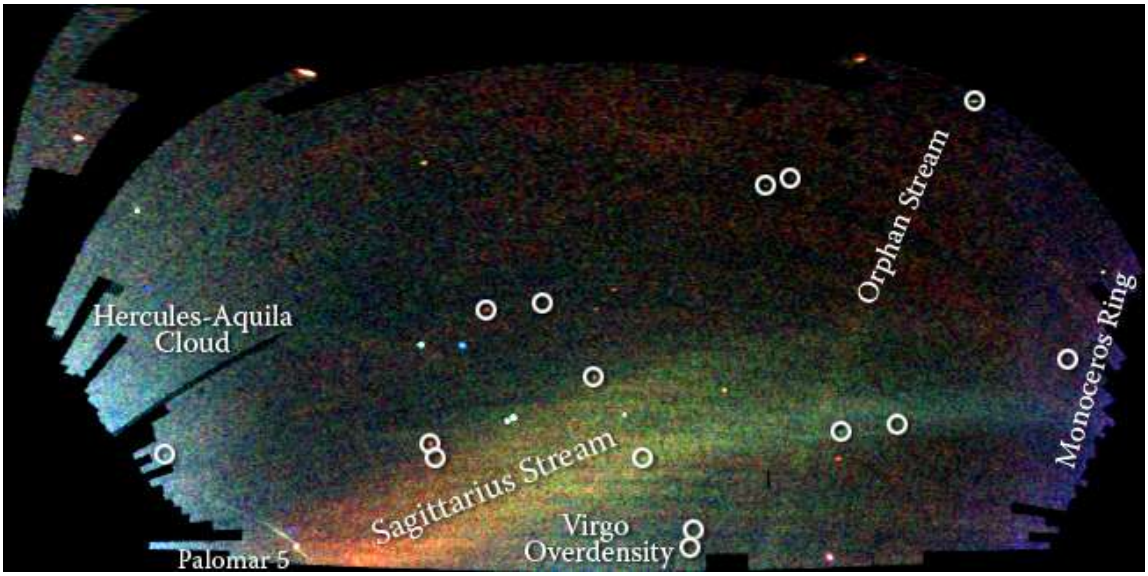


Figura 1.3: "Field of Streams". Mapa de estrellas en las regiones exteriores de la Vía Láctea, derivado a partir de las imágenes de SDSS, en una proyección tipo Mercator. El color indica la distancia de las estrellas (en azul se presentan las estrellas más cercanas con $20.0 < r \leq 20.66$, en verde estrellas con $20.66 < r \leq 21.33$ y rojo para estrellas más lejanas con $21.33 < r \leq 22.00$) y la intensidad indica la densidad de estrellas en el cielo. Las estructuras visibles en este mapa incluyen corrientes o *streams* de estrellas provenientes de la galaxia enana de Sagitario, la corriente Orphan que cruza la corriente de Sagitario, la corriente de Monoceros que rodea el disco de la Vía Láctea, rastros de estrellas que están siendo desprendidas del cúmulo globular Palomar 5, y excesos de estrellas encontradas hacia las constelaciones de Virgo y Hércules. Los círculos encierran nuevos compañeros de la Vía Láctea descubiertos por el SDSS; dos de ellos son cúmulos globulares y los otros son galaxias enanas tenues (fuente: Belokurov et al., 2006b & Sloan Digital Sky Survey).

1.1.3. Estrellas gigantes rojas como trazadoras de subestructuras

Una gigante roja es una estrella que ha agotado el suministro de hidrógeno en su núcleo y ha comenzado la fusión termonuclear de hidrógeno en una capa que rodea al núcleo de helio (Sparke & Gallagher, 2007). Durante su vida en la fase de secuencia principal, las estrellas convierten el hidrógeno del núcleo en helio. Su etapa en la secuencia principal termina cuando ya casi todo el hidrógeno en el núcleo ha sido fusionado. Una vez agotado el hidrógeno en el núcleo, éste se contrae. Este proceso hace que se den las condiciones de temperatura y presión para reanudar la quema de hidrógeno en la capa que rodea el núcleo. Cuando el núcleo se contrae, a su vez, aumenta su temperatura, y las capas exteriores de la estrella se expanden. En este proceso de enfriamiento y expansión, la estrella se denomina subgigante. Cuando la envoltura de la estrella se enfría lo suficiente, la estrella deja de expandirse, su luminosidad comienza a aumentar y la estrella asciende por la rama gigante roja del diagrama Hertzsprung-Russell (Prialnik, 2000; Zeilik & Gregory, 1998; Campante et al., 2017).

En la Figura 1.4 se muestra la posición de las estrellas gigantes rojas en el diagrama H-R de la misión Gaia (Gaia Collaboration et al., 2018).

Las estrellas gigantes rojas ofrecen un gran potencial para estudiar subestructuras en la Galaxia, sobre todo en el halo, e incluso a mayores distancias dentro del Grupo Local. Esto es debido a que son lo suficientemente numerosas para trazar subestructuras, a diferencia de las estrellas de la rama horizontal, que se utilizan para detectar poblaciones pero no son tan abundantes; y lo suficientemente brillantes como para ser detectadas a grandes distancias, a diferencia de las estrellas enanas que son numerosas pero intrínsecamente débiles. Adicionalmente, las estrellas gigantes K se encuentran en poblaciones de todas las edades y metalicidades, por lo que cualquier subestructura es potencialmente trazable por este tipo de estrellas (Majewski, 2004). Con estrellas gigantes se pueden explorar grandes volúmenes en partes externas de la Galaxia de manera eficiente, incluso con telescopios pequeños. Sin embargo, es necesario contar con un método que sea capaz de identificar las estrellas gigantes de las estrellas enanas más cercanas (Majewski et al., 2000).

Existen diversos estudios de la subestructura de la Galaxia y el Grupo Local, por

medio de estrellas gigantes. Entre ellos está la serie de trabajos "Exploring Halo Substructure with Giant Stars", que consiste en un estudio de la estructura del halo de la Vía Láctea y de los halos de otras galaxias del Grupo Local, según lo trazado por las estrellas gigantes que los componen (Majewski et al., 2000), incluyendo estudios de las galaxias satélites enanas esferoidales de la Vía Láctea Ursa Menor, Carina, Leo I, Leo II, Sculptor y Sagitario (Majewski, 2004), de la corriente de marea de *omega* Centauri Majewski et al. (2012), entre otros. También se ha estudiado la subestructura en el halo de M31 (Ferguson et al., 2002; McConnachie et al., 2009; Ibata et al., 2014), la estructura del halo de la Galaxia con estrellas gigantes K de SEGUE (Janesh et al., 2016) y de LAMOST (Yang et al., 2019) y estudios de subestructuras en el anticentro de la Galaxia (Li et al., 2021).

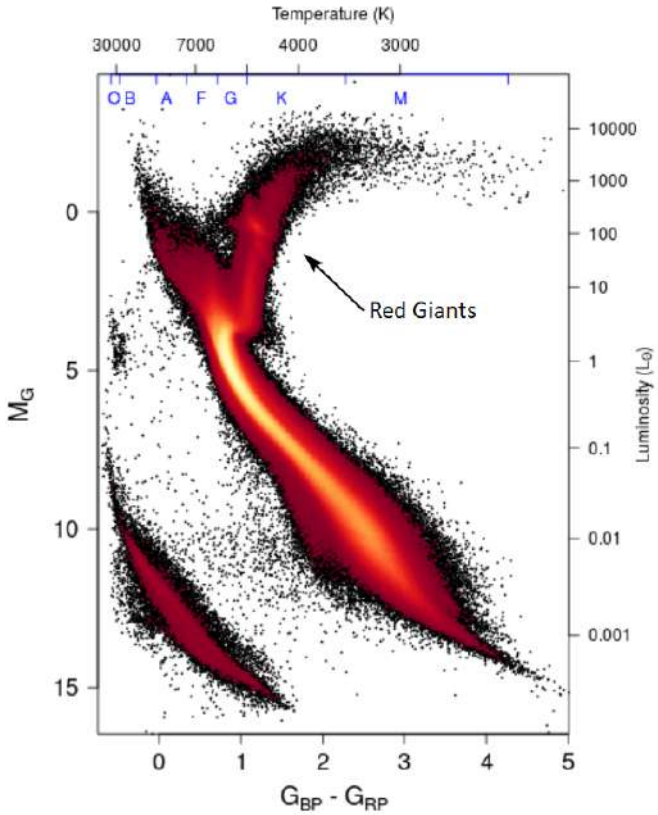


Figura 1.4: Diagrama Hertzsprung–Russell observacional M_G vs G_{BP} - G_{RP} de 4,276,690 estrellas con baja extinción (E(B - V) < 0.015 mag) de Gaia DR2. La escala de colores representa la raíz cuadrada de la densidad de estrellas. Los equivalentes aproximados de temperatura y luminosidad de las estrellas se indican en los ejes superior y derecho, respectivamente. Adicionalmente se indica la secuancia de las estrellas gigantes rojas (*Red Giants*) (fuente: Gaia Collaboration et al., 2018).

1.2. Southern Photometric Local Universe Survey

La llegada de los grandes surveys durante las últimas décadas nos ha provisto de grandes volúmenes de datos con información de millones de objetos astronómicos, de todo o gran parte del cielo en diferentes longitudes de onda, abriendo una nueva era en la astronomía. En este ámbito, sin dudas el Sloan Digital Sky Survey (SDSS, York et al. 2000) marcó un gran precedente, dejando en evidencia la importancia de los surveys fotométricos como proovedores de información homogénea de millones de objetos en el cielo en múltiples bandas. Otros ejemplos de estos surveys son: Two Micron All-Sky Survey (2MASS; Skrutskie et al. 2006), VISTA Hemisphere Survey (VHS; McMahon et al. 2013), Pan-STARRS (Chambers et al., 2016), Dark energy Survey (DES; Dark Energy Survey Collaboration et al. 2016), Gaia (Gaia Collaboration et al., 2016) o Southern Photometric Local Universe Survey (S-PLUS; Mendes de Oliveira et al. 2019). Otra ventaja es que los objetos detectados, y finalmente incluidos en los datos de un relevamiento fotométrico de gran campo, sólo se ven sesgados por la magnitud límite de las bandas y la resolución de los instrumentos. Estos relevamientos son capaces de proveer información de la distribución de energía espectral de los objetos (SED) y su utilidad destaca en trabajos para objetivos científicos que no requieran cierta resolución espectral, es decir, que la información espectroscópica no sea estrictamente necesaria en una primera instancia (Cenarro et al., 2019). Además, los relevamientos fotométricos son de alta utilidad para suministrar la información necesaria para la selección de objetos de surveys espectroscópicos (por ejemplo, Angeloni et al. 2019).

En específico, y por pertinencia con este trabajo, el Southern Photometric Local Universe Survey (S-PLUS) es un survey fotométrico óptico en ejecución que comenzó sus operaciones en 2016 y que cubrirá un área en total de $\sim 9,300$ deg^2 (para 2021 llevaba cubierto un $\sim 24\,\%$ dle área planificada) con el sistema fotométrico de 12 filtros de Javalambre (Cenarro et al., 2019) utilizando el telescopio robótico dedicado T80-South (T80S) de 0.826 m de apertura y configuración Ritchey-Chretien, ubicado en el Cerro Tololo Inter-american Observatory (CTIO) en el norte de Chile. Posee una cámara CCD de 9232 $\times$ 9216 píxeles de 10 μm y una escala de 0.55 arcsec-píx, lo que

resulta en un campo de 1,4 × 1,4 grados (Mendes de Oliveira et al., 2019; Almeida-Fernandes et al., 2021). Tanto el sistema de filtros, como el telescopio y la cámara son idénticos a los utilizados en el Javalambre Photometric Local Universe survey (J-PLUS), llevado a cabo con el Javalambre Auxiliary Survey Telescope (T80/JAST) en el Observatorio Astrofísico de Javalambre (OAJ) en España (Cenarro et al., 2019). Esto vuelve a S-PLUS un survey complementario a J-PLUS en el hemisferio sur. A la fecha, entre los grandes relevamientos fotométricos en el hemisferio sur, S-PLUS es aquel que cuenta con el mayor número de bandas fotométricas (Almeida-Fernandes et al., 2021).

1.2.1. Sistema fotométrico de Javalambre

El sistema fotométrico de Javalambre cubre el rango óptico completo y parte del infrarrojo cercano (entre 3,500 y 10,000 Å) y se compone de una combinación de 5 filtros de banda ancha y 7 filtros de banda angosta (Cenarro et al., 2019) (Figura 1.5).

Los 5 filtros de banda ancha corresponden a los filtros *ugriz*, los 4 últimos similares a los de SDSS (Fukugita et al., 1996), y el filtro *u* está especialmente diseñado para el sistema de Javalambre, contando con mayor eficiencia que la banda *u* de SDSS (Mendes de Oliveira et al., 2019). Los 7 filtros de banda angosta (F378, F395, F410, F430, F515, F660, F861)[1] están localizados sobre características espectrales prominentes, las cuales se indican en la Tabla 1.1. Estas características proveen de un escenario ideal para la clasificación y caracterización de poblaciones estelares (Gruel et al., 2012; Marín-Franch et al., 2012; Cenarro et al., 2019).

Este sistema fotométrico está optimizado para la clasificación estelar (ya que permite reconstruir la SED estelar) y para la determinación de los parámetros estelares: temperatura efectiva (T_{eff}), gravedad superficial (log g) y metalicidad ([Fe/H]) (Gruel et al., 2012; Marín-Franch et al., 2012). Los diferentes parámetros de la atmósfera estelar implican una variación en el ancho equivalente de una línea espectral y por lo tanto, una variación del flujo en la región de la línea y por ende en el pasabanda fotométrico. Mediante este mecanismo es que la sensibilidad a los parámetros este-

[1]Notar que en algunas fuentes los filtros de banda angosta de este sistema están indicados por una sintaxis que antepone J0 al número del filtro. Por lo tanto, el filtro F861, por ejemplo, también es demonimado J0861.

Tabla 1.1: Características de los filtros del sistema de Javalambre (fuente: S-PLUS, 2019).

Nombre del filtro	λ_{eff} (Å)	$\Delta\lambda$ (Å)	Comentarios
u	3,536	352	u Javalambre
F378	3,770	151	[OII]
F395	3,940	103	Ca H+K
F410	4,094	201	Hδ
F430	4,292	201	banda G
g	4,751	1,545	g SDSS
F515	5,133	207	triplete de Mgb
r	6,258	1,465	r SDSS
F660	6,614	147	Hα
i	7,690	1,506	i SDSS
F861	8,611	408	triplete de Ca
z	8,831	1,182	z SDSS

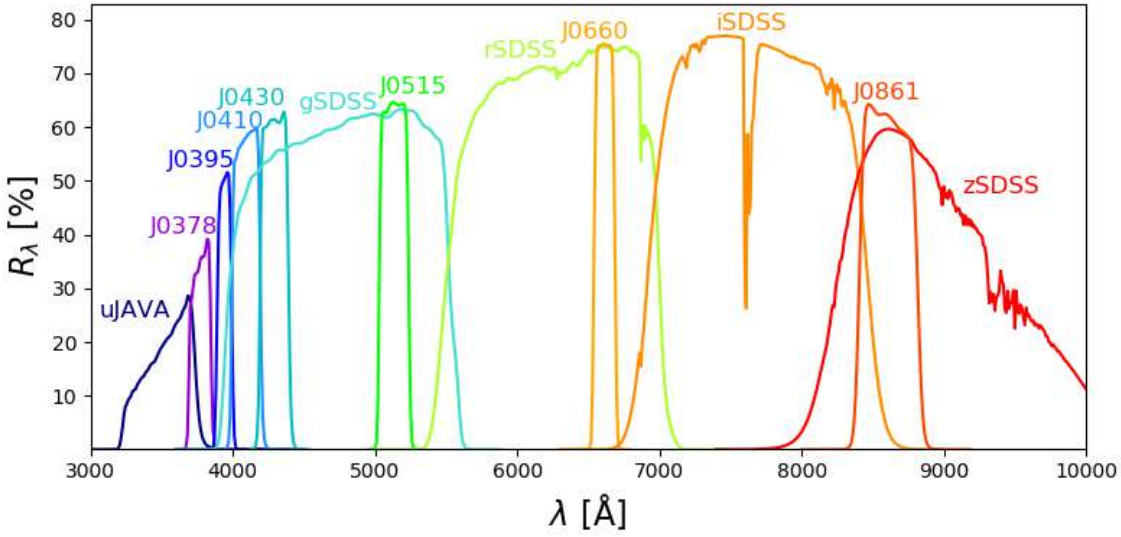


Figura 1.5: Curvas de transmisión total del sistema de filtros de Javalambre. En el eje vertical se encuentra la eficiencia y en el eje horizontal la longitud de onda en Angstroms. La eficiencia incluye las contribuciones de la transmisión del filtro, la transmisión de la atmósfera, la eficiencia del CCD y las curvas de reflectividad del espejo primario. Este sistema de filtros fotométricos está compuesto por 7 filtros de banda angosta (F378, F395, F410, F430, F515, F660, F861) y 5 filtros de banda ancha (u, g, r, i, z) (fuente: S-PLUS, 2019).

lares es útil desde el punto de vista de los surveys fotométricos. Con respecto a las bandas de S-PLUS, por ejemplo, de hace mucho tiempo es conocida la dependencia del triplete de magnesio con la gravedad superficial (Öhman, 1934; Thackeray, 1939). En la Figura 1.6 se muestra esta sensibilidad en el espectro de una enana y una gigante tipo K. A su vez, existen estudios que han mostrado la dependencia entre las líneas del triplete de Calcio y la metalicidad de gigantes rojas en cúmulos estelares (Armandroff & Da Costa, 1991; Cole et al., 2004; Warren & Cole, 2009). En la Figura 1.7 se observa la variación de las líneas del triplete de calcio con la metalicidad. En la Figura 1.8 se muestra la respuesta de las líneas de Ca II H+K frente a la variación de la temperatura efectiva y la metalicidad y la sensibilidad de Hγ a la temperatura efectiva. Es importante remarcar que el tamaño de las características espectrales mencionadas anteriormente son lo suficientemente prominentes para ser medidas fotométricamente.

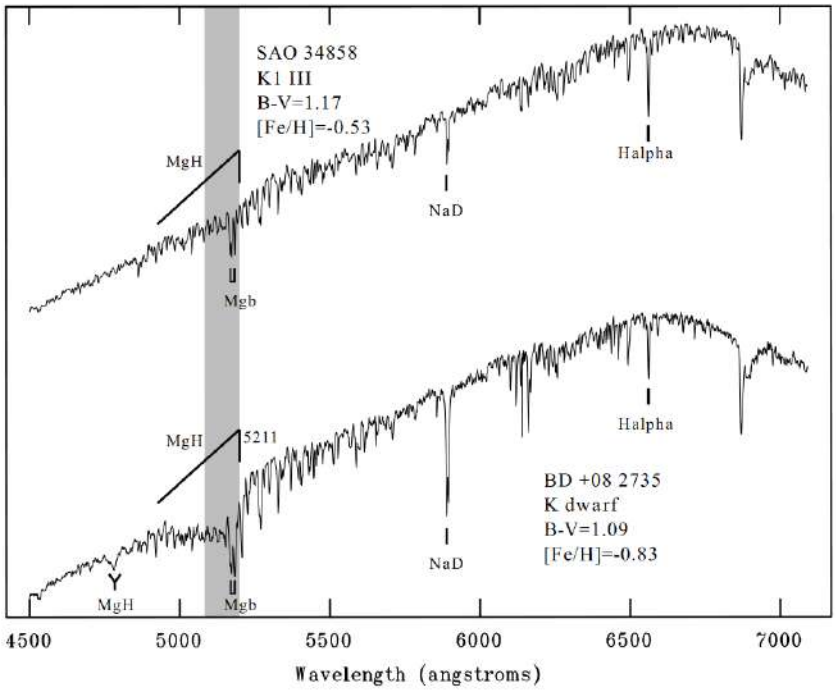


Figura 1.6: Comparación del espectro de una estrella gigante y enana K con abundancias similares. Se muestra la dependencia del triplete de magnesio en la gravedad superficial (fuente:Majewski et al., 2000).

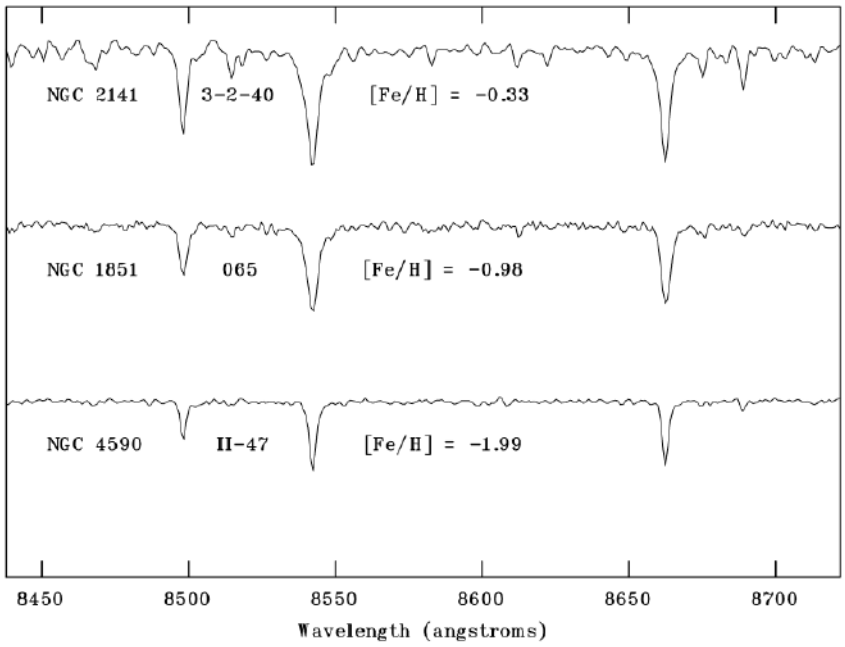


Figura 1.7: Comparación del espectro de estrellas gigantes rojas de temperaturas y gravedades similares en tres cúmulos estelares centrado en el triplete de Calcio. Se observa la variación del triplete de calcio con las metalicidades (fuente: Cole et al., 2004).

1.2.2. Sub-surveys y catálogos fotométricos

El survey S-PLUS está dividido en 5 sub-surveys (Mendes de Oliveira et al., 2019). Estos son:

- Main Survey (MS): cubre un área de ~ 8000 deg^2 bajo una estrategia motivada por los requerimientos del área de la astronomía extragaláctica. Incluye las áreas de la Stripe 82 y las Nubes de Magallanes.
- Ultra Short Survey: cubriendo las mismas áreas que el Main Survey, y con menores tiempos de exposición, está orientado a la búsqueda e identificación de estrellas pobres en metales por la importante información que éstas poseen respecto a la formación y evolución de la Vía Láctea.
- Variability Fields Survey: observando en campos que ya hayan sido cubiertos por el Main Survey, incluye observaciones de variables cataclísmicas, binarias eclipsantes, AGN, super novas (SNe), asteroides, eventos de ondas gravitacionales, entre otros.

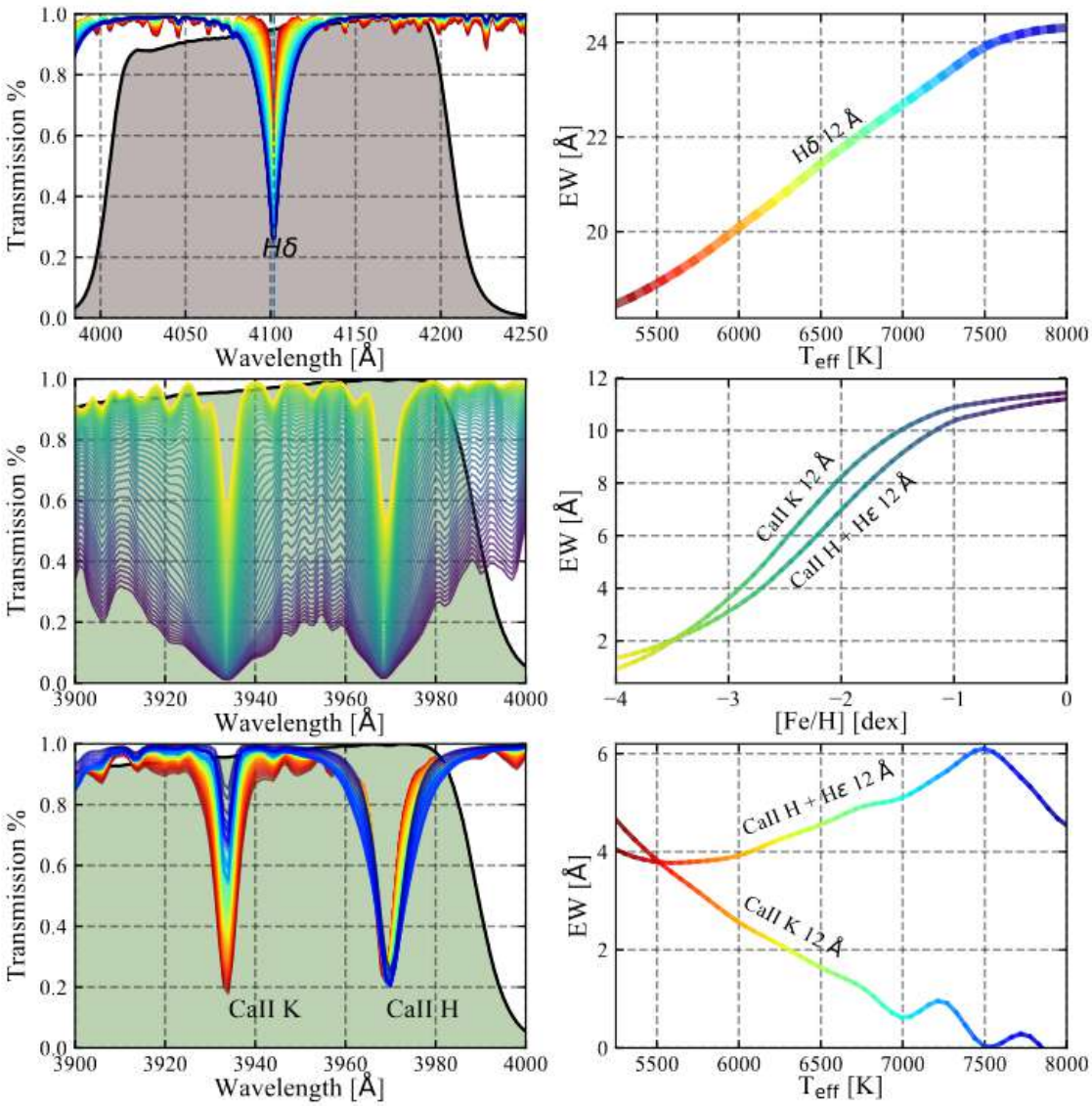


Figura 1.8: Panel superior: sensibilidad de Hγ a la temperatura efectiva en una estrella de log g = 2.5 y [Fe/H] = -2.5. Panel intermedio: variación en las líneas de Ca II H+K con la metalicidad de una estrella de T_{eff} = 5,000 K y log g = 1. Panel inferior: respuesta de la línea Ca II H+K al aumento de la temperatura efectiva. (fuente: Whitten et al., 2019a).

- Galactic Survey: cubre ~ 1420 deg^2 en la zona del plano Galáctico, incluyendo regiones en el bulbo y disco. Se orienta principalmente al estudio de estrellas variables y cúmulos abiertos; en ambos casos, el sistema de filtros de Javalambre ayuda a determinar las características de los objetos, ya sean tipos espectrales, edades, metalicidades y masas del cúmulo.
- Marble Field Survey (MFS): compuesto por un conjunto de campos que son visitados con mayor frecuencia en noches cuando la visibilidad es demasiado pobre para observar MS, es decir, peor que 2". Los objetos seleccionados para el MFS son las Nubes de Magallanes Grande y Pequeña (LMC y SMC), M83, el Grupo Dorado y el cúmulo Hydra. El MFS es adecuado para la identificación y caracterización de estrellas variables y el estudio de galaxias cercanas.

Los datos de S-PLUS son liberados regularmente a través de Data Releases. Hasta la fecha, hay dos lanzamientos públicos, más un tercer Data Release de acceso interno. El primer Data Release (DR1) contiene observaciones entre 2016 y 2018 y fue publicado el 2 de julio de 2019. Incluye las observaciones en un área de 336 deg^2 en la Stripe 82[2], contando con ~ 3 millones de detecciones. El último lanzamiento público es el segundo Data Release (DR2), liberado el 30 de marzo de 2021. El DR2 contiene las observaciones realizadas desde al año 2016 hasta marzo de 2020, es decir, incluye las observaciones de la DR1 y cubre un área total de 950.5 deg^2, incluyendo más de 30M detecciones. Adicionalmente, se encuentra disponible el tercer Data Release interno (iDR3) para los miembros de la colaboración. Este último cuenta con > 50 millones de detecciones (Almeida-Fernandes, 2020). La cobertura de S-PLUS iDR3 se muestra en la Figura 1.9.

Los catálogos fotométricos de S-PLUS incluyen la siguiente información para cada fuente (Mendes de Oliveira et al., 2019; Almeida-Fernandes et al., 2021):

- Identificación: campo e IDs;
- Astrometría: coordenadas y posiciones;
- Fotometría: magnitudes en distintos tipos de apertura, errores y señal ruido para cada filtro;

[2]Región rectangular entre las coordenadas $0^o <$ RA $< 60^o$, $300^o <$ RA $< 360^o$ y -1.26$^o <$ DEC $< 1.26^o$ (Alam et al., 2015), que ha sido extensamente observada por diferentes proyectos y en distintas longitudes de onda.

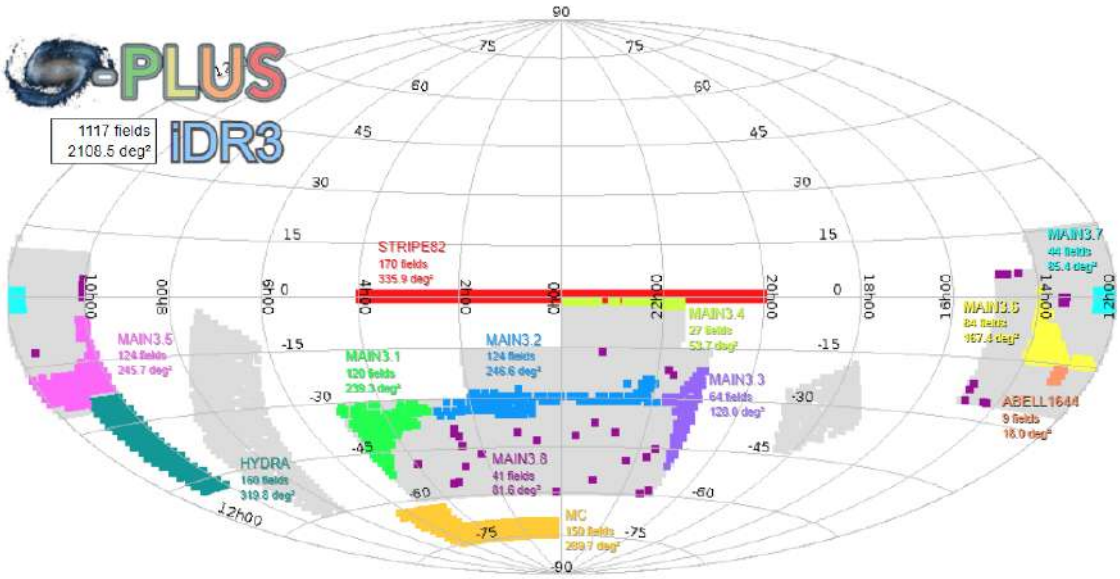


Figura 1.9: Cobertura completa en el cielo del proyecto S-PLUS en coordenadas ecuatoriales. Los diferentes colores señalan los 1117 campos observados hasta el 3rd internal Data Release del survey, que incluye 2108.5 deg^2, mientras que las áreas en gris corresponden a campos aún no observados (fuente: Almeida-Fernandes, 2020).

- Información morfológica: elipticidad, ángulo de posición, estelaridad, entre otras;
- Flags.

Por otro lado, SPLUS incluye información adicional por medio de los *Value Added Catalogs* (VACs), tales como redshifts fotométricos, clasificación estrella/galaxia, entre otros.

1.3. Técnicas computacionales

1.3.1. Machine learning

El aprendizaje automático (o automatizado) o *machine learning* es "*el campo de estudio que le da a las computadoras la capacidad de aprender sin estar explícitamente programadas*" (Samuel, 2000). Por otro lado, Mitchell (1997) define *machine learning* como "*el estudio de algoritmos informáticos que mejoran automáticamente a través*

de la experiencia", o puesto de otra manera: "*se dice que un programa de computadora aprende de una experiencia E con respecto a alguna tarea T y alguna medida de rendimiento P, si su rendimiento en T, medido por P, mejora con experiencia E.*". Por ejemplo, pensemos en un programa computacional que implemente un algoritmo de *machine learning* que es capaz de reconocer imágenes de perros y gatos. La tarea T es reconocer si una imagen corresponde a un perro o a un gato, la medida de rendimiento P podría ser la fracción de imágenes correctamente clasificadas y la experiencia vendría dada por el catálogo o set de entrenamiento. Este último, en este caso, correspondería a un conjunto de imágenes clasificadas con anterioridad a partir de las cuales el algoritmo aprende.

En *machine learning* las características, características de entrada, o simplemente entradas (*features*, *input features* o *features*), son aquellas variables independientes que describen al conjunto de datos y que son ingresadas al sistema para que éste "aprenda" a partir de ellas. Los objetivos o salidas (*targets* u *outputs*) son los resultados previstos por el sistema de aprendizaje automatizado a partir de las variables de entradas y pueden ser valores tanto categóricos como continuos (Bhattacharjee, 2017).

El aprendizaje automatizado resulta muy útil y práctico para resolver problemas cuyas soluciones pueden ser muy complejas o muy demandantes en tiempo, o bien para obtener información de grandes cantidades de datos.

1.3.1.1. Diferentes tipos de sistemas de aprendizaje automatizado

Existen diferentes tipos de sistemas de *machine learning*. Por ejemplo, está el aprendizaje supervisado y el no supervisado. En el primero los datos de entrenamiento incluyen las soluciones, denominadas como etiquetas o *labels*. En cambio, en el segundo las soluciones no son entregadas previamente al algoritmo. Por otro lado, está el aprendizaje por batch (o *batch learning*) y el aprendizaje en línea (*online learning*). Los sistemas de *batch learning* son entrenados con todos los datos disponibles y luego aplica lo aprendido; si se quieren agregar nuevos datos se debe entrenar desde cero. En el aprendizaje en línea el sistema es capaz de aprender sobre los nuevos datos, de forma incremental y progresiva, a medida que los recibe (Géron, 2017).

Las tareas de aprendizaje supervisado más comunes son la regresión y clasificación: los modelos de regresión se utilizan para predecir resultados cuantitativos (valores numéricos continuos) y los modelos de clasificación para predecir resultados cualitativos (clases o categorías) (Hastie et al., 2009). Los problemas de clasificación más comúnmente trabajados son los de clasificación binaria, en donde el sistema distingue entre dos clases o categorías, y la clasificación multiclases, en donde el sistema es capaz de diferenciar entra más de dos clases. Existen problemas de regresión múltiple en donde el sistema utiliza múltiples características para hacer una predicción, regresión univariable donde se predice un valor único para cada ejemplo[3], o regresión multivariable en los cuáles se predicen múltiples valores para cada ejemplo (Géron, 2017).

1.3.1.2. Set de entrenamiento, de validación y de pruebas

Un algoritmo de *machine learning*, entonces, realiza predicciones en base a los datos que le fueron administrados. Sin embargo, para conocer la capacidad de un modelo de generalizar[4] éste debe ser aplicado a nuevos datos. Es por esto que el set de datos dispuestos suele dividirse en tres conjuntos diferentes (Sammut & Webb, 2010; Hastie et al., 2009; Sarang, 2020):

- Set de entrenamiento: datos utilizados para ajustar el modelo, son los datos a partir de los cuales el modelo aprende;
- Set de validación: datos utilizados para estimar los errores de las predicciones durante la selección del modelo (determinación de los mejores hiperparámetros durante el entrenamiento);
- Set de pruebas[5]: datos utilizados para evaluar el desempeño del modelo después del entrenamiento.

En la Figura 1.10 se muestra un diagrama del desarrollo de un sistema de aprendizaje automatizado supervisado utilizando un set de entrenamiento, un set de validación y

[3]Los ejemplos o *examples* en machine learning corresponden a un conjunto de características de un objeto o evento dispuesto para ser procesado por el sistema de machine learning.

[4]En *machine learning*, la generalización es la capacidad del modelo para dar salidas sensibles a conjuntos de entradas que no había visto antes.

[5]Cabe destacar que los términos "set de validación" y "set de pruebas" en algunos casos son utilizados indistintamente. En este trabajo, para dichos conjuntos de datos se consideran las definiciones entregadas en la Sección 1.3.1.2.

un set de pruebas. El modelo se entrena a partir de los datos del set de entrenamiento y se evalúa bajo el set de validación en un proceso iterativo. El modelo que presenta mejores resultados es el modelo óptimo que luego se evalúa con el set de pruebas.

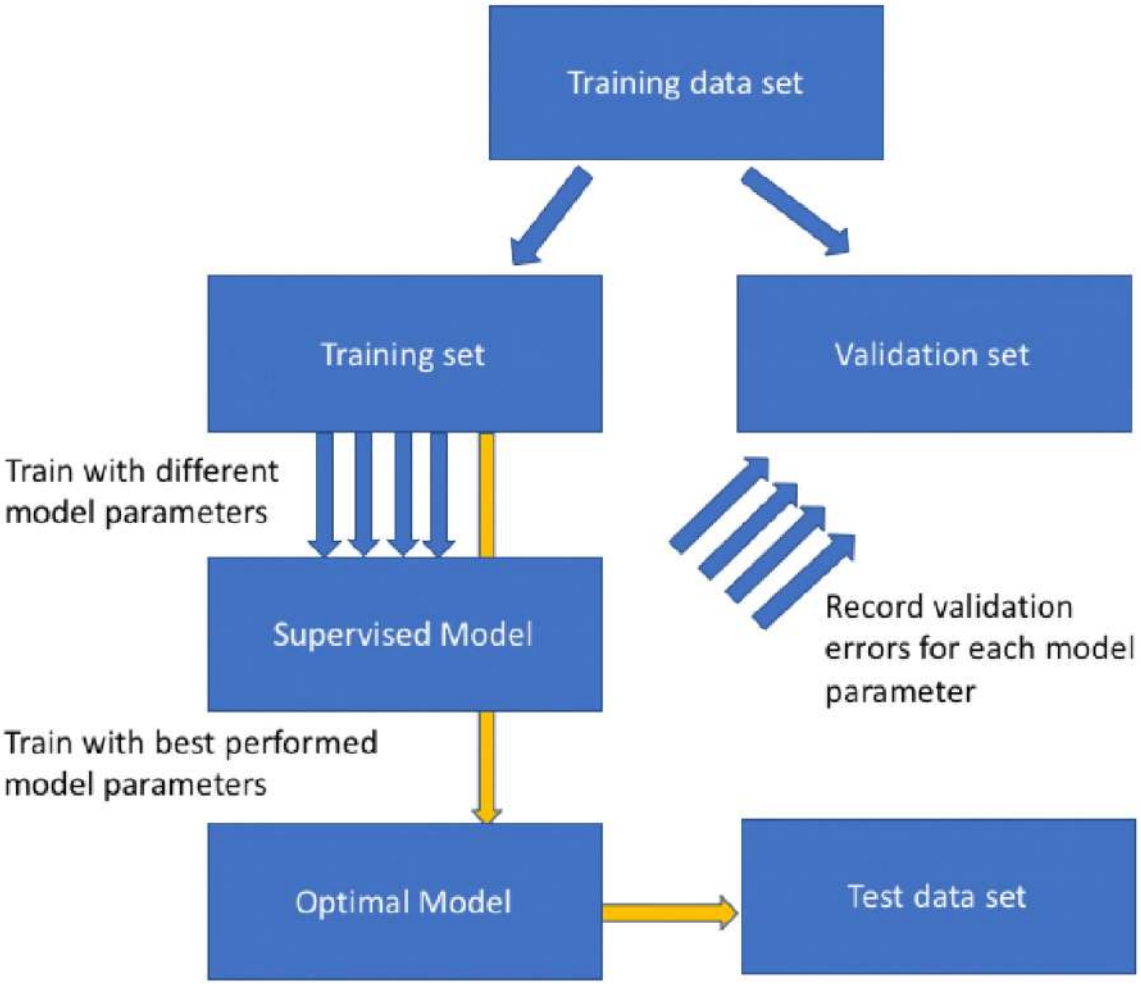


Figura 1.10: Diagrama de flujo para la selección de un modelo de aprendizaje automatizado supervisado. Las flechas azules indican el proceso de validación y las flechas amarillas indican el proceso de entrenamiento final y el proceso de prueba sobre el set de pruebas. Figura de Xu & Goodacre (2018).

1.3.1.3. Métricas o medidas de rendimiento

El rendimiento de un modelo se evalúa con diferentes métricas. Se utilizan diferentes métricas para distintos tipos de modelos y, en algunos casos, se suelen utilizar varias métricas de manera simultánea para evaluar un modelo, ya que una sola podría no dar suficiente información.

Las métricas más comunes para problemas de clasificación son:

- Matriz de confusión o matriz de error: forma de visualización de las predicciones de un modelo. Cada entrada de la matriz contiene el número de predicciones hechas por el modelo donde clasificó las clases correcta o incorrectamente. Un caso particular de la matriz de confusión es en la clasificación binaria, en donde una clase se asigna como positiva y otra como negativa. Como se muestra en la Figura 1.11 las celdas incluyen:
  - *True positives* o verdaderos positivos (TP): número de predicciones que el clasificador predice correctamente como positivas.
  - *True negatives* o verdaderos negativos (TN): número de predicciones que el clasificador predice correctamente como negativas.
  - *False positives* o falsos positivos (FP): número de predicciones que el clasificador predice incorrectamente como positivas.
  - *False negatives* o falsos negativos (FN): número de predicciones que el clasificador predice incorrectamente como negativas.

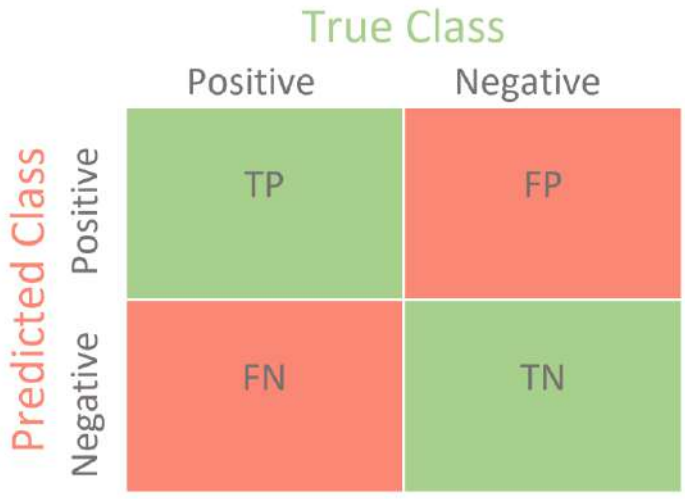


Figura 1.11: Matriz de confusión para clasificación binaria. Los elementos de la diagonal poseen las predicciones correctas, mientras que los elementos fuera de la diagonal son las muestras mal clasificadas (fuente: Mohajon, 2020).

- *Accuracy* o exactitud: corresponde al número de predicciones correctas dividido

por el total de predicciones.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.2}$$

- Precisión: fracción de predicciones positivas que son realmente positivas.

$$precision = \frac{TP}{TP + FP} \tag{1.3}$$

- *Sensitivity, recall, true positive rate*, sensibilidad o exhaustividad (TPR): fracción de muestras positivas correctamente identificadas por el modelo.

$$recall = \frac{TP}{TP + FN} \tag{1.4}$$

- *Specificity, true negative rate* o especificidad (TNR): fracción de muestras negativas correctamente identificadas por el modelo.

$$TNR = \frac{TN}{TN + FP} \tag{1.5}$$

- *F1-score*: media harmónica de la precisión y sensibilidad.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{1.6}$$

Nótese que en problemas de clasificación multiclase (con más de dos categorías) o donde simplemente no se tratan las clases como positiva o negativa, se utilizan en general las mismas métricas; precisión para cada clase y exhaustividad o *recall* para cada clase (en lugar se sensibilidad y selectividad).

A diferencia de las métricas utilizadas en los modelos de clasificación, los modelos de regresión deben ser capaces de trabajar con valores continuos. Las más comunes son:

- *Mean squared error* o error cuadrático medio (MSE):

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{1.7}$$

- *Mean absolute error* o error absoluto medio (MAE):

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{1.8}$$

donde, en ambos casos, m es el número total de ejemplo, y_i e $\hat{y}_i$ son las predicciones y los valores reales del i-ésimo ejemplo, respectivamente.

Generalmente se quiere conocer qué tan bien se desempeña el algoritmo de *machine learning* sobre datos que nunca haya visto antes, por esto las métricas se evalúan sobre el set de pruebas (Goodfellow et al., 2016).

1.3.1.4. Limitaciones

A pesar de que las técnicas de *machine learning* han sido transformadoras en diversas áreas, éstas poseen limitaciones producto de diferentes factores. Entre ellos se encuentran (Géron, 2017):

- Datos insuficientes para entrenar un algoritmo apropiadamente: se necesitan muchos datos para que la mayoría de los algoritmos de aprendizaje automático funcionen correctamente.
- Set de entrenamiento no representativo: generará un algoritmo incapaz de generalizar.
- Datos de entrenamiento de baja calidad: hacen que para el sistema sea más difícil identificar patrones.
- Set de entrenamiento sesgado: por ejemplo, un sesgo en el color de piel, género o cultura.
- Características irrelevantes: los datos de entrenamiento deben contener suficientes características relevantes y no demasiadas irrelevantes para poder aprender.
- Sobreajuste u *overfitting*: el modelo se ajusta muy bien a los datos de entrenamiento, ya que "ha aprendido mucho" de ellos, pero es débil al momento de generalizar; por lo que no es capaz de realizar inferencias (o predicciones) confiables cuando se le presentan nuevos datos.

- Subajuste o *underfitting*: el modelo es demasiado simple o "no ha aprendido lo suficiente" y por lo tanto es incapaz de generalizar.

En la Figura 1.12 se ilustra el problema de *overfitting* y *underfitting* en problemas de clasificación y regresión.

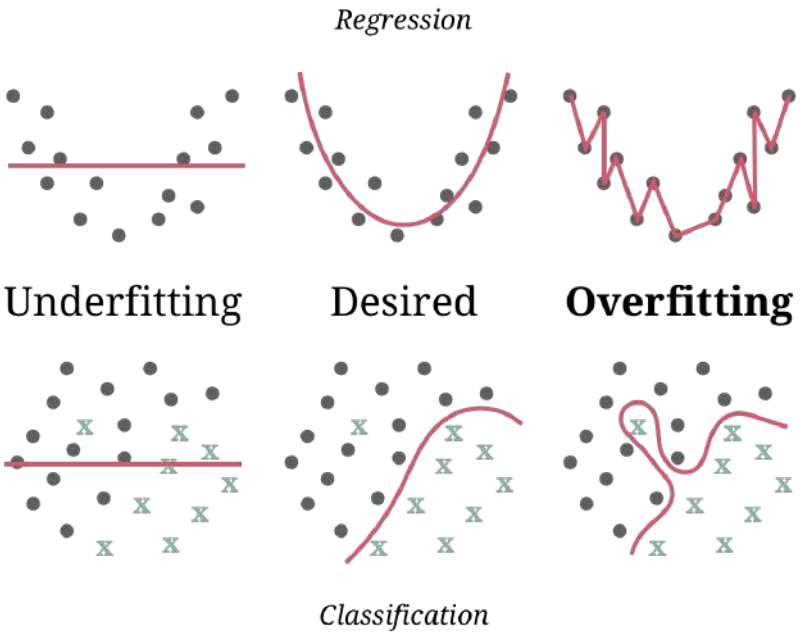


Figura 1.12: Representación de un modelo que produce un subajuste o *underfitting* (izquierda), sobreajuste u *overfitting* (derecha) y un buen ajuste en un problema de regresión (paneles superiores) y de clasificación (paneles inferiores). Los datos están representados por los puntos grises (y las cruces en el problema de clasificación) y los ajustes del modelo por las líneas rojas (fuente: The R Bootcamp, 2019).

1.3.2. Redes neuronales artificiales y Deep learning

El *deep learning* es un área dentro del *machine learning*, que escencialmente se basa en redes neuronales profundas (de 3 o más capas) (IBM Cloud Education, 2020). Las redes neuronales artificiales o *artificial neural networks* (ANNs), comúnmente referidas simplemente como "*neural networks*"(NNs), corresponden a un modelo computacional basado en las redes neuronales biológicas que constituyen el sistema nervioso de los animales (Sammut & Webb, 2010).

El cerebro humano y animal trabaja de una manera completamente diferente a la de

una computadora digital. El cerebro es una computadora o sistema de procesamiento de información altamente complejo, no lineal y paralelo; presenta la capacidad de organizar a las neuronas para realizar ciertos cálculos, ya sea reconocimiento de patrones, control motor, percepción, etc, a una velocidad mucho mayor que las computadoras digitales. En el caso de las redes neuronales, para lograr un buen rendimiento emplean una interconexión masiva de sus "células"denominadas neuronas o unidades de procesamiento. Por esto, se puede atribuir la siguiente definición de redes neuronales: "*Una red neuronal es un procesador distribuido masivamente en paralelo compuesto por unidades de procesamiento simples que tiene una propensión natural a almacenar conocimiento experiencial y ponerlo a disposición para su uso*"(Haykin et al., 1999).

Una red neuronal artificial toma los datos de entrada, los transforma calculando una suma ponderada sobre las entradas y aplica una función no lineal a esta transformación para calcular un estado intermedio. Estos pasos ocurren en cada capa, el componente de más alto nivel en una ANN. Los estados intermedios se utilizan como entrada en la siguiente capa. A partir de este proceso es que una ANN aprende, las salidas de cada capa ingresan a la siguiente hasta llegar a la capa de salida donde se genera una predicción. La primera y la última capa de una red se denominan capas de entrada y salida (*input layer* y *output layer*), respectivamente, y todas las capas intermedias se denominan capas ocultas o *hidden layers* (Dettmers, 2015).

Las ANNs fueron introducidas por Mcculloch & Pitts (1943), quienes presentaron un modelo computacional simplificado de cómo las neuronas biológicas podrían trabajar juntas en cerebros animales para realizar cálculos complejos usando lógica proposicional y planteando la primera arquitectura de una red neuronal artificial.

Rosenblatt (1957) inventó el *perceptrón*, una arquitectura de ANNs basada en unidades denominadas *threshold logic unit* (TLU) que puede utilizarse para clasificaciones binarias. Un perceptrón está compuesto por una capa de TLUs totalmente conectada[6] a las unidades de entrada. Configurar múltiples capas de perceptrones abre la posibilidad de resolver problemas más complejos. Estas ANNs se llaman perceptrones multicapas o *multilayer perceptrons* (MLPs). Los MLPs son un tipo de *feedforward neural network* (FNN), que son aquellas en que el flujo de información va en una sola dirección (Zell, 1997).

[6]Una capa se dice que está totalmente conectada (*fully conected*) o que es una capa densa cuando todas sus neuronas están conectadas a cada neurona de la capa anterior (Géron, 2017).

Rumelhart et al. (1985) en su artículo, presentan un entrenamiento de algoritmos con retropropagación o *backpropagation*, el que hoy se conoce como gradiente descendente o *gradient descent*. Este algoritmo sigue los siguientes pasos:

- *Forward pass* o predicción: las características de entrada se pasan a la capa de entrada de la red y son enviadas a la primera capa oculta. El algoritmo calcula la salida de cada neurona para cada ejemplo y pasa los resultados a la siguiente capa. Este proceso se repite para cada una de las capas ocultas hasta llegar a la capa de salida con una predicción. Todos los resultados intermedios son conservados.
- Cómputo de errores: el algoritmo mide el error de salida comparando las predicciones con los valores reales mediante una función de error.
- *Backward pass*: el algoritmo determina las contribuciones de error provenientes de cada conexión en la capa anterior, y así sucesivamente hasta llegar a la capa de entrada. Dicho de otra manera, este paso va propagando el gradiente de error hacia atrás a través de la red.
- *Gradient descent*: se ajustan o actualizan todos los pesos de la red para reducir el error y que por lo tanto, las predicciones sean más precisas.

Estos pasos se repiten en un proceso iterativo, ya sea un número predeterminado de veces o hasta que la red converge en una solución esperada. El número de pasos por el set de entrenamiento que el algoritmo ha completado es denominado como número de épocas (Géron, 2017).

1.3.2.1. Funciones de activación

Las funciones de activación son aquellas que operan en cada unidad, generando una transformación no lineal en ellos(Dettmers, 2015). Al ocupar *backpropagation*, es necesario que las funciones que trabajan en cada neurona tengan derivadas distintas de cero para poder llevar a cabo el paso de *gradient descent*. Las funciones de activación más comúnmente utilizadas son:

- Función logística o sigmoide:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1.9}$$

- Tangente hiperbólica:

$$\sigma(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{1.10}$$

- *Rectified Linear Unit*, ReLU:

$$\sigma(z) = max(0, z) \tag{1.11}$$

En la Figura 1.13 se muestran estas funciones de activación.

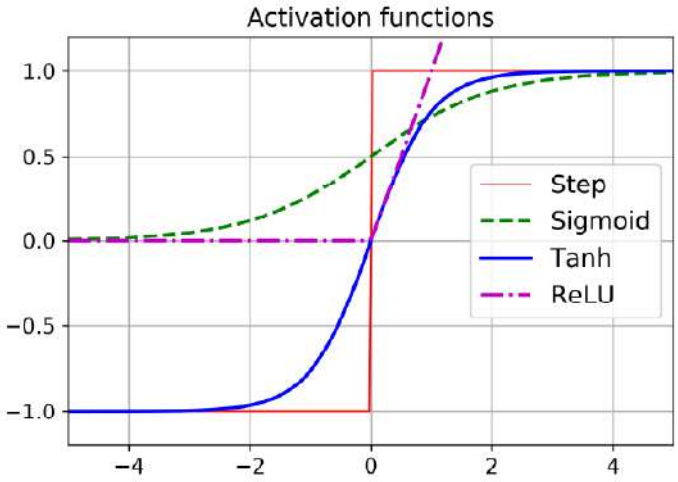


Figura 1.13: Funciones de activación más comúnmente utilizadas en sistemas de *deep learning* (fuente: Géron, 2017).

1.3.2.2. Arquitectura de una red neuronal artificial

La arquitectura de las redes neuronales artificiales cuenta con una serie de ajustes que controlan el funcionamiento del algoritmo y que son fijados previos al entrenamiento y se mantienen constante durante este proceso. Estos ajustes se conocen como hiperparámetros (Goodfellow et al., 2016). En la Tabla 1.2 se muestran los hiperparámetros utilizados en las ANNs y los valores que suelen tomar.

En la Figura 1.14 se muestra un esquema de la arquitectura de una red neuronal artificial de 4 capas totalmente conectadas (la capa de entrada no se considera en el número total de capas de una NN).

Tabla 1.2: Resumen de la arquitectura de una ANN.

Hiperparámetro	Valores típicos
N° de neuronas en la capa de entrada	N° de características de entrada
N° de capas ocultas	Variable
N° de neuronas por capa oculta	Variable
N° de neuronas en la capa de salida	N° de elementos a predecir
Funciones de activación de las capas ocultas	ReLU, tanh, sigmoide u otras. Pueden variar de capa en capa
Función de activación de la capa de salida	Ninguna, ReLU, tanh, softmax u otras. Depende del problema
Función de pérdida	MAE, MSE, cross-entropy u otras. Depende del problema

Las salidas u *outputs* de una capa totalmente conectada se suelen denotar por:

$$\begin{aligned} A^{[l]} &= g^{[l]}(W^{[l]}A^{[l-1]} + b^{[l]}) \\ A^{[0]} &= X \\ A^{[L]} &= \hat{y} \end{aligned} \tag{1.12}$$

donde A es el vector de salida, g la función de activación, W es la matriz de los pesos y b el vector de bias. El superíndice $[l]$ indica que los valores o vectores están asociados a la capa l. $[0]$ es la capa de entrada o *input layer* y $[L]$ es la capa de salida u *output layer*. X es la matriz de las características de entrada o *input features* e $\hat{y}$ son las predicciones.

Una ANN con muchas capas ocultas se denomina red neuronal profunda o *deep neural network* (DNN). El campo del *deep learning* estudia las DNN y otros modelos que incluyen computaciones profundas (Géron, 2017).

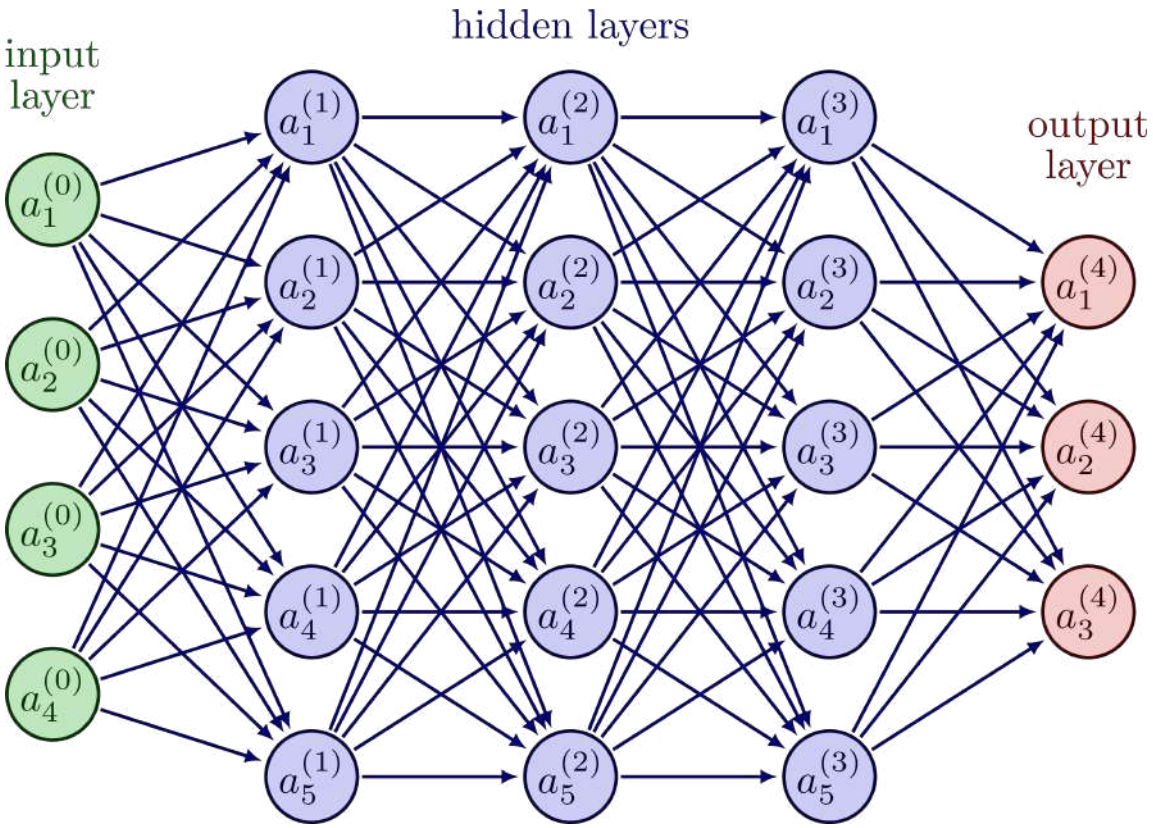


Figura 1.14: Representación de la arquitectura de una red neuronal artificial. Esta red tiene 4 capas totalmente conectadas. Tiene 3 capas ocultas, cada una con 5 unidades o neuronas. La capa de entrada tiene 4 unidades y la capa de salida tiene 3.

1.3.2.3. Aplicaciones

El desarrollo del *deep learning* y las NNs durante las últimas décadas se ha visto fortalecido principalmente por la gran cantidad de datos disponibles para entrenarlas y la mejora de las capacidades de las computadoras desde la década de 1990 hasta la actualidad (Géron, 2017).

Actualmente, las tecnologías de aprendizaje automático y aprendizaje profundo impulsan muchos aspectos de la sociedad moderna: reconocimiento de voz (*e.g. Apple Siri, Amazon Alexa, Microsoft Cortana, Google Now*), procesamiento del lenguaje (*e.g. Google Translate*), sistemas de recomendación de contenido (*e.g. Spotify, Netflix, Twitter, Instagram*), productos o servicios (*e.g.* anuncios publicitarios) de interés del usuario, reconocimiento facial (*e.g. Facebook*), asistentes de conducción (*e.g. Google Maps, Uber*), vehículos autómonos (*e.g. Tesla*), descubrimiento de fármacos y toxicología, bioinformática, análisis de imágenes y diagnósticos médicos, detección de fraudes financieros y muchos más (LeCun et al., 2015; Balas et al., 2019).

1.4. Avances en casos astrofísicos con algoritmos de machine learning

Como se mencionó anteriormente, los conjuntos de datos astronómicos están experimentando un rápido aumento en tamaño y complejidad, lo que introduce a la astronomía en la era del *big data*. En vista de este acelerado crecimiento es que en el ámbito de la astronomía se han empezado a desarrollar herramientas automatizadas para detectar, clasificar, caracterizar y analizar objetos o conjuntos de objetos astronómicos. Los algoritmos automatizados han tenido una creciente popularidad durante los últimos años en la astronomía y ya han sido implementados en una gran variedad de tareas (Baron, 2019). Algunos ejemplos de estudios llevados acabo con técnicas de *deep learning* que incluyen la implementación de redes neuronales en astronomía son: Yip et al. (2019) en la detección de exoplanetas, Soumagnac et al. (2015) en la separación entre estrellas y galaxias, Walmsley et al. (2020) en la clasificación de galaxias con redes neuronales bayesianas, Lochner et al. (2016) y Mahabal et al. (2017) en la clasificación fotométrica de supernovas y de curvas de luz, respec-

tivamente, Das & Sanders (2019) en la determinación de distancias, masas y edades espectroscópicas de estrellas gigantes rojas, Bilicki et al. (2018) en la determinación de *redshifts*, el estudio cosmológico por medio de redes neuronales convolusionales de Fluri et al. (2019), y muchos más.

Por otro lado, y por pertinencia con esta tesis, la determinación de metalicidades estelares también ha sido desarrollada por medio de técnicas de *machine learning*.

En este sentido, hay que considerar que las metalicidades de estrellas han sido históricamente determinadas por medio de observaciones espectroscópicas que permiten medición y posterior modelamiento de las líneas espectrales. Sin embargo, la obtención de espectros estelares es una tarea que demanda mucho tiempo en telescopios grandes, además de un análisis más complejo. Es por esto que surge la necesidad de contar con métodos alternativos para la determinación de metalicidades estelares. Otros método para determinar las metalicidades estelares es mediante el uso de índices fotométricos. Estos índices fotométricos miden el flujo proveniente en la región espectral de una línea por fotometría con un filtro de banda angosta; para posteriormente estimar la intensidad de la línea comparando la intensidad del continuo espectral estimado por fotometría con un filtro de banda ancha. Posteriormente, los índices son calibrados utilizando estrellas con parámetros estelares conocidos por medio de espectros y aplicados a la muestra que quiera estudiarse (e.g., Geisler, 1986). Si bien la precisión de los métodos fotométricos es menor, eso se compensa con la gran cantidad de objetos que pueden estudiarse a la vez y que la precisión es suficiente para realizar diversos estudios.

En los últimos años se han implementado técnicas de *machine learning* para determinar metalicidades estelares fotométricas. Por ejemplo, Miller (2015) desarrolló un método para determinar metalicidades estelares a partir de los filtros *ugriz* de SDSS por medio de distintos algoritmos de aprendizaje automatizado (*K-nearest neightbours*, *random forest* y *support vector machines*). Este método determinó metalicidades estelares con un error cuadrático medio ~ 0.29 dex en relación a las mediciones espectroscópicas. En otra contribución, Thomas et al. (2019) presentaron un algoritmo que, a partir de la información fotométrica de Canada-France Imaging Survey (CFIS), Pan-STARRS 1 (PS1) y Gaia, es capaz de discriminar estrellas entre enanas y gigantes y determinar sus distancias y metalicidades. Los algoritmos que desarrollaron logran identificar más del 70 % de las estrellas gigantes en el set de

entrenamiento y las estimaciones de metalicidades tienen incertidumbres de $\sim$ 0.2 dex para [Fe/H] < -1.2. Whitten et al. (2019b), por su parte, desarrollaron una metodología para determinar los parámetros estelares T_{eff} y [Fe/H] por medio de redes neuronales a partir de la fotometría del survey J-PLUS, con un particular énfasis en estrellas de baja metalicidad. Las estimaciones de metalicidades y temperaturas estelares mediante este método tienen una dispersión de $\sim$ 0.25 dex y $\sim$ 91 K, respectivamente.

Capítulo 2

Objetivos

El objetivo general de esta tesis es crear una herramienta capaz de derivar metalicidades estelares de estrellas gigantes rojas mediante el entrenamiento de un algoritmo de regresión basado en aprendizaje automatizado, utilizando la fotometría de S-PLUS, que a su vez deberá clasificar las estrellas según gravedad superficial entre estrellas gigantes y enanas. El enfoque en las estrellas gigantes rojas permitirá utilizar los resultados de este trabajo para trazar subestructuras en la galaxia.

Objetivos específicos:

- Generar un catálogo de estrellas curado (i.e., limpio) para el entrenamiento, validación y pruebas de la herramienta.
- Crear la arquitectura del algoritmo de clasificación para la discriminación entre estrellas gigantes y enanas.
- Crear la arquitectura del algoritmo de regresión para la derivación de metalicidades a partir de la información fotométrica.
- Investigar las características espectrales de S-PLUS capaces de discriminar entre estrellas gigantes/enanas y la derivación precisa de metalicidades para gigantes rojas (y para enanas).
- Identificar estrellas rojas, discriminando entre gigantes y enanas y determinar las metalicidades de las estrellas gigantes rojas (y de las enanas).
- Demostrar que el modelo generado es capaz de producir metalicidades, en calidad y cantidad, que permitan determinar la distribución de metalicidades de

las estrellas gigantes rojas de S-PLUS.

- Realizar una inspección de la distribución obtenida de las estrellas que fueron clasificadas como gigantes rojas.
- Comparar con resultados obtenidos de otros autores.

Capítulo 3

Datos y Metodología

En este capítulo se presenta el conjuntos de datos utilizado en la presente tesis y la metodología seguida. En este trabajo se ocupan dos conjuntos de datos: el primero es aquel que se utiliza como set de entrenamiento, de validación y de pruebas en los algoritmos desarrollados que incluye información fotométrica de S-PLUS e información espectroscópica de APOGEE. El segundo consiste en el catálogo completo de estrellas de S-PLUS tras aplicar ciertos cortes de calidad.

La metodología seguida en esta tesis de forma resumida contempló los siguientes pasos:

- Creación del catálogo de entrenamiento, de validación y de pruebas para utilizar en los algoritmos diseñados, además de aplicar cortes de calidad a los catálogos de S-PLUS.
- Construcción, entrenamiento y evaluación de algoritmos capaces de discriminar estrellas entre gigantes y enanas a partir de índices fotométricos. Se desarrollaron redes neuronales artificiales y algoritmos de *random forest* para este propósito.
- Construcción, entrenamiento y evaluación de algoritmos capaces de determinar metalicidades estelares a partir de índices fotométricos. Se desarrollaron redes neuronales artificiales por separado para estrellas gigantes, enanas y para todo el conjunto, esperando que al trabajar con gigantes y enanas por separado se obtengan mejores resultados.
- Clasificación y determinación de metalicidades al catálogo completo de S-

PLUS.

Cada uno de estos pasos se describen detalladamente en las siguientes Secciones.

3.1. Set de datos

3.1.1. Fotometría de S-PLUS

En este trabajo se utilizaron los catálogos fotométricos de S-PLUS Partial Internal Data Release 3 versión n4 (S-PLUS iDR3-n4), obtenidos a partir de datos compartidos por la colaboración de S-PLUS. Los catálogos del iDR3-n4 incluyen observaciones en un área del cielo correspondiente a 2,214 $\deg^2$ distribuídos en 1,107 campos. Estos campos, a su vez, se encuentran distribuidos en 11 subregiones con RA $< 10°$ definidas para su calibración (ver Figura 1.9[1]). Estas subregiones, junto con el número de campos y de fuentes en cada una, se encuentran en la Tabla 3.1.

A continuación se describen los datos contenidos en S-PLUS iDR3-n4.

Los catálogos fotométricos multibanda son generados por medio del software SExtractor (Bertin, E. & Arnouts, S., 1996) en una imagen reducida combinada, correspondiente a la suma de las bandas g, r, i, z.

Para cada banda, la fotometría se reporta para seis tipos de aperturas definidas de la siguiente manera:

- `AUTO`: Apertura con escala adaptativa. Derivada del algoritmo de 'primer momento' de Kron. Mejor magnitud total para fuentes extendidas;
- `PETRO`: Apertura con escala adaptativa elíptica. Derivada a partir del estimador fotométrico Petrosiano. Mejor magnitud para la estimación de propiedades físicas de fuentes extendidas;
- `ISO`: Magnitudes isofotales, con el área isofotal definida como el número de píxeles con valores que exceden el umbral de 3 sigma respecto al *background*;

[1]En la Figura 1.9 se señala que S-PLUS iDR3 cuenta con 1,117 campos en 2,108.5 $\deg^2$, pero la versión n4 cuenta con 1,107 campos en 2,214 $\deg^2$.

Tabla 3.1: Cantidad de datos en cada subregión de S-PLUS iDR3-n4.

Subregión	N° de campos	N° total de fuentes	N° de fuentes con `CLASS_STAR` > 0.92
STRIPE82	170	10,973,634	1,564,747
MC[a]	149	-	-
Hydra	160	15,612,384	4,317,901
MAIN3.1	120	3,668,201	344,355
MAIN3.2	124	6,456,155	491,273
MAIN3.3	64	4,258,948	694,253
MAIN3.4	27	1,398,195	174,276
MAIN3.5	124	7,392,950	1,036,853
MAIN3.6	84	4,803,662	549,531
MAIN3.7	44	2,392,837	194,402
MAIN3.8	41	2,270,136	279,404
ABELL1644[b]	-	511,253	64,494
Total	1,107	59,738,355	9,711,489

[a] En la versión n4 no se incluyen los campos en las Nubes de Magallanes.
[b] ABELL1644 no es una subregión en sí, pero viene incluída en un catálogo separado. Por esto se considera dentro de esta lista.

- `APER_3`: Apertura circular fija de 3 arcosegundos de diámetro. Ideal para calibraciones y comparaciones con otros datos;
- `APER_6`: Apertura circular fija de 6 arcosegundos de diámetro. Ideal para comparaciones con otros datos;
- `PSTOTAL`: Apertura circular fija de 3 arcosegundos de diámetro que incluye corrección de apertura. Mejor magnitud para la estimación de propiedades físicas de fuentes puntuales.

Información adicional de estos tipos de apertura se encuentra en el sitio de SExtractor (Bertin, E. & Arnouts, S., 1996).

Para este trabajo se utilizaron las magnitudes `PStotal`, ya que son las que mejor representan la magnitud total de las fuentes puntuales (Almeida-Fernandes et al., 2021). Cabe destacar que los catálogos fotométricos vienen separados en archivos diferentes para cada campo. En ellos se entregan las magnitudes con las correcciones de *zero point* y donde las magnitudes de las fuentes no detectadas están establecidas como $m = 99.00$.

La astrometría de S-PLUS iDR3-n4 está vinculada a las posiciones de las estrellas en 2MASS y las coordenadas ecuatoriales RA y DEC están en la época J2000.

Los catálogos también incluyen las *flags* de SExtractor que indican la posible contaminación por fuentes vecinas, píxeles saturados, objetos truncados, etc. Estas flags están incluídas tanto para la imagen de detección como para cada filtro (`PhotoFlagDet` y `PhotoFlag_[Filter]`[2], respectivamente) y contienen 8 bits con información relevante sobre el proceso de extracción de la fuente, en orden de importancia:

- 1: El objeto tiene vecinos brillantes y cercanos que producen bias en la fotometría o píxeles malos;
- 2: El objeto estaba originalmente 'mezclado' con otro;
- 4: Al menos un píxel del objeto está saturado;
- 8: El objeto está truncado (cerca del límite de la imagen);
- 16: La apertura del objeto está incompleta o corrompida;

[2]El demarcador `[Filter]` en el nombre de un parámetro indica que éste existe para los 12 filtros del sistema de Javalambre.

- 32: Los datos isofotales del objeto están incompletos o corruptos;
- 64: Se produjo un *overflow* de memoria durante el *deblending*;
- 128: Se produjo un *overflow* de memoria durante la extracción.

En cuanto a la discriminación de las fuentes entre estrellas y galaxias, S-PLUS iDR3-n4 incluye el parámetro `CLASS_STAR` de SExtractor. La clasificación se basa en una red neuronal consistente de un perceptrón multicapa, entrenada mediante aprendizaje supervisado para estimar la probabilidad de que la detección de SExtractor sea una fuente puntual o un objeto extendido basado en su morfología. El output de la red corresponde a un 'índice de estelaridad' de valores entre 0 y 1, donde 0 significa que el objeto es más probablemente una galaxia y 1 que es más probable que se trate de una estrella (Bertin, E. & Arnouts, S., 1996).

En este trabajo se consideraros las siguientes columnas de los catálogos de S-PLUS iDR3-n4:

- `Field`: Nombre del campo de observación;
- `ID`: Identificador del objeto propio de S-PLUS iDR3;
- `RA`: Ascensión Recta en J2000 (grados);
- `DEC`: Declinación en J2000 (grados);
- `CLASS_STAR`: Clasificador estrella/galaxia en la imagen de detección;
- `PhotoFlag_[Filter]`: Flag fotométrica de la fuente en la imagen de cada filtro;
- `[Filter]_PSTOTAL`: Magnitud PSTOTAL AB para cada filtro;
- `e_[Filter]_PSTOTAL`: Error en la magnitud PSTOTAL AB para cada filtro;
- `s2n_[Filter]_PSTOTAL`: Señal ruido de la magnitud PSTOTAL en cada banda, definidos como el flujo dividido por el error de este.

La condición que se adoptó para considerar las fuentes de los catálogos como estrellas fue `CLASS_STAR` > 0.92. Este es el valor que utilizaron Almeida-Fernandes et al. (2021); Whitten et al. (2019a) para considerar las estrellas en el survey S-PLUS. Los números de fuentes que cumplen esta condición para cada catálogo se muestran en la Tabla 3.1.

Los catálogos fotométricos de S-PLUS iDR3-n4 no presentan correcciones por extinción del medio interestelar. A continuación se detalla el procedimiento para realizar dicha corrección.

3.1.1.1. Corrección por extinción interestelar

La extinción es la absorción y el *scattering* de la radiación electromagnética producida por los granos de polvo compuestos por elementos pesados originados en las estrellas y liberados al medio interestelar mediante explosiones o vientos estelares. La extinción hace que los objetos se vean más rojos de lo que realmente son, fenómeno llamado enrojecimiento o *reddening* interestelar, cuyos efectos en diferentes longitudes de onda vienen dados en las leyes de extinción (Draine, 2003; Schlafly & Finkbeiner, 2011).

El enrojecimiento se traduce cuantitativamente como el exceso de color, definido como la diferencia entre el índice de color observado de un objeto y su índice de color intrínseco (denotado generalmente con un subíndice 0). Este último corresponde al valor teórico del índice de color si es que la radiación electromagnética del objeto no se viera afectada por la extinción del medio interestelar. El exceso de color (para el color $B-V$, que es el color más comúnmente utilizado históricamente), entonces viene dado por:

$$E(B-V) = (B-V)_{obs} - (B-V)_0 \tag{3.1}$$

La extinción en una banda χ, A_χ, usualmente se estima de la siguiente forma:

$$A_\chi = R_\chi * E(B-V) \tag{3.2}$$

donde R_χ es la extinción en la banda χ relativa a $E(B-V)$ (*total-to-selective extinction ratio*). Para la Vía Láctea se tiene un valor típico de $R_V = 3.1$ (Schultz & Wiemer, 1975; Cardelli et al., 1989), aunque este número varía según la línea de visión.

Las correcciones por extinción, es decir, las magnitudes fotométricas desenrojecidas

de S-PLUS se obtuvieron de la siguiente forma:

$$\chi_0 = \chi - \kappa_\chi * E(B-V) \tag{3.3}$$

donde χ_0 son las magnitudes desenrojecidas en la banda χ, χ las magnitudes observadas (sin desenrojecer), R_χ el coeficiente de extinción en la banda χ y $E(B-V)$ los excesos de color. Se usaron $E(B-V)$ del mapa de polvo de Schlegel et al. (1998) (SFD98) obtenidos mediante el package *dustmaps* (Green, 2018), aplicando luego la recalibración indicada por Schlafly & Finkbeiner (2011) (SF11), de manera que:

$$E(B-V)_{SF11} = 0.86 * E(B-V)_{SFD98} \tag{3.4}$$

La limitación es que estos valores de $E(B-V)$ están obtenidos de un mapa de polvo bidimensional, por lo tanto no se toman en cuenta las distancias a las fuentes, sino que suponen la contribución del polvo en la dirección dada. Sin embargo, los mapas tridimensionales disponibles están limitados a ciertas declinaciones y su precisión tiene una alta dependencia a las estimaciones de distancias, por lo que pueden estar sujetos a grandes incertezas. Además, no es prioridad contar con mapas tridimensionales mientras se utilizan con datos a altas latitudes galácticas, como es el caso de este trabajo, Por estas razones, y para tener la mayor cantidad de datos desenrojecidos de forma homogénea, es que se utilizaron los $E(B-V)_{SF11}$.

Para los coeficientes de extinción, se utilizaron los valores determinados por López-Sanjuan et al. (2021) determinados para cada uno de los filtros del survey J-PLUS usando la técnica de pares de estrellas de Yuan et al. (2013) (ver Tabla 3.2). Coeficientes de extinción determinados para J-PLUS ya han sido utilizados para realizar correcciones por extinción a la fotometría de S-PLUS. Por ejemplo, en Whitten et al. (2021) utilizaron los coeficientes de extinción determinados por López-Sanjuan et al. (2019), también a partir de los datos de J-PLUS. Esto no supone un problema, ya que J-PLUS y S-PLUS cuentan con telescopios, detectores y sistema de filtros idénticos.

Tabla 3.2: Coeficientes de extinción para los filtros del sistema de Javalambre (fuente: López-Sanjuan et al., 2019).

Nombre del filtro	$\kappa_\chi = A_\chi \,/\, E(B-V)$	Comentarios
u	4.916	u Javalambre
F378	4.637	[OII]
F395	4.467	Ca H+K
F410	4.289	Hδ
F430	4.091	banda G
g	3.629	g SDSS
F515	3.325	triplete de Mgb
r	2.527	r SDSS
F660	2.317	Hα
i	1.825	i SDSS
F861	1.470	triplete de Ca
z	1.363	z SDSS

3.1.2. Creación del catálogo de entrenamiento, de validación y de pruebas

El set de datos que se utilizó para luego conformar el catálogo de entrenamiento, de validación y de pruebas fue construido mediante un *cross-match*[3] entre los catálogos fotométricos de S-PLUS iDR3-n4 y el catálogo de datos espectroscópicos de Apache Point Observatory Galactic Evolution Experiment (APOGEE) de Sloan Digital Sky Survey (SDSS) DR17 (Abdurro'uf et al., 2021), de aquí en adelante APOGEE DR17. Es importante destacar que la fotometría en las diferentes bandas de S-PLUS corresponden a las características de entrada o *input features* x y la APOGEE entrega las etiquedas o *labels* y.

APOGEE es un survey espectroscópico estelar de gran escala realizado en el infrarrojo cercano (específicamente en la banda H, en el rango $1.5\mu m < \lambda < 1.7\mu m$) que adquiere espectros de muy alta resolución de R $\sim$ 22,500. Es uno de los programas en SDSS-III y IV que cuenta con dos fases: APOGEE-1 (Majewski et al., 2017) y

[3]Identificación de mediciones en dos o más catálogos que corresponden a un mismo objeto, generalmente siendo identificadas mediante sus coordenadas: los objetos separados entre sí por una distancia menor a un valor arbitrario definido corresponden a una coincidencia.

APOGEE-2 (Majewski et al. in prep.). Sus espectrógrafos se encuentran ubicados en el Telescopio Sloan Foundation y el Telescopio New Mexico State University (NMSU) en Apache Point Observatory (APO) en Nuevo México, Estados Unidos (APOGEE-1 y APOGEE-2N) y el Telescopio Irénée du Pont de Las Campanas Observatory (LCO) en Atacama, Chile (APOGEE-2S).

Como los datos de SDSS DR17 no estuvieron disponibles públicamente hasta el día 6 de Diciembre de 2021, antes de esa fecha se construyó un set de datos compuesto por 9,658 estrellas producto del *cross-match* entre los catálogos de S-PLUS iDR3 n4 y las estrellas de APOGEE de SDSS DR16 (Ahumada et al., 2020). Este set de datos inicial permitió establecer la arquitectura para la manipulación de datos e ir probando diferentes algoritmos. Por esta razón, este set de datos previos obtenido a partir de SDSS DR16 no será tratado en mayor profundidad. Es importante considerar que APOGEE DR16 incluye por primera vez información de APOGEE-2S, aunque contiene relativamente pocos datos en el hemisferio Sur. Es por esto último que se esperó a la salida de SDSS DR17 para construir el set de datos que se utilizaría finalmente en los modelos de aprendizaje automatizado. APOGEE DR17 contiene más datos en el hemisferio Sur, por lo que permite elaborar un catálogo de entrenamiento, de validación y de pruebas más numeroso como resultado del *overlap* con el footprint de S-PLUS. Esto tendría un impacto positivo en los algoritmos, ya que las técnicas de deep learning se ven beneficiadas al contar con mayores cantidades de datos de entrenamiento.

APOGEE DR17 contiene aproximadamente 2,6 millones de espectros de visitas individuales con datos e información de 657,135 objetivos únicos, incluyendo todos los datos de SDSS-III/APOGEE y SDSS-IV/APOGEE-2. Los principales productos de datos son espectros por visitas individuales y por visitas combinadas, mediciones de velocidad radial, parámetros atmosféricos (entre ellos, temperatura efectiva, gravedad superficial y metalicidad) y abundancias de elementos individuales. Los parámetros estelares y las abundancias químicas son determinados a partir de la APOGEE Stellar Parameters and Chemical Abundance Pipeline (ASPCAP, García Pérez et al. 2016)[4]. La distribución de los campos pertenecientes al survey se muestran en la Figura 3.1.

En este trabajo se usó el archivo allStar, específicamente `allStardr17synspec_rev1.fits`,

[4]Información adicional del catálogo se encuenta en el sitio de SDSS Datamodel: allStar.

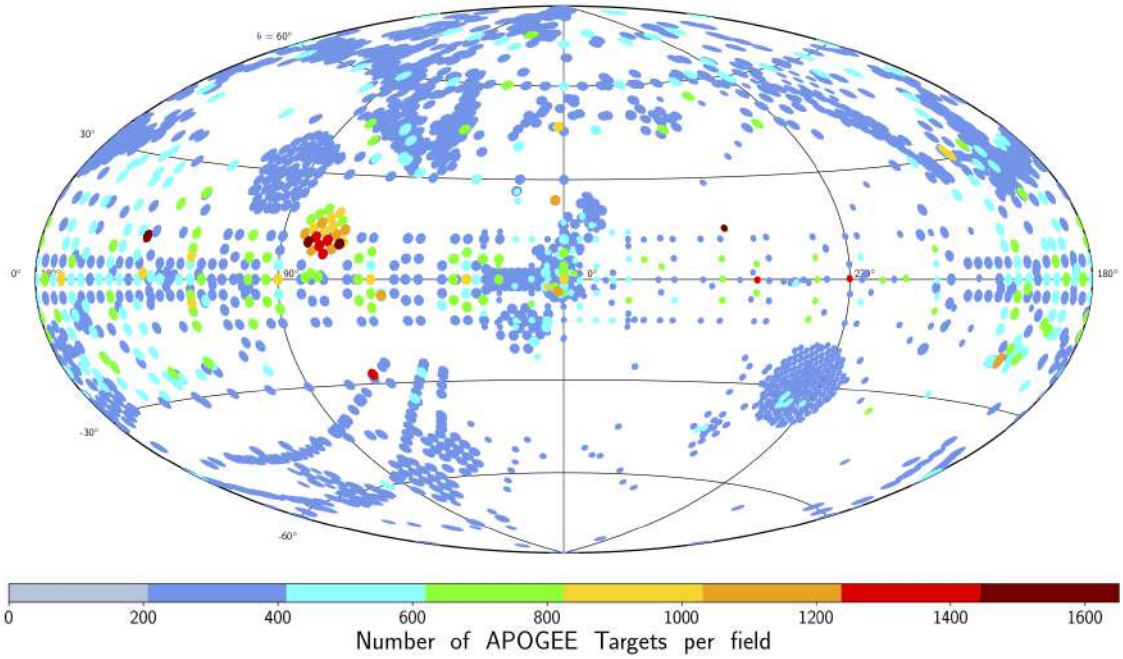


Figura 3.1: Distribución de los campos de APOGEE en el DR17 presentado en coordenadas Galácticas. Los colores indican el número de objetivos observados por campo (fuente: Abdurro'uf et al. 2021).

que es el catálogo que contiene los flujos de cada visita a un objeto combinados en un único espectro, esto implica que la señal ruido de dicho espectro final es más alta y permite obtener resultados más precisos, además de tener una entrada por cada estrella única. Este archivo entrega la información mencionada anteriormente: velocidad radial, parámetros atmosféricos y abundancias de elementos individuales. En concreto se utilizaron para el catálogo de entrenamiento, de validación y de pruebas las columnas con los parámetros estelares `TEFF`, `LOGG` y `M_H`, estos son, la temperatura efectiva en Kelvin, el logaritmo de la gravedad superficial en cgs y la metalicidad en dex. En este punto es importante mencionar que [M/H] es la metalicidad total ajustada al espectro completo y por lo tanto tiene contribución de varios metales, mientras que la cantidad [Fe/H] es una medición directa de las líneas de Fe. Aunque las diferencias entre ambos valores pueden ser muy pequeñas, en este trabajo se utiliza el parámetro [M/H], ya que es más representativo de la metalicidad total de la estrella.

El *cross-match* entre los catálogos de S-PLUS iDR3-n4 y APOGEE DR17 fue realizado utilizando Astropy (Astropy Collaboration et al., 2013, 2018), seleccionando una

separación máxima de 3.0 arcosegundos entre las posiciones en coordenadas ecuatoriales, ascensión recta (RA) y declinación (DEC), de los objetos. Cabe destacar que dentro de este rango se búsqueda para cada estrellas se encontró una única coincidencia. Resultaron 22,871 estrellas con $0.02 < \log g < 5.39$, $-2.45 < [\mathrm{M/H}] < 0.50$ y $3{,}160 < T_{eff} < 13{,}961$. En Abdurro'uf et al. (2021) se indica que para estrellas enanas frías ($T_{eff} < 3{,}500$ K) y para las estrellas calientes ($T_{eff} > 7{,}000$ K) se presentan problemas sistemáticos en la determinación de los parámetros estelares. Por esta razón, es que para el set de datos que conforma el catálogo de entrenamiento, de validación y de pruebas se consideraron sólo las estrellas dentro del rango de temperaturas efectivas entre 3,500 y 7,000 K, quedando así 21,148 estrellas. De estas 21,148 estrellas se descartaron 288 que no poseen información de metalicidad. También se eliminaron las entradas que tienen magnitudes en los filtros de S-PLUS iguales a 99.00 (ya que las magnitudes en todos los filtros se utilizaron en el desarrollo de los algoritmos en un principio), pues este es el valor asignado a fuentes no detectadas, obteniéndose una muestra de 20,735 estrellas. De esta muestra se seleccionaron las estrellas que tuvieran una señal ruido `s2n_[Filter]_PSTOTAL` > 50 en todas las bandas de S-PLUS para definir el set de datos que compone los catálogos de entrenamiento, de validación y de pruebas. Adicionalmente, se inspeccionaron las magnitudes límites que alcanza esta muestra, y a partir de esta información de descartó una estrella con F378 (mag) ~ 35. Esta estrella se descartó.

El conjunto de datos contiene 18,062 estrellas. En la Figura 3.2 se muestra la distribución de estos datos en el plano ($\log g$, T_{eff}) o diagrama de Kiel, incluyendo información de las metalicidades. En la Figura 3.3 se muestra un diagrama pareado que enseña las relaciones entre los parámetros T_{eff}, $\log g$ y [M/H]. Los errores típicos de estos parámetros son $\sigma_{Teff} \sim 26$, $\sigma_{logg} \sim 0.03$ y $\sigma_{[M/H]} \sim 0.01$. Adicionalmente, la distribución espacial del catálogo en coordenadas Galácticas se presenta en la Figura 3.4, aquí se observa que la mayoría de las estrellas están ubicadas en la zona del Stripe 82. Esta muestra tiene una magnitud límite r (mag) < 16.

Cabe destacar que no se realizaron cortes por señal ruido de APOGEE DR17, ya que se inspeccionaron los errores de los parámetros estelares en función de la señal ruido, e incluso en señales ruido bajas (*i.e.* SNR < 50) los errores típicos eran de $\sigma_{Teff} \sim 80$, $\sigma_{logg} \sim 0.09$ y $\sigma_{[M/H]} \sim 0.03$, errores mucho menores a los que se espera obtener por medio de un método fotométrico, incluso basado en redes neuronales. Es por esto que se decidió no perder estos datos que representan un $\sim 7\,\%$ de la

muestra.

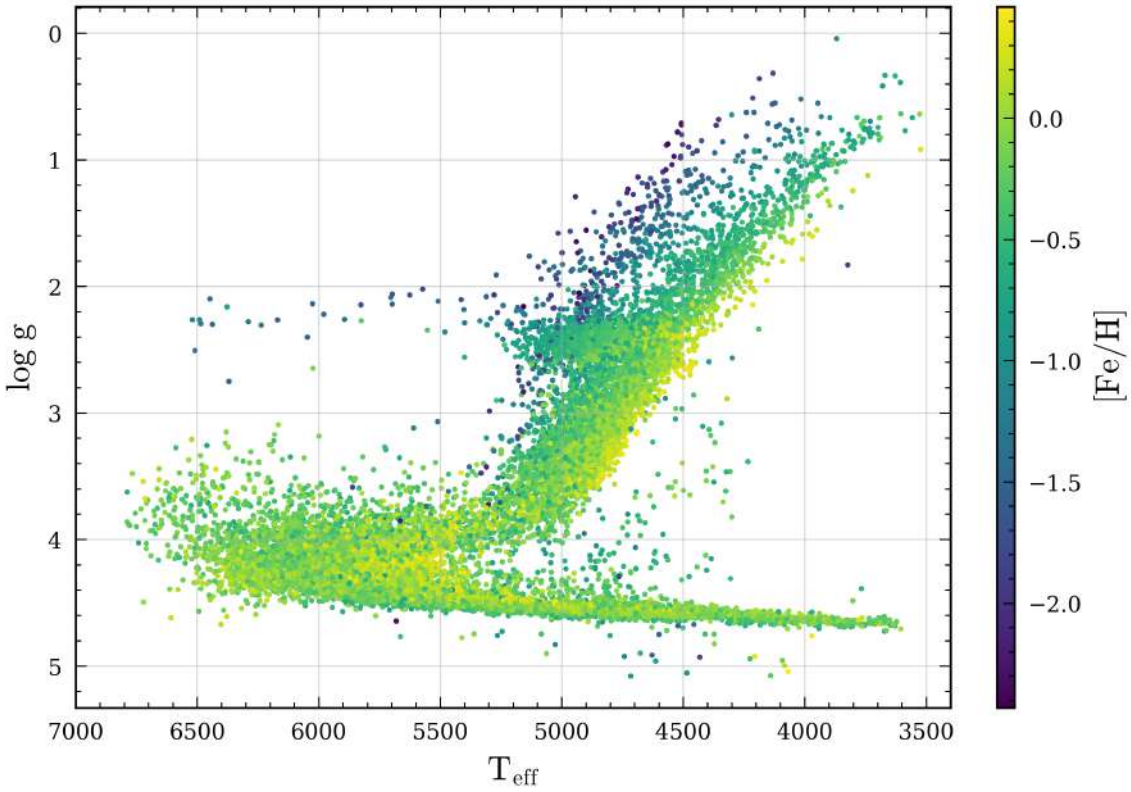


Figura 3.2: Distribución de las 18,062 estrellas resultantes del cross-match entre S-PLUS iDR3 y APOGEE DR17 para los catálogos de entrenamiento, de validación y de pruebas en el diagrama de Kiel. El mapa de color representa las metalicidades.

Este set de datos de 18,062 estrellas se dividió entre el catálogo de entrenamiento, de validación y de pruebas, como se indica en la Tabla 3.3. La selección de las estrellas que pertenecen a cada catálogo fue hecha de manera aleatoria y las proporciones fueron elegidas según lo típico para este tipo de problemas (Géron, 2017).

3.1.2.1. Creación de labels para la separación de estrellas gigantes y enanas en el set de datos

Para implementar un algoritmo de clasificación basado en aprendizaje supervisado (que sea capaz de discriminar estrellas entre gigantes y enanas a partir de colores fotométricos) es necesario contar con *labels* que indiquen a qué clase o categoría pertenece cada objeto. Por esto, es necesario contar con algún criterio que identifique cada estrella del set de datos como gigante o enana, y que registre en una columna

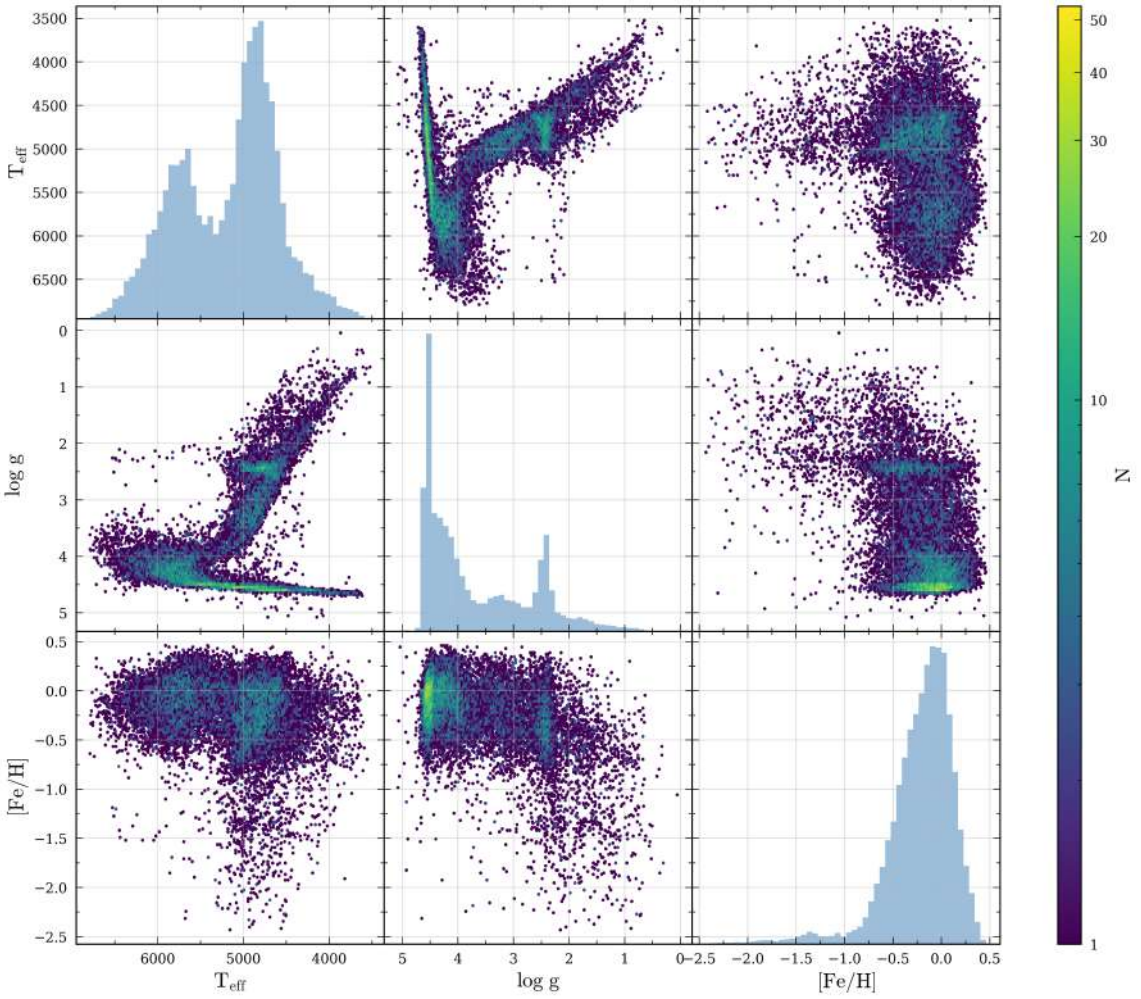


Figura 3.3: Diagrama pareado que muestra las relaciones y distribuciones de los parámetros T_{eff}, log g y [M/H] de las estrellas del set de entrenamiento, de validación y de pruebas. Las gráficas en la diagonal muestran las distribuciones de dichos parámetros, mientras que las gráficas fuera de la diagonal muestran las relaciones entre dichos parámetros. Las tres gráficas de la parte inferior izquierda son las mismas que las tres gráficas de la parte superior derecha de la matriz pero con los ejes invertidos, y en ellas el mapa de color está en escala logarítmica e indica en número de estrellas por bin en escala logarítmica.

Tabla 3.3: Número de estrellas que conforman los catálogos de entrenamiento, de validación y de pruebas.

Catálogo	Porcentaje	N° de estrellas
Entrenamiento	72 %	13,004
Validación	8 %	1,446
Pruebas	20 %	3,612

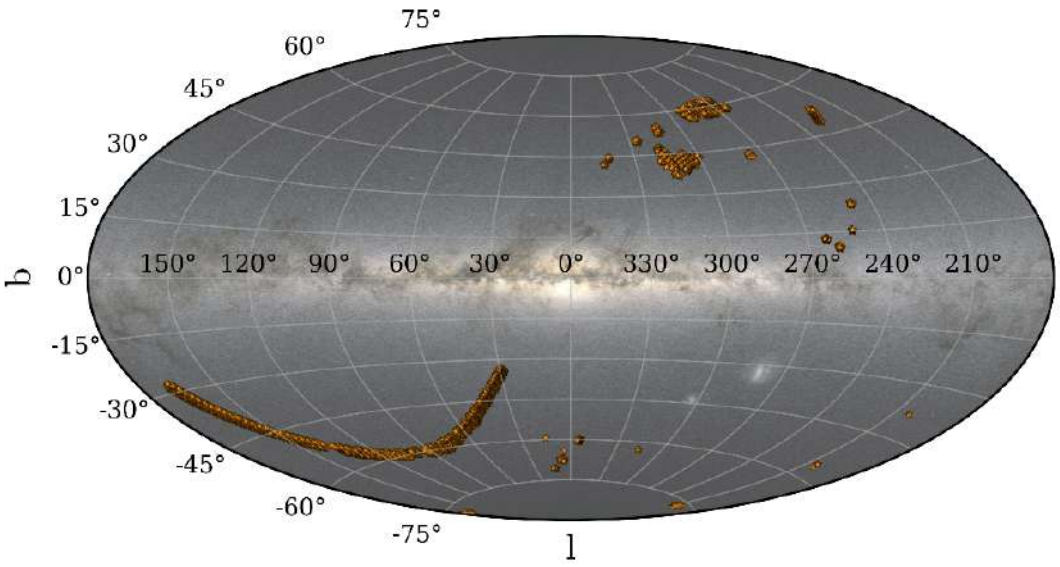


Figura 3.4: Distribución espacial de los campos conteniendo las estrellas del set de entrenamiento, de validación y de pruebas en coordenadas Galácticas en proyección Aitoff sobre la imagen del cielo de ESA/Gaia/DPAC.

del set de datos la clase o categoría a la que pertenece cada estrella (por ejemplo, denotado por ceros y unos, como se hizo en este caso).

Para determinar los parámetros estelares de los espectros de las estrellas APOGEE, ASPCAP compara las observaciones con una gran biblioteca de espectros sintéticos organizados en grandes cuadrículas o *grids* y se determina el espectro sintético que mejor se adapta a cada observación. APOGEE DR17 cuenta con 5 *sub-grids* o subcuadrículas para distintos tipos espectrales (que cubren distintos rangos en el diagrama H-R): dos para estrellas gigantes (GK y M) y tres para enanas (F, GK y M) (Jönsson et al., 2020). La subcuadrícula de espectros sintéticos que se ajusta mejor a una estrella es el parámetro `ASPCAP_GRID`. En la Figura 3.5 se observa la distribución de las estrellas definidas como gigantes o enanas según su `ASPCAP_GRID`. Aquí se muestra cómo la base de la RGB está poblada por estrellas clasificadas ya sea como gigantes o enanas.

Es importante destacar que, entre todas las categorías posibles, en este trabajo se están considerando a las estrellas como pertenecientes a una categoría entre gigantes o enanas; por ejemplo, no se consideraron categorías adicionales para estrellas subgigantes o para la rama asintótica de las gigantes (AGB). Respecto a esto, en primer lugar, si se quisiera desarrollar un modelo que clasifique las estrellas entre más

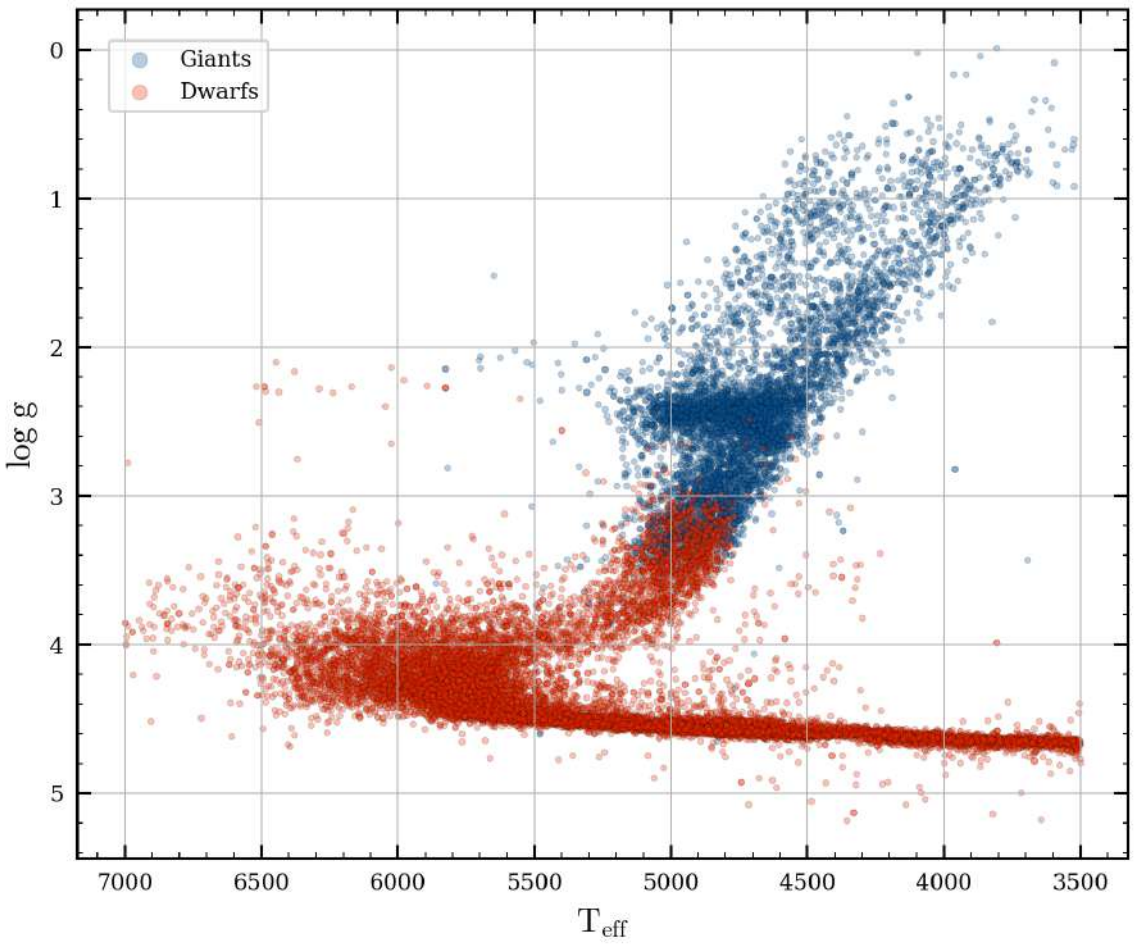


Figura 3.5: Distribución de estrellas enanas (en rojo) y gigantes (en azul) del set de entrenamiento, de validación y de pruebas según el *best-fit spectrum* en el plano (T_{eff}, log g).

categorías, se requeriría de datos adicionales de cada clase (no disponibles en esta oportunidad) para entrenar un modelo robusto que las logre identificar. En segundo lugar, para los fines de este proyecto no es necesario contar con una clasificación más detallada, puesto que la importancia de la clasificación radica en identificar las estrellas gigantes rojas entre el resto de las estrellas.Las estrellas gigantes y enanas pueden verse muy similares fotométricamente, y al no conocer las distancias a las mismas, puede no ser sencillo distinguir cuáles de ellas son en realidad enanas ubicadas en la vecindad Solar y cuáles son gigantes ubicadas a mayores distancias. Las estrellas clasificadas como gigantes rojas en `ASPCAP_GRID` son efectivamente gigantes rojas, sin embargo, las enanas son mas dudosas porque hay una superposición de estrellas enanas y gigantes en la base de la rama de las gigantes, en $\log g > 3.0$, como se observa en la Figura 3.5. Además, `ASPCAP_GRID` identifica como enanas a las estrellas ubicadas en $2 < \log g < 3$ y $T_{eff} > 5{,}800$ en la Figura 3.5, pero estas tienen una gravedad superficial muy baja para ser enanas. Por estas dos últimas razones es que no se utilizó el parámetro `ASPCAP_GRID` de APOGEE.

Para establecer la separación entre las estrellas gigantes y enanas se entrenó un algoritmo de *clustering* (o agrupamiento) sobre las 18,062 estrellas de los catálogos de entrenamiento, de validación y de pruebas. Los algoritmos de *clustering* permiten agrupar conjuntos de objetos de un set de datos de modo que los objetos de un mismo *cluster* o grupo tengan más similitud entre sí que con los objetos de los otros grupos. En particular, para este caso se entrenó el algoritmo de agrupamiento jerárquico aglomerativo (*agglomerative hierarchical clustering*) `AgglomerativeClustering` del módulo `sklearn.cluster` (Michel et al., 2019) de scikit-learn (Pedregosa et al., 2011). Este algoritmo realiza un agrupamiento jerárquico utilizando una estrategia "*bottom-up*": cada observación comienza en su propio *cluster* y los *clusters* se van fusionando sucesivamente. Para este problema, se implementó el algoritmo de forma que identificara dos grupos en el plano ($T-eff$, $\log g$), concretamente aplicando el algoritmo con los parámetros por defecto a los datos de T_{eff} y $\log g$ estandarizados con `MinMaxScaler` del módulo `sklearn.preprocessing` (Gramfort et al., 2019). Los resultados obtenidos por este algoritmo para el set de entrenamiento, de validación y de pruebas se presenta en la Sección 4.1.

Una vez construido el set de entrenamiento, de validación y de pruebas (con las *labels* de gigante o enana incluidas) se procedió a la clasificación de gigantes y enanas a partir de colores fotométricos.

3.2. Algoritmo para clasificación de estrellas entre gigantes y enanas

En esta Sección se describe el procedimiento utilizado para la construcción, el entrenamiento y la evaluación de diferentes algoritmos automatizados capaces de determinar si una estrella es gigante o enana a partir de la fotometría de S-PLUS, tomando como catálogo de entrenamiento, de validación y de pruebas la muestra descrita en la Sección 3.1.2. Este catálogo cuenta con la información fotométrica de S-PLUS, a partir de la cual se computaron los colores fotométricos a utilizarse como características de entrada, y también cuenta con las etiquetas determinadas por el algoritmo de *clustering* descrito en la Sección 3.1.2.1 a partir de los parámetros estelares espectroscópicos T_{eff} y log g de APOGEE DR17.

La clasificación de estrellas gigantes y enanas corresponde a un problema de clasificación binaria. Esta tarea se llevó a cabo con sistemas de aprendizaje supervisado, ya que el set de entrenamiento utilizado, además de incluir las características de entrada (colores fotométricos), incluye las *labels* determinadas por el algoritmo de *clustering* que identifican a cada estrella como gigante o enana (indicado por una columna con ceros y unos, para enanas y gigantes, respectivamente, determinadas mediante la metodología descrita en la Sección anterior). Concretamente, se utilizaron redes neuronales profundas entrenadas con el set de datos descrito en la Sección 3.1.2. También se implementó por separado un algoritmo de *random forest*, ya que dicho algoritmo suele tener un alto desempeño en problemas de clasificación. Finalmente, se selecciona la solución que presente los mejores resultados.

Todas las redes neuronales fueron desarrolladas con el lenguaje de programación Python, con la librería TensorFlow (Abadi et al., 2015) en su versión 2.7 y el algoritmo de *random forest* con la librería scikit-learn (Pedregosa et al., 2011).

3.2.1. Redes neuronales

3.2.1.1. Características de entrada

La combinación de colores fotométricos a utilizar para entrenar un algoritmo con los "mejores resultados"[5] no es conocida a priori (Reimers et al., 2020). Si bien se conoce la sensibilidad de algunos filtros a ciertos parámetros estelares (por ejemplo, el filtro F515 es sensible principalmente a log g, Majewski et al. 2000; o los filtros F410 y F395 son sensibles a T_{eff} y [M/H], respectivamente, Whitten et al. 2021), no se puede determinar exactamente qué incidencia tendría cada filtro en un modelo que determine dichas relaciones.

Para encontrar la mejor combinación de colores fotométricos que responda al problema tratado en este trabajo, se realizó un procedimiento empírico probando todas las combinaciones posibles según las condiciones descritas a continuación:

- Considerar todos los colores determinados a partir de cada filtro de banda angosta y el filtro de banda ancha al que se superpone (ver Figura 1.5), ya que estos colores se comportan como índices espectrales y finalmente lo que se mide con estos índices es el flujo de cada línea espectral. Estos colores son: **(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F660 - r) y (F861 - z)**.
- Considerar los colores determinados a partir de todas las combinaciones posibles entre los filtros de banda ancha. Estos colores también se comportan como índices espectrales y generalmente entregan información de temperatura, o en algunos casos de metalicidad (sobre todo en bandas azules). Estos colores son: **(u - g), (u - r), (u - i), (u - z), (g - r), (g - i), (g - z), (r - i), (r - z) e (i - z).**

Bajo estas consideraciones se definieron diferentes combinaciones de colores fotométricos de entrada para entrenar las redes neuronales que fueron diseñadas. Así se puede comparar el rendimiento de una red al utilizar distintas características de entrada y finalmente elegir a partir de las métricas los colores de las estrellas de S-PLUS que se utilizarán para identificar las estrellas gigantes y enanas. Se determinaron todas

[5] Los resultados son evaluados y comparados cuantitativamente a partir de diferentes métricas.

las combinaciones posibles a partir de los colores mencionados anteriormente, que cuentan con 4, 5, 6, 7 y 8 colores fotométricos. Cada combinación contó con 3, 4, 5, 6 o 7 colores determinados por los índices de color definidos por los filtros de banda angosta y su respectivo banda ancha y además un color compuesto de filtros banda ancha. No se agregaron más colores de fitros banda ancha, ya que la información que aportan generalmente es la sensibilidad a la temperatura efectiva ya presente y considerar más de estos colores aumentaría la necesidad de cómputo del problema. Se obtuvieron 350 + 350 + 210 + 70 + 10 combinaciones de 4, 5, 6, 7 y 8 colores, respectivamente, que dan en total 990 combinaciones de colores que representan las distintas alternativas de características de entrada para entrenar las redes neuronales que fueron desarrolladas.

3.2.1.2. Arquitectura e hypertuning

El diseño de las redes neuronales artificiales consta de una serie de hiperparámetros, es decir, aquellos parámetros que controlan el proceso de entrenamiento de un modelo (TensorFlow, 2022). Probar combinaciones de hiperparámetros individualmente para luego determinar la arquitectura óptima de un modelo es una tarea que podría demandar mucho tiempo. Por esto se utilizó KerasTuner (O'Malley et al., 2019), una librería en el lenguaje Python que ayuda a determinar el conjunto óptimo de hiperparámetros a utilizar sobre un espacio de búsqueda previamente definido. Este proceso se denomina ajuste de hiperparámetros o *hypertuning*.

Se realizó el *hypertuning* 5 veces, para una combinación seleccionada para cada número de colores de entrada (4, 5, 6, 7 y 8 colores, siendo 5 combinaciones de colores seleccionadas en total), ya que se estimó que las mejores arquitecturas de las redes iban a variar al trabajar con distintos números de *input features*. No se hizo el ajuste para las 990 combinaciones de colores de entrada, ya que, a pesar de que este proceso es más rápido que el ajuste manual, sigue tardando varios minutos para cada ajuste en las máquinas que se disponían para el desarrollo de este proyecto[6].

El proceso de *hypertuning* con KerasTuner cuenta con varias etapas (Invernizzi et al., 2019), las que son descritas a continuación:

[6]El equipo utilizado para el desarrollo de los algoritmos fue un *notebook* HP OMEN, con un procesador Intel Core i5 de 4 núcleos junto a 8 GB de RAM, utilizando el sistema operativo Windows 11 Pro.

Ajuste de la arquitectura del modelo

Se creó un modelo para ser ajustado y a su vez se definió el espacio de búsqueda de hiperparámetros. Esto se llevó a cabo mediante una función que devuelve un modelo compilado (TensorFlow, 2021), en la cual se indican los hiperparámetros que se desean ajustar y los rangos o alternativas entre las cuales se buscará la mejor arquitectura.

Los parámetros que se ajustaron fueron:

- Número de capas ocultas: entre 1 y 5;
- Número de unidades de cada capa oculta: múltiplos de 32 entre 32 y 512;
- Función de activación para cada capa oculta: relu, sigmoid o tanh;
- Capas de *dropout*[7]: con fracciones entre 0.0 y 0.3;
- *Learning rate*: 1e-2, 1e-3 o 1e-4.

Notar que el número de unidades en las capas ocultas, sus funciones de activación y la fracción de las capas de *dropout* después de cada capa oculta dependen también del número de capas ocultas, lo que también es un hiperparámetro a ser ajustado. Por lo tanto, el número de hiperparámetros a ajustar puede ser entre 5 y 17 (dependiendo del número de capas ocultas).

En cuanto a las entradas de cada red, se añadió una capa de preprocesamiento para normalizar las características de entrada. Si bien, al trabajar con datos estructurados la importancia de la normalización de las características de entrada es más evidente en problemas en donde éstas tienen distintos órdenes de magnitud (Géron, 2017), en general se recomienda llevar a cabo este paso, ya que disminuyen el tiempo de convergencia de los algoritmos (Grus, 2019). Se utilizó una capa `tf.keras.layers.Normalization` (Chollet & Omernick, 2020). Esta capa escala los datos de entrada, centrándolos en 0 con una desviación estándar de 1.

Las capas de salida de todas las redes neuronales contaron con dos unidades (ya que se están clasificando los datos entre dos clases) y una función de activación softmax (Keras, 2021a). Esta función se suele aplicar en problemas de clasificación con

[7] Una capa de *dropout* establece aleatoriamente las unidades de entrada en 0 con una frecuencia que corresponde a la tasa o fracción asignada para la capa, lo que ayuda a evitar el sobreajuste.

categorías mutuamente excluyentes, funciona asignando un "puntaje de confianza" a cada clase y aquella con puntaje más alto es la clase que predice (Géron, 2017; Radečić, 2020).

Los parámetros para la compilación del modelo fueron los siguientes:

- `optimizer`: Adam, un método de *gradient descent* estocástico computacionalmente eficiente (Kingma & Ba, 2017);
- `loss`: Sparse Categorical Cross entropy, función de pérdida probabilística utilizada en problemas que involucran dos o más clases identificadas por números enteros (Keras, 2021b);
- `metrics`: *accuraccy* o exactitud.

Los parámetros de las capas de entrada, de salida y de la compilación del modelo fueron fijados en la función anteriormente mencionada.

Determinación de tuner class

Luego de tener definido el espacio de búsqueda de parámetros, se debe elegir la *tuner class* que corresponde al *tuning algorithm* o algoritmo de ajuste. Se seleccionó la clase Hyperband, ya que ha demostrado ser más rápida en una amplia variedad de problemas de aprendizaje automatizado (Li et al., 2018).

Para iniciar el tuner se deben especificar algunos parámetros (Keras, 2019), los más importantes son:

- `hypermodel`: función descrita en el punto anterior;
- `objective`: objetivo a optimizar, se infiere de las métricas predefinidas en el módulo si se debe minimizar o maximizar. Dependerá del tipo de problema: para la clasificación se usó la exactitud o *accuracy* (se debe maximizar);
- `max_epochs`: número máximo de épocas para entrenar cada modelo en el proceso de ajuste. Se fijó en 20 épocas.

Preparación del set de datos

Como se mencionó anteriormente, el proceso de *hypertuning* se realizó 5 veces, correspondiendo a las combinaciones de 4, 5, 6, 7 y 8 colores de entrada. Para elegir las combinaciones usadas en cada proceso se seleccionaron aquellas conformadas por los colores compuestos por las bandas con mayor SNR medio entre todas las estrellas. Es decir, para el proceso de *hypertuning* para las combinaciones con 4 colores de entrada se utilizó la combinación de 4 colores dentro de las 990 cuyas bandas tenían mayores SNR medios y así para cada cantidad de colores. Las combinaciones de colores de entrada para cada ajuste de hiperparámetros se encuentran en la Tabla 3.4.

Tabla 3.4: Colores de entrada utilizados para realizar el ajuste de hiperparámetros.

Número de colores	Colores de entrada
4	(F660 - r), (F861 - z), (F515 - g), (r - i)
5	(F660 - r), (F861 - z), (F515 - g) (F430 - g), (R - i)
6	(F660 - r), (F861 - z), (F515 - g) (F430 - g), (F410 - g), (r - i)
7	(F660 - r), (F861 - z), (F515 - g) (F430 - g), (F410 - g), (F395 - g), (r - i)
8	(F660 - r), (F861 - z), (F515 - g) (F430 - g), (F410 - g), (F395 - g) (F378 - u), (r - i)

El set de entrenamiento está separado en las variables `x_train` e `y_train`, el set de validación en `x_val` e `y_val` y el set de pruebas en `x_test` e `y_test`. Las variables "`x`" corresponden a las entradas (los colores fotométricos) y las variables "`y`" a las *labels* (identificación de estrellas como gigantes o enanas representadas en los datos como unos y ceros, respectivamente).

Búsqueda de los mejores hiperparámetros

La función que posee la arquitectura de la red y el espacio de búsqueda de hiperparámetros es llamada repetidas veces por KerasTuner durante el proceso de ajuste con diferentes hiperparámetros en cada una de las pruebas que se realizan (Invernizzi et al., 2019). Los modelos son ajustados, evaluados y las métricas son registradas por KerasTuner. Mediante este procedimiento, el *tuner* va indagando en el espacio de búsqueda anteriormente definido hasta encontrar la mejor configuración de hiperparámetros. Adicionalmente, se añadió un *callback* a este proceso. Un *callback* es un objeto (en contexto computacional) que realiza ciertas acciones durante el entrenamiento. En este caso se utilizó para hacer un *earlystopping* que hace que el modelo deje de entrenar cuando una métrica definida ha dejado de mejorar (Keras, 2020). En este caso se monitoreó el valor de la función de pérdida en el set de validación con una paciencia (número de épocas sin mejora después de las cuales se detendrá el entrenamiento) de 5.

Una vez terminado el procedimiento descrito, se obtiene el mejor modelo seleccionado según la métrica *accuracy* para cada proceso de *hypertuning*. Luego de repetir este paso para las 5 combinaciones de colores de entrada elegidas para el *hypertuning*, se obtuvieron los modelos con los hiperparámetros optimizados para 4, 5, 6, 7 y 8 colores fotométricos de entrada. Estos modelos son guardados para posteriormente ser aplicados a las demás combinaciones de colores, como se explica a continuación.

3.2.1.3. Entrenamiento con diferentes combinaciones de colores de entrada

Para seleccionar el mejor modelo se debe realizar el entrenamiento y evaluación de cada una de las posibles combinaciones. Para ello, se definió una función que itera a lo largo de la lista que contiene las 990 combinaciones de colores (definidas en la Sección 3.2.1.1). Para cada combinación esta función realiza las siguientes tareas:

- Define `x_train`, `y_train`, `x_val`, `y_val`, `x_test` e `y_test` de forma aleatoria (procurando que los colores utilizados usen siempre el mismo set de datos aleatorios, ya que esto permite una mejore comparación de resultados) en las proporciones que se indican en la Tabla 3.3.

- Asigna uno de los modelos previamente ajustados según el número de colores de la combinación.
- Entrena el modelo con los datos referenciados por las variables `x_train` e `y_train`.
- Evalúa el modelo sobre los datos contenidos en `x_test`, determinando la pérdida o *loss* y la exactitud o *accuracy*.
- Realiza predicciones con `x_test`.
- A partir de las predicciones, define:
  - Número de gigantes correctamente identificadas;
  - Número de gigantes identificadas como enanas;
  - Número de enanas correctamente identificadas;
  - Número de enanas identificadas como gigantes;
  - *Recall* o exhaustividad para gigantes: fracción de gigantes correctamente clasificadas;
  - *Recall* o exhaustividad para enanas: fracción de enanas correctamente clasificadas;
  - Precisión de gigantes: fracción de las estrellas identificadas como gigantes correctamente identificadas;
  - Precisión de enanas: fracción de las estrellas identificadas como enanas correctamente identificadas.
- Genera una tabla que registra las métricas obtenidas para cada una de las 990 combinaciones de colores.

3.2.1.4. Determinación de la red con mejor rendimiento

A partir de la tabla generada por la función descrita en la Sección 3.2.1.3, se hizo una comparación entre el rendimiento de las diferentes ANNs y las diferentes combinaciones de colores de entrada utilizadas.

Los parámetros que se utilizaron para determinar la mejor ANN fueron: exhaustividad, precisión para estrellas gigantes y enanas y exactitud para los modelos. Para todas estas métricas un valor mayor implica un mejor desempeño.

3.2.1.5. Clasificación de las estrellas de S-PLUS

Una vez definida la red neuronal y los colores de entrada que en su conjunto obtienen el mejor rendimiento al momento de clasificar las estrellas entre gigantes y enanas, se aplicó dicho algoritmo a todas las estrellas (`CLASS_STAR` $>$ `0.92`) de los catálogos de S-PLUS (ver Tabla 3.1). De estas estrellas, al igual que con las que constituyeron los set de entrenamiento, de validación y de pruebas, se consideraron sólo aquellas con SNR > 50 en cada una de las bandas que se utilizaron en la ANN. Adicionalmente, se eliminaron aquellos datos con magnitud igual a 99.0 en alguno de los filtros utilizados, ya que este es el valor para fuentes no detectadas.

3.2.2. Algoritmo de random forest

Los algoritmos de *random forest* (Breiman, 2001) son reconocidos por su buen desempeño al trabajar con problemas de clasificación (Caruana et al., 2008). Es por esto que, alternativamente a las redes neuronales, se elaboró un algoritmo de este tipo para abordar el problema de clasificación de estrellas gigantes y enanas.

Este algoritmo se contruyó utilizando `RandomForestClassifier` del módulo `sklearn.ensemble` (Louppe et al., 2019). Se utilizó la misma combinación de colores de entrada determinada en la Sección 3.2.1.4, y además se usaron las mismas variables `x_train` e `y_train` para entrenar el modelo y `x_test` e `y_test` para evaluarlo. Cabe destacar que en los algoritmos de *random forest* no se realizó un ajuste de hiperparámetros, por lo cual no se utilizó el set de validación. De esta forma se pudo comparar el desempeño de ambos algoritmos, redes neuronales y *random forest*, en base a los mismos datos.

3.3. Algoritmo para la determinación de metalicidades estelares

En esta Sección se describe el procedimiento utilizado para la construcción, el entrenamiento y la evaluación de diferentes algoritmos automatizados capaces de derivar metalicidades estelares a partir de la fotometría de S-PLUS, tomando como catálogo de entrenamiento, de validación y de pruebas la muestra descrita en la Sección 3.1.2. Este catálogo cuenta con la información fotométrica de S-PLUS, a partir de la cual se computaron los colores fotométricos a utilizarse como características de entrada, y también cuenta con las metalicidades de APOGEE DR17, que funcionan como las etiquetas en estos algoritmos.

La determinación de las metalicidades de las estrellas corresponde a un problema de regresión. Al igual que para la clasificación, esta tarea se realizó mediante sistemas de aprendizaje supervisado con un set de entrenamiento que incluye colores fotométricos como características de entrada y las metalicidades espectroscópicas de APOGEE DR17 como *targets* o *labels*. Se realizaron 3 conjuntos de redes neuronales: uno entrenado sólo con las estrellas gigantes, el segundo con el set de estrellas completo y otro entrenado sólo con las estrellas enanas. Se estimó que entrenar modelos por separado para estrellas gigantes y enanas podría resultar en un mejor rendimiento en cuanto a la precisión de los resultados. Por esta razón, se entrenaron los tres conjuntos por separado, para así poder comparar los desempeños de cada grupo (ver Sección 4.2.3). A pesar de que la determinación de metalicidades de estrellas enanas no es parte de los objetivos principales de este estudio, esto se realizó de todas formas, ya que contar con dicha información complementa el análisis de las estrellas de S-PLUS y se logra de manera relativamente sencilla reutilizando los programas computacionales ya implementados.

3.3.1. Redes neuronales

3.3.1.1. Características de entrada

Como se mencionó en la Sección 3.2.1.1, la combinación óptima de colores de entrada para cada problema no es conocida previamente. Entonces se siguió un proceso idéntico al descrito en dicha Sección para determinar las 990 combinaciones de colores fotométricos disponibles como características de entrada.

3.3.1.2. Arquitectura e hypertuning

Para determinar la mejor arquitectura de los algoritmos se realizó un proceso de *hypertuning* con KerasTuner análogo al descrito en la Sección 3.2.1.2 para la tarea de clasificación.

En este caso el *hypertuning* se realizó 15 veces: 5 veces para cada uno de los 3 grupos de redes neuronales y subconjuntos de datos. De estas 5 se realizó una vez el *hypertuning* para cada número de colores de input (4, 5, 6, 7 y 8). A continuación, se describen los pasos seguidos para este proceso. El procedimiento para los 3 conjuntos de redes neuronales fue el mismo, sólo cambian los sets de datos utilizados que incluyen sólo enanas, sólo gigantes y gigantes+enanas.

Ajuste de la arquitectura del modelo

Se elaboró una función computacional que entrega un modelo compilado con el espacio de búsqueda de hiperparámetros definido. El espacio de búsqueda es el mismo que el definido para las redes de clasificación. La diferencia es que esta función está diseñada para el problema de regresión: la capa de salida es una capa densa de una unidad (ya que se va a predecir un sólo parámetro, la metalicidad), utilizando como función de pérdida MAE y definiendo las métricas MAE y MSE. Al igual que para la clasificación, esta función incluye una capa de preprocesamiento para normalizar las características de entrada, `tf.keras.layers.Normalization` y un optimizador Adam.

Determinación de tuner class

Al igual que en la Sección 3.2.1.2 se utilizó la clase *Hyperband* (Li et al., 2018).

En este caso los parámetros se definieron de la siguiente manera:

- `hypermodel`: función descrita en el paso anterior;
- `objective`: *loss* o pérdida (a minimizar);
- `max_epochs`: 20 épocas.

Preparación del set de datos

Se utilizaron sets de datos diferentes para cada conjunto de redes neuronales. Sólo se utilizaron las estrellas identificadas como gigantes por el algoritmo de *clustering* (ver Sección 3.1.2.1) para el conjunto que determina metalicidades de estrellas gigantes. Se utilizaron las estrellas restantes (que son las identificadas como enanas por el algoritmo de *clustering*) para el conjunto que determina metalicidades de estrellas enanas y para las redes que determinan metalicidades de todas las estrellas se utilizó todo el set de datos descrito en 3.1.2. En todos los casos se utilizaron las mismas fracciones indicadas en la Tabla 3.3 para configurar los set de entrenamiento, de validación y de pruebas, pero para los primeros dos conjuntos la cantidad de datos es diferente.

Los colores fotométricos de entrada en las variables `x_train`, `x_val` y `x_test` para cada número de colores son los indicados en la Tabla 3.4. Las variables `y_train`, `y_val` e `y_test` contienen las metalicidades de APOGEE DR17, parámetro que los algoritmos descritos en esta Sección buscan predecir.

Búsqueda de los mejores hiperparámetros

El proceso es el mismo descrito en la Sección 3.2.1.2. Este proceso se realizó por separado para cada conjunto de redes neuronales y 5 veces para cada conjunto (uno por número de colores de entrada), dando un total de 15 modelos ajustados. Cada modelo fue guardado para luego ser aplicado a las distintas combinaciones de entrada, según el número de colores utilizados. Se añadió un *callback earlystopping*,

monitoreando la pérdida en el set de validación con una paciencia igual a 5.

3.3.1.3. Entrenamiento con diferentes combinaciones de colores de entrada

Se definió una función similar a la descrita en la Sección 3.2.1.3 que itera las 990 combinaciones de colores realizando las siguientes acciones:

- Define `x_train`, `y_train`, `x_val`, `y_val`, `x_test` e `y_test` en las razones que se indican en la Tabla 3.3 (procurando que los colores utilizados usen siempre el mismo set de datos aleatorios para cada conjunto de redes que utiliza sólo enanas, sólo gigantes y gigantes+enanas, ya que esto permite una mejor comparación de resultados) de forma aleatoria.
- Según el número de colores y el set de datos utilizado (sólo gigantes, sólo enanas o todas las estrellas) le asigna el modelo con los hiperparámetros ajustados correspondiente.
- Entrena el modelo con los datos referenciados en las variables `x_train` e `y_train`.
- Evalúa el modelo sobre `x_test`, determinando la pérdida, MAE y MSE.
- Genera una tabla que registra las métricas obtenidas para cada combinación de colores.

Esta función se ejecutó 3 veces, una utilizando los datos de las estrellas gigantes, otra con los datos de las enanas y la tercera con los datos de todas las estrellas.

3.3.1.4. Determinación de la red con mejor rendimiento

A diferencia del proceso realizado para determinar la mejor red para la clasificación de estrellas, en este caso primero se debe identificar si efectivamente elaborar modelos por separado para estrellas gigantes y enanas resulta mejor para la determinación de sus metalicidades que ocupar una única red que determine las metalicidades de todas las estrellas. Para esto se compararon los rendimientos de los 3 conjuntos de redes sobre todas las combinaciones de colores a la vez y luego se seleccionaron las redes con los mejores rendimientos. Los rendimientos se midieron con las métricas

MAE y MSE. Estas métricas son medidas de error, por lo que un valor menor de MAE y MSE representan un mejor desempeño.

3.3.1.5. Aplicación a las estrellas de S-PLUS

Las redes neuronales con mejor desempeño en la determinación de metalicidades estelares fueron aplicadas a todas las estrellas (`CLASS_STAR` $>$ `0.92`) de los catálogos de S-PLUS (ver Tabla 3.1).

Una vez obtenidas las metalicidades de las estrellas gigantes y enanas, se realizó un *cross-match* en un radio de 1.0 arcosegundo con el catálogo de distancias foto-geométricas (ya que en general son más precisas) de Bailer-Jones et al. (2021) de Gaia Early Data Release 3 (Gaia EDR3, Gaia Collaboration et al. 2021), ya que dentro de este radio se lograron coincidencias para más del 90 % de las fuentes. A partir de estas distancias determinaron las coordenadas Galactocéntricas X, Y, Z y R (Jurić et al., 2008), como se indica en las siguientes ecuaciones:

$$\begin{aligned} X =& R_{\odot} - \ d\ cos(l)\ cos(b), \\ Y =& \ d\ sin(l)\ cos(b), \\ Z =& \ d\ sin(b), \\ R =& \sqrt{X^2 + Y^2} \end{aligned} \tag{3.5}$$

donde l y b son la longitud y latitud galáctica, d las distancias y $R_{\odot} = 8$ kpc es la distancia del Sol al centro Galáctico adoptada (Reid, 1993; Camarillo et al., 2018). El centro del plano (X,Y) se ubica en el centro Galáctico. El eje X apunta hacia el Sol, el eje Z hacia el Polo Norte Galáctico y el eje Y es perpendicular a ambos, de tal manera que la rotación del Sol sigue el sentido positivo del eje Y. El plano (X,Y) es paralelo al plano de la Galaxia y en Z = 0 está contenido el Sol. R es la distancia al centro de la Galaxia. Las coordenadas X,Y,Z componen el sistema de coordenadas cartesianas Galactocéntrico y R,Z son coordenadas cilíndricas Galactocéntricas.

Con esta información se pudo determinar la distribución de metalicidad de las estrellas de la Galaxia contenidas en la submuestra del catálogo de SPLUS iDR3-n4. Sin embargo, hay que considerar que a pesar de la precisión de Gaia, las incertidum-

bres en los paralajes de las fuentes más distantes y débiles pueden llegar a ser muy grandes y estas repercuten en la determinación de distancias de Bailer-Jones et al. (2021).

Capítulo 4

Resultados y Discusión

En este capítulo se presentan los resultados obtenidos a partir de la aplicación de las metodologías presentadas en el Capítulo anterior. Esto es, los resultados del algoritmo de *clustering*, los algoritmos diseñados para la clasificación de estrellas entre gigantes o enanas y los algoritmos diseñados para la determinación de metalicidades de estrellas gigantes, enanas y para todas las estrellas. Adicionalmente se presentan los resultados de la aplicación de estos algoritmos sobre las estrellas de S-PLUS iDR3 junto con un análisis preliminar de la información obtenida.

4.1. Algoritmo de clustering para seleccionar estrellas gigantes y enanas

En esta sección se presentan los resultados de la aplicación del algoritmo `AgglomerativeClustering` en las 18,063 estrellas del set de datos que compone el catálogo de entrenamiento, de validación y de pruebas (ver Figura 4.1). El algoritmo fue aplicado con los parámetros establecidos por defecto en dicho módulo[1].

En la Figura 4.1 se presenta la distribución de los dos grupos que identificó el modelo: el azul correspondiente a las estrellas gigantes y el rojo a estrellas enanas. El *cluster* de las gigantes está compuesto por 7,356 estrellas y el de las enanas por 10,707.

[1]Los parámetros definidos por defecto se encuentran disponibles en la documentación de Scikit Learn, accesibles aquí.

Como se mencionó en la Sección 3.1.2.1, parte de este proyecto apunta a desarrollar un algoritmo de clasificación que sea capaz de reconocer las estrellas entre dos grupos, que básicamente estén representados por las aglomeraciones notadas en el diagrama de contorno en la Figura 4.1. En dicha Figura se puede observar cómo el uso del algoritmo de *clustering* cumple esta función, no así mediante el criterio de ASPCAP que se muestra de la Figura 3.5.

Hay que considerar que las estrellas ubicadas en la zona de $2 < \log g < 3$ y 5,500 $< T_{eff} <$ 6,600 K en la Figura 4.1, que el algoritmo de *clustering* ha agrupado con las gigantes, son muy calientes para ser estrellas gigantes rojas. Sin embargo, este conjunto de $\sim$ 30 estrellas no está siquiera cerca de ser representativo dentro del grupo de las 7,365 estrellas seleccionadas como gigantes, por lo que no se espera que tenga mayor incidencia en el algoritmo que clasifica las estrellas.

En la Figura 4.2 se muestra la distribución de probabilidad de las metalicidades espectroscópicas de APOGEE DR17 de las estrellas gigantes y enanas clasificadas por el algoritmo de *clustering*. Es importante conocer esta información, ya que la precisión y solidez de un modelo de aprendizaje automatizado depende de los datos con los cuáles fue entrenado (Hettiarachchi et al., 2005). El rango de las características del set de entrenamiento, en este caso metalicidades, también es importante a tener en cuenta, ya que los algoritmos suelen presentar grandes errores al momento de extrapolar (Wang et al., 2020).

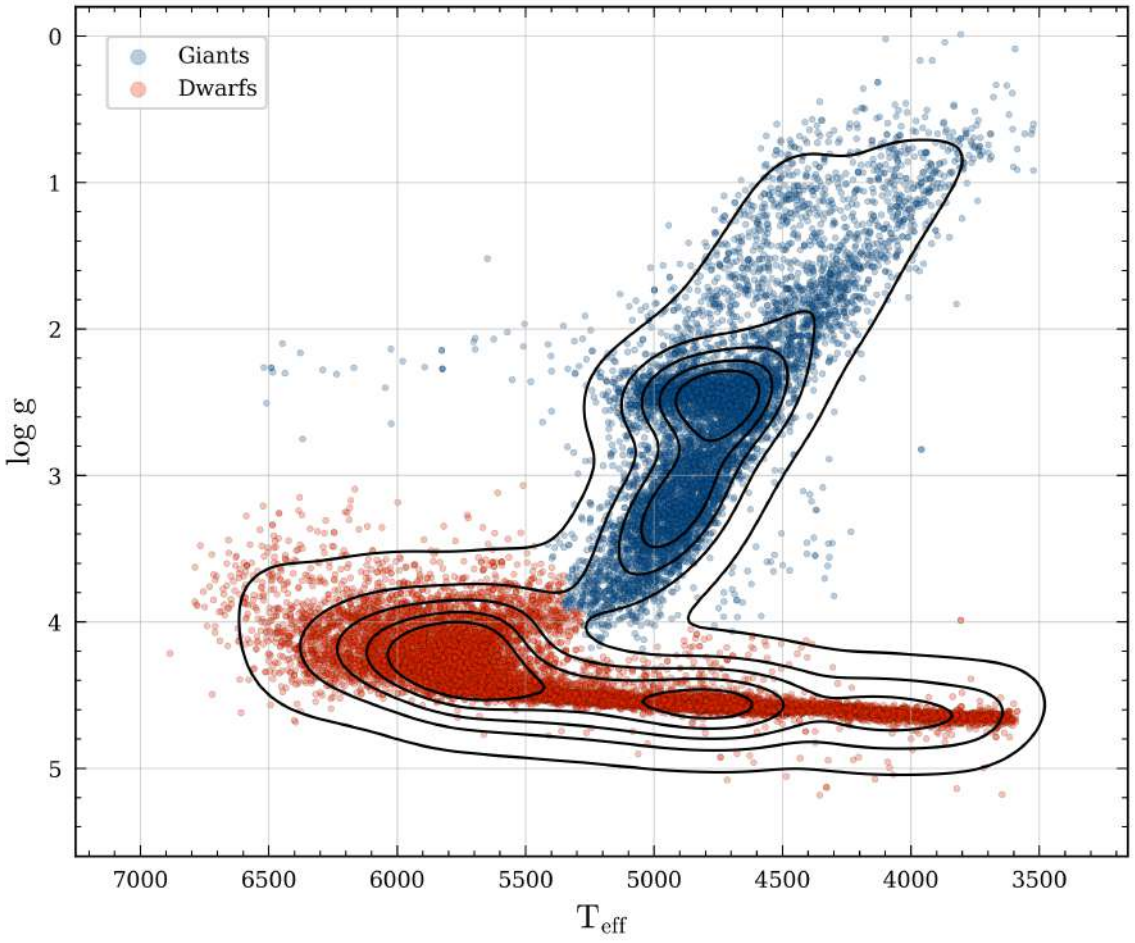


Figura 4.1: Distribución de estrellas enanas (en rojo) y gigantes (en azul) determinada por el algoritmo `AgglomerativeClustering` en el plano (T_{eff}, log g). Las líneas negras representan la densidad de objetos contenidos en ellas. Nótese cómo la separación producida por el algoritmo coincide con la zona de menor densidad, delineando por los contornos, entre ambas clases. Nótese, además, la similitud de los resultados obtenidos por este algoritmo con la información presentada en la Figura 3.5.

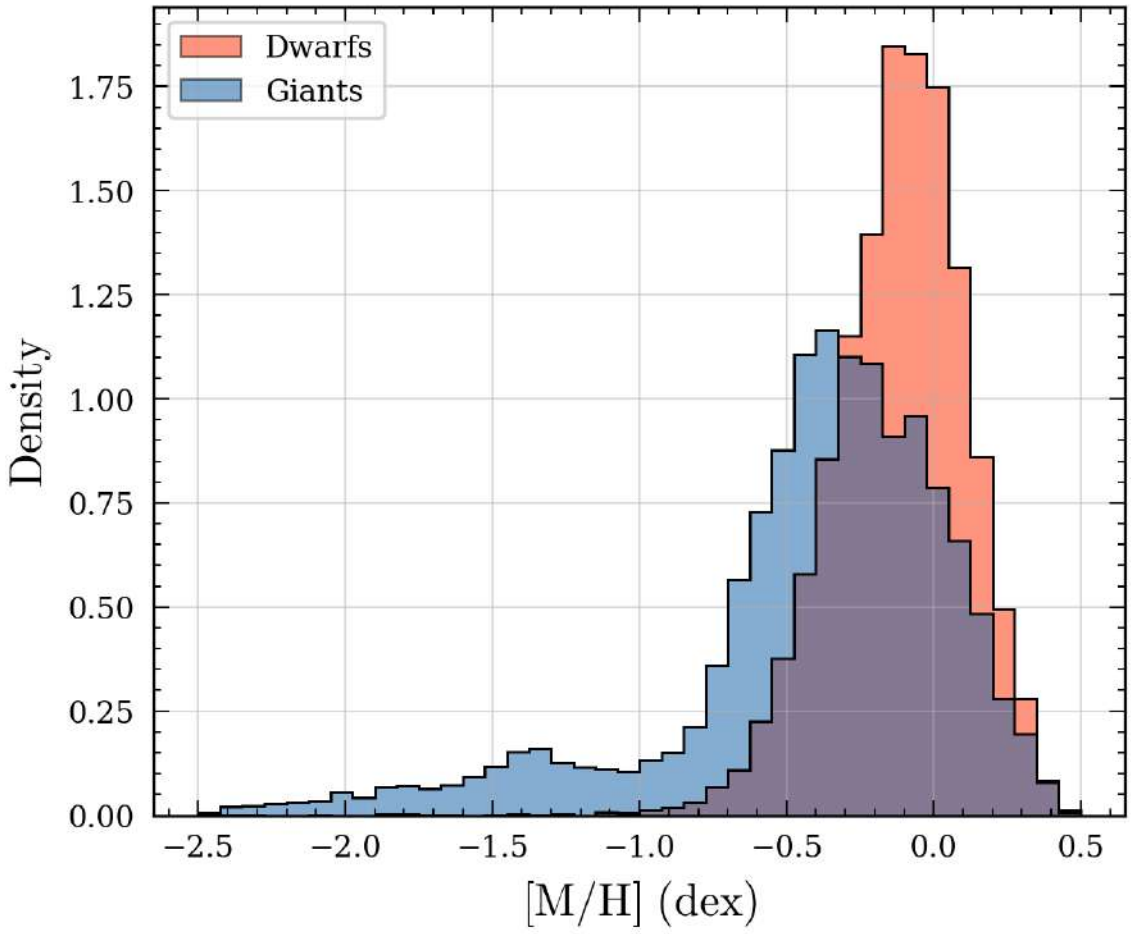


Figura 4.2: Distribución de probabilidad de metalicidades de estrellas gigantes (azul) y enanas (rojo) en el set de entrenamiento, de validación y de pruebas. Las metalicidades espectroscópicas de las estrellas gigantes presentan errores típicos de $\sigma \sim 0.009$ dex y las enanas de $\sigma \sim 0.01$ dex.

4.2. Algoritmos de clasificación de estrellas entre gigantes y enanas

4.2.1. Catálogo de entrenamiento, de validación y de pruebas

Luego de aplicar el algoritmo de *clustering* al set de datos de 18,062 estrellas descrito en la Sección 3.1.2, se identificaron las etiquetas de los catálogos de entrenamiento, de validación y de pruebas (estos catálogos fueron determinados de forma aleatoria según las proporciones indicadas en la Tabla 3.3). La cantidad de estrellas gigantes y enanas en cada set se muestra en la Tabla 4.1.

Tabla 4.1: Cantidad de estrellas gigantes y enanas en los sets de entrenamiento, de validación y de pruebas utilizados en los algoritmos de clasificación.

Catálogo	N° gigantes	N° enanas	N° total de estrellas
Entrenamiento	5,296	7,708	13,004
Validación	589	857	1,446
Pruebas	1,471	2,141	3,612

4.2.2. Arquitectura de las redes neuronales

Para determinar la arquitectura de las ANN desarrolladas, se realizó un ajuste de hiperparámetros o *hypertuning*, siguiendo el procedimiento indicado en la Sección 3.2.1.2.

Los hiperparámetros de las ANN determinados mediante este proceso, para cada número de colores de entrada, se muestran en la Tabla 4.2. Notar que, en todas las redes KerasTuner, determinó el *learning rate* óptimo igual a 0.001, una función de activación relu para la primera capa oculta y, excepto para la red que toma 8 colores de entrada, estableció 4 capas ocultas. Nótese también que los hiperparámetros de las redes que ocupan 5 y 7 colores de entrada quedaron con los mismos hiperparámetros. Cabe destacar que este procedimiento se repitió con estas últimas dos combinaciones

de colores para asegurarse de que no hubiese algún error de digitación dentro del código.

4.2.3. Determinación de las redes neuronales con mejor desempeño

Para la determinación de las mejores redes neuronales se utiliza la función descrita en la Sección 3.2.1.3 se generó un registro del desempeño de las ANNs entrenadas con las 990 combinaciones de colores de entrada determinadas en la Sección 3.2.1.1. A continuación se detallan los pasos para este procedimiento.

En primer lugar, se comparó la tasa de gigantes correctamente clasificadas con la tasa de enanas correctamente clasificadas (exhaustividad o *recall*, ver Ecuación 1.4). En la Figura 4.3 se muestran los valores obtenidos para la exhaustividad al evaluar los 990 entrenamientos realizados para gigantes y enanas. Se observa que, en general, la exhaustividad de las enanas es mayor que para las estrellas gigantes. Esto último indica que las redes estarían identificando una mayor fracción de enanas correctamente que de gigantes, llegando a identificar en varios casos más del 98 % de las enanas de la muestra, mientras que las redes con mayor exhaustividad de gigantes logran identificar el $\sim 96\,\%$ de las gigantes de la muestra. Cabe destacar que un gráfico como el de la Figura 4.3 por sí sólo no indica qué colores aplicados a las redes producen un mejor o peor rendimiento, sino que simplemente muestran el rendimiento general de los algoritmos desarrollados y permiten ir descartando aquellos con un desempeño inferior. Para comparar el rendimiento de diferentes combinaciones de colores hay que consultar individualmente en el registro los números asociados a éstas.

En primera instancia, se descartaron aquellas redes con combinaciones de colores de entrada que lograran identificar menos del 96 % de las gigantes correctamente, quedando así 48 combinaciones. Entre estos entrenamientos, el que reconoció la menor fracción de estrellas enanas logró identificar el 97.4 % de ellas. De las redes que quedaron preseleccionadas, se aplicó un nuevo corte: se descartaron aquellas que identificaran menos del 96.5 % de las estrellas gigantes (no se descartaron inicialmente las redes que identificaron menos del 96.5 % de las gigantes para poder monitorear en paralelo el rendimiento con las estrellas enanas). De aquí resultaron 5 redes neuronales, cuyos colores de entrada, exhaustividad, precisión, exactitud (ver Ecuaciones 1.4, 1.3

Tabla 4.2: Hiperparámetros determinados por KerasTuner para las combinaciones de colores indicadas en la Tabla 3.4.

Número de colores de entrada		4	5	6	7	8
	Capas ocultas	4	4	4	4	5
1° capa	Unidades	384	224	192	224	96
	Función de activación	relu	relu	relu	relu	relu
	Fracción de dropout	0.3	0.0	0.1	0.0	0.1
2° capa	Unidades	320	128	64	128	256
	Función de activación	relu	tanh	tanh	relu	relu
	Fracción de dropout	0.0	0.2	0.1	0.2	0.0
3° capa	Unidades	288	384	384	384	288
	Función de activación	tanh	relu	sigmoid	tanh	relu
	Fracción de dropout	0.2	0.1	0.2	0.1	0.1
4° capa	Unidades	160	352	128	352	352
	Función de activación	tanh	sigmoid	sigmoid	sigmoid	sigmoid
	Fracción de dropout	0.2	0.1	0.1	0.1	0.1
5° capa	Unidades	-	-	-	-	32
	Función de activación	-	-	-	-	relu
	Fracción de dropout	-	-	-	-	0.0
	Tasa de aprendizaje	0.001	0.001	0.001	0.001	0.001

y 1.2, respectivamente) y *loss* (`SparseCategoricalCrossentropy`) se encuentran en la Tabla 4.3.

Estos cinco entrenamientos tienen rendimientos muy similares, con diferencias de no más del 1 % en las diferentes métricas obtenidas. Además, todas estas redes tienen en común los colores de entrada (F378 - u), (F395 - g), (F430 - g), (F515 - g), (F861 - z).

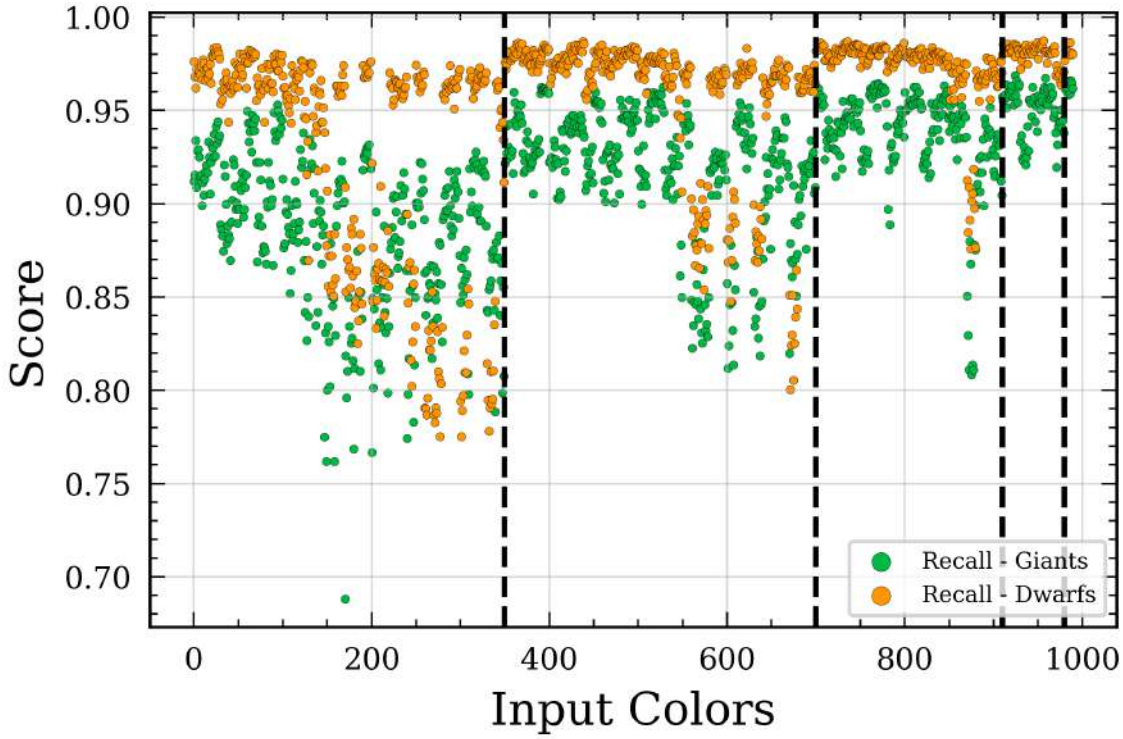


Figura 4.3: Comparación de la exhaustividad o *recall* para gigantes y enanas en las redes neuronales de clasificación con diferentes combinaciones de colores como características de entrada. El eje horizontal indica el número identificatorio único asociado a cada una de las combinaciones de colores, identificadas con números enteros entre 0 y 989. En verde se muestra la exhaustividad de las gigantes y en naranjo la exhaustividad de las enanas. Las cuatro líneas verticales segmentadas separan las combinaciones conformadas por distintas cantidades de colores: los valores del eje horizontal entre 0 y 349 indican combinaciones compuestas por cuatro colores, entre 350 y 699 combinaciones con cinco colores, entre 700 y 909 combinaciones con seis colores, entre 910 y 979 combinaciones con siete colores y entre 980 y 989 combinaciones con ocho colores. Nótese que, por lo tanto, las líneas segmentadas adicionalmente están dividiendo la gráfica según la arquitectura de la ANN utilizada para los entrenamientos y las pruebas.

Tabla 4.3: Resultados en el set de pruebas de las redes neuronales cuyos colores de entrada identifican más del 96.5 % de las estrellas gigantes del set de pruebas de manera correcta.

N°	Colores	Exhaustividad		Precisión		Exactitud	Pérdida
		Gigantes	Enanas	Gigantes	Enanas		
924	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - r)	0.968601	0.978585	0.968601	0.978585	0.974536	0.081594
926	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - z)	0.965870	0.984171	0.976536	0.976895	0.976751	0.076669
958	(F378 - u), (F395 - g), (F430 - g), (F515 - g), (F660 - r), (F861 - z), (r - z)	0.967235	0.980912	0.971879	0.977726	0.975367	0.086882
985	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F660 - r), (F861 - z), (g - i)	0.965870	0.981378	0.972509	0.976830	0.975090	0.089401
987	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F660 - r), (F861 - z), (r - i)	0.965870	0.979516	0.969842	0.976787	0.973983	0.100769

A partir de los resultados obtenidos en la evaluación, se eligió la red que utilizó como colores de entrada la combinación número 926 (de aquí en adelante ANN-C926[2]) para clasificar las fuentes estelares de S-PLUS. Si bien esta red no es la que logra identificar la mayor fracción de estrellas gigantes de la muestra, es la que las clasifica de forma más precisa y que tiene una mayor exactitud, además de tener una menor pérdida.

La Figura 4.4 muestra, de forma resumida, la arquitectura del modelo ANN-C926, además de la visualización de `tf.keras.utils.plot_model` (Keras, 2021c). Este modelo cuenta con cinco capas totalmente conectadas y, luego de cada capa densa, (a excepción de la capa de salida) cuenta con una capa de *dropout*.

La Figura 4.5 muestra la pérdida y la exactitud de este modelo para los set de entrenamiento y de validación en función del número de épocas. Cabe destacar que todos los modelos fueron entrenados en 50 épocas, valor determinado a partir de diferentes entrenamiento. Al correr los entrenamientos en un mayor número de épocas se producía sobreajuste, es decir, el modelo empezaba a memorizar los datos de entrenamiento y empezaba a presentar resultados menos precisos sobre el set de validación. Esto se deduce, ya que al asignar un mayor número de épocas, en las gráficas de rendimiento se observa un buen desempeño sobre el set de entrenamiento, pero no para el set de validación. Esto indica que el algoritmo, es decir, el modelo no es capaz de generalizar a datos nuevos.

La Figura 4.6a presenta la matriz de confusión de la red, en donde se pueden visualizar y comparar las predicciones del modelo utilizando el set de pruebas con las clases reales (clases establecidas por el algoritmo de *clustering*). Es importante recordar que estos datos no han sido utilizados en el entrenamiento, por lo que teóricamente mostrarían el desempeño real de la red. El algoritmo logró identificar correctamente 3,524 estrellas que corresponden al 97.6 % del set de pruebas. Adicionalmente, en la Tabla 4.6b se muestran los valores que alcanzan la exhaustividad, precisión y exactitud y las cantidades de estrellas en cada categoría. El algoritmo logró reconocer el 97.1 % de las estrellas gigantes y el 97.9 % de las enanas. Además, el 97.0 % de las estrellas que el modelo clasificó como gigantes y el 98.0 % de las estrellas que el modelo clasificó como enanas, estaban correctamente identificadas.

[2]Significado del código: ANN - Artificial Neural Network, C - Clasificación, 926 - número de la combinación.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization (Normalizatio  (None, 5)                11
 n)

 dense (Dense)               (None, 224)               1344

 dropout (Dropout)           (None, 224)               0

 dense_1 (Dense)             (None, 128)               28800

 dropout_1 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 384)               49536

 dropout_2 (Dropout)         (None, 384)               0

 dense_3 (Dense)             (None, 352)               135520

 dropout_3 (Dropout)         (None, 352)               0

 dense_4 (Dense)             (None, 2)                 706

=================================================================
Total params: 215,917
Trainable params: 215,906
Non-trainable params: 11
```

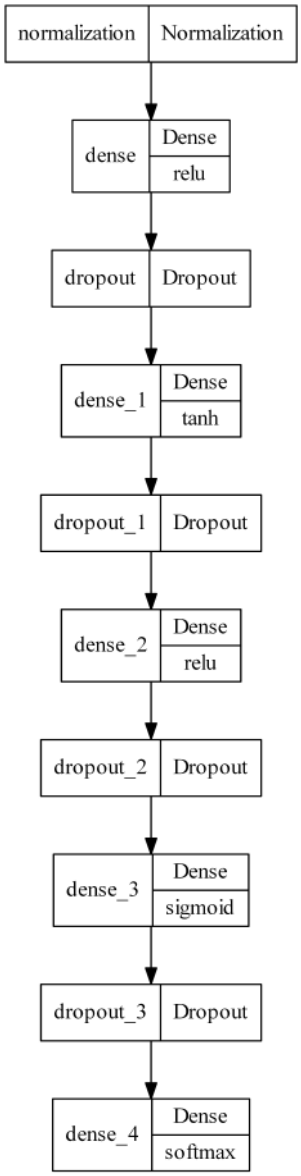


Figura 4.4: Arquitectura del modelo ANN-C926 para clasificación de estrellas. Izquierda: representación resumida del modelo, generada por medio de la función `summary()` de TensorFlow. Cada fila representa una capa con un nombre único asignado. Cada capa tiene una salida (*output*) cuya forma se muestra en la columna "Output Shape"; estas salidas son las entradas de las capas siguientes. La columna "Param #"muestra el número de parámetros entrenados en cada capa. Para las capas densas este número es igual a `output_channel_number` × `(input_channel_number + 1)`. El número total de parámetros es igual al número de parámetros entrenables y no entrenables. Los parámetros no entrenables en este modelo son los de la capa de normalización. Derecha: representación visual del modelo ANN-C926 generado con `tf.keras.utils.plot_model`, la que además muestra las funciones de activación empleadas en cada capa densa.

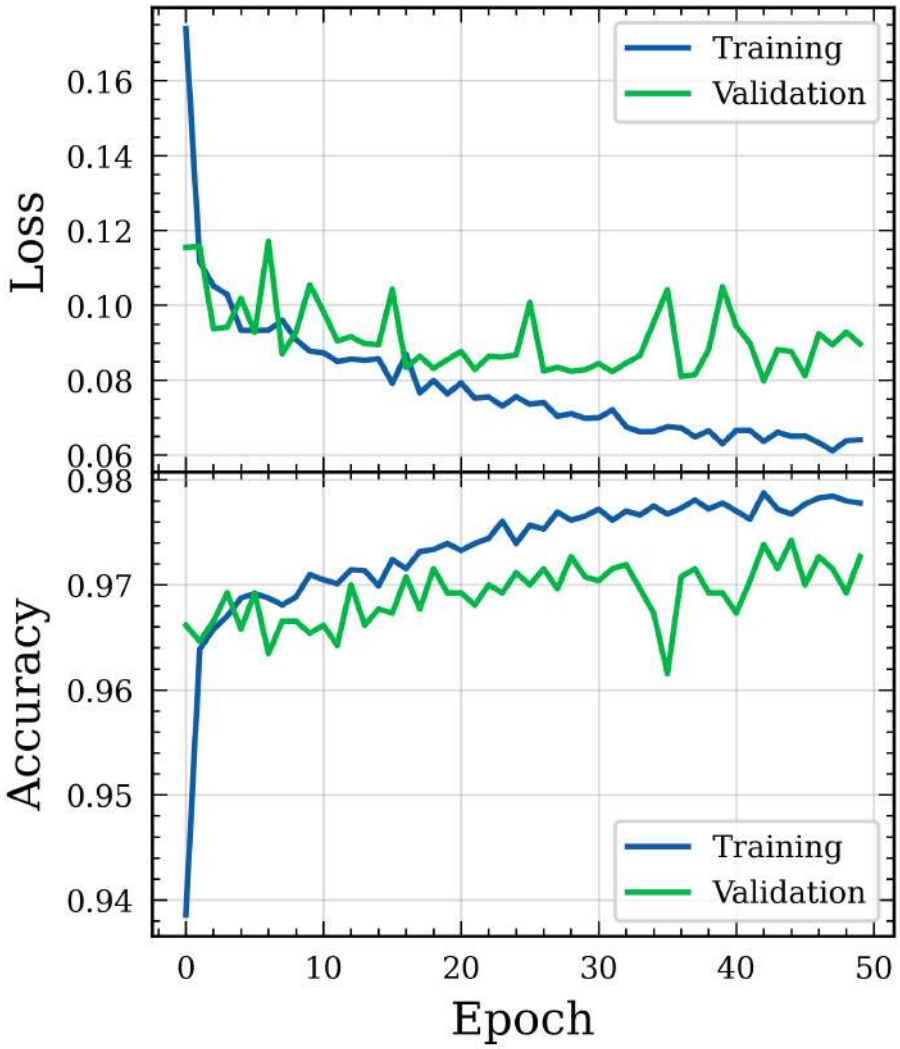
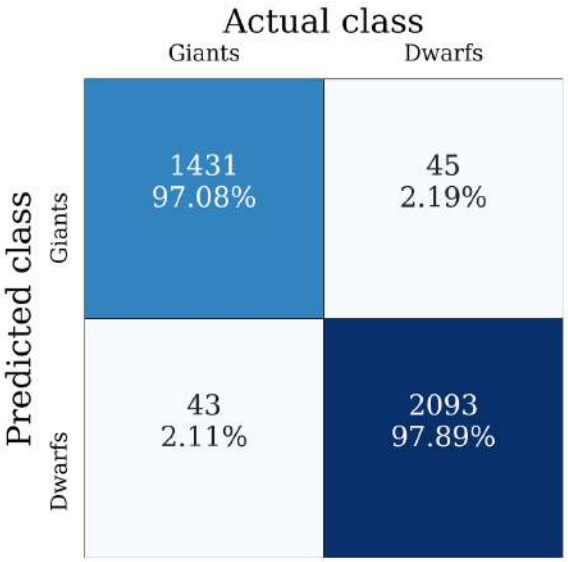


Figura 4.5: Rendimiento del modelo 926 para clasificación de estrellas en 50 épocas. En el panel superior se muestra la pérdida en cada época y en el panel inferior la exactitud del modelo por época. Ambos parámetros se muestran para el set de entrenamiento (azul) y para el set de validación (verde).

Esta red, por lo tanto, logra identificar más estrellas enanas y de forma más precisa que estrellas gigantes. De hecho, predijo que el set de datos contaba con 1,476 estrellas gigantes, siendo que este contaba con 1,474, es decir, sobrestimó ligeramente la cantidad de gigantes y, por lo tanto, subestimó la cantidad de estrellas enanas. Es por esto que las 5 redes que fueron preliminarmente seleccionadas, que logran identificar más del 96.5 % de las gigantes correctamente (ver Tabla 4.3) a pesar de identificar mejor las estrellas enanas, siguen siendo más óptimas para la identificación de las estrellas gigantes que las demás redes entrenadas.

En la Figura 4.7 se muestran las predicciones del modelo ANN-C926 en el diagrama de Kiel, con los parámetros T_{eff} y log g del set de pruebas APOGEE DR17. Se observa que, si bien la gran mayoría de las estrellas identificadas como gigantes están ubicadas en la RGB, unas pocas quedan dispersas entre la MS. Para las estrellas que la red clasificó como enanas ocurre algo similar, algunas se encuentran en la *subgiant branch*, pero sobretodo en la zona de superposición que se aprecia en la Figura 3.5. Estos errores son aceptables, ya que en esta zona ambas poblaciones poseen características similares.



(a) Matriz de confusión

Métrica	Gigantes	Enanas
Exhaustividad	0.971	0.979
Precisión	0.970	0.980
Exactitud	0.976	

Categoría	Real	Predicha
Gigantes	1,474	1,476
Enanas	2,138	2,136
Total	3,612	

(b) Métricas y número de estrellas

Figura 4.6: (a) Matriz de confusión de las predicciones de la ANN-C926 sobre el set de pruebas. (b) Arriba: métricas de la red ANN-C926 obtenidos posterior a la evaluación del set de pruebas para las estrellas gigantes y enanas. Abajo: cantidad real, predicha por la red de estrellas y número total.

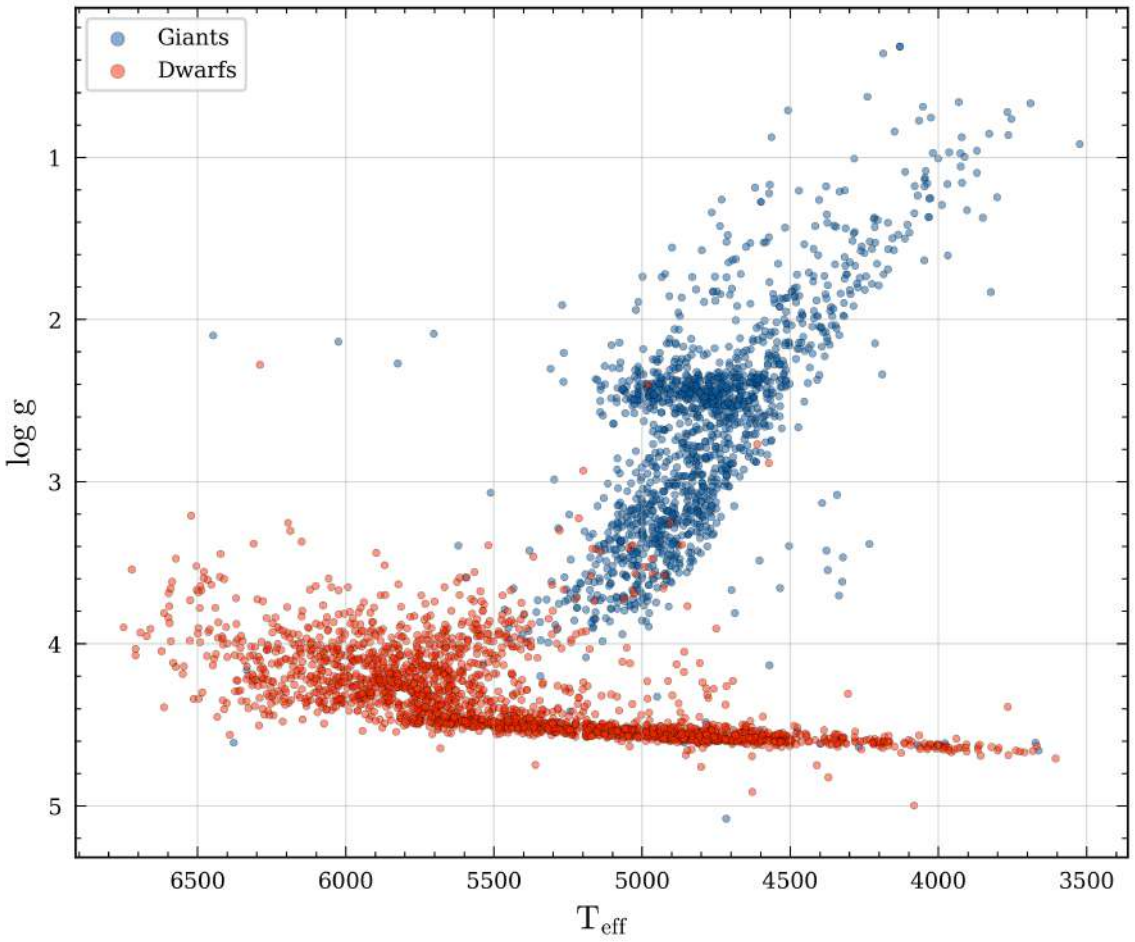


Figura 4.7: Diagrama de Kiel con las predicciones de la red neuronal ANN-C926 en el set de pruebas. Estrellas que el modelo clasificó como enanas se muestran en rojo y las que el modelo clasificó como gigantes en azul.

4.2.4. Desempeño del algoritmo de random forest

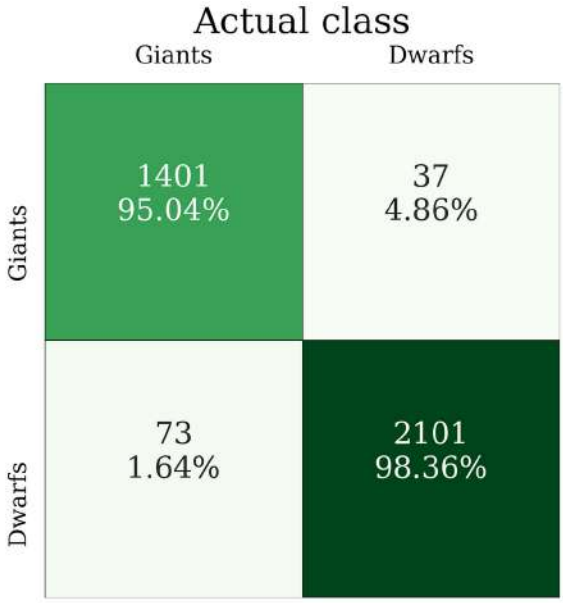
Como se mencionó en la Sección 3.2.2, se implementó un algoritmo de *random forest* para el problema de clasificación. Este algoritmo fue entrenado y probado con los mismos catálogos de entrenamiento y de pruebas utilizados en la ANN-C926 (sin el conjunto de validación, ya que no se hizo un ajuste de hiperparámetros), es decir, con las mismas estrellas, los mismos colores de entrada y mismas *labels* (ver Tabla 4.3).

El rendimiento del algoritmo se muestra en la Figura 4.8. En dicha Figura, se presenta la matriz de confusión y las métricas obtenidas luego de realizar las predicciones con las entradas del set de pruebas y compararlas con las categorías reales de cada estrella. También se presentan las cantidades de estrellas de cada clase, las predicciones de cada clase y el número total de estrellas. Este algoritmo logró identificar el 95.1 % de las estrellas gigantes y el 98.3 % de las enanas correctamente. Entre las estrellas que el modelo identificó como gigantes, el 97.4 % fueron correctamente clasificadas y el 96.7 % de las estrellas clasificadas como enanas correspondían a esta categoría.

En comparación con la red neuronal entrenada con la combinación de colores 926, ANN-C926, este algoritmo de *random forest* logra identificar una menor fracción de estrellas gigantes y una mayor de estrellas enanas (una diferencia de 30 y 8 estrellas, respectivamente). Las clasificaciones de las estrellas gigantes por el algoritmo de *random forest* son un 0.4 % más precisas y las de estrellas enanas un 1.3 % menos precisas que las clasificaciones hechas por la red neuronal. La exactitud de ambos modelo tiene una diferencia de un 0.6 %, siendo la red neuronal más exacta que el método de *random forest*.

En la Figura 4.9 se muestra la distribución de las estrellas del set de pruebas y las predicciones en el diagrama de Kiel. En comparación a la Figura 4.7, se observan menos estrellas que la red predijo como enanas en la zona de la RGB (una RGB más contaminada).

En síntesis, la red neuronal presenta un mejor rendimiento que el algoritmo de *random forest*, a pesar de que la diferencia entre los rendimientos de ambos modelos medidos con las diferentes métricas no llega a ser superior a un 2.0 %. Es por esta razón, y porque además, el rendimiento en la clasificación de las estrellas gigantes



(a) Matriz de confusión

Métrica	Gigantes	Enanas
Exhaustividad	0.951	0.983
Precisión	0.974	0.967
Exactitud	0.970	

Categoría	Real	Predicha
Gigantes	1,474	1,438
Enanas	2,138	2,174
Total	3,612	

(b) Métricas y número de estrellas

Figura 4.8: (a) Matriz de confusión de las predicciones del algoritmo de *random forest* con la combinación de colores 926 sobre el set de pruebas. (b) Arriba: métricas del algoritmo de *random forest* con la combinación de colores 926 sobre el set de pruebas para las estrellas gigantes y enanas. Abajo: cantidad real, predicha por el algoritmo y total de estrellas.

es mejor en el modelo de ANN, que se determinó que la red neuronal entrenada con la combinación de colores (F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - z) fue elegida para clasificar las estrellas de S-PLUS. Los primeros seis índices fotométricos corresponden a las líneas [OII], Ca H+K, banda G, triplete de Mg y triplete de Ca, respectivamente. El triplete de magnesio es una característica sensible a la gravedad superficial estelar, por lo que era esperable que la mejor red incluyera el color (F515 - g).

4.3. Clasificación con redes neuronales de las estrellas de S-PLUS

Al catálogo completo de estrellas (`CLASS_STAR` ¿0.92) de S-PLUS iDR3 se le aplicaron los mismos cortes de calidad que se utilizaron para determinar el catálogo de entrenamiento, de validación y de pruebas. Esto es, se restaron las estrellas con SNR < 50 en las magnitudes fotométricas y además se descartaron las estrellas con

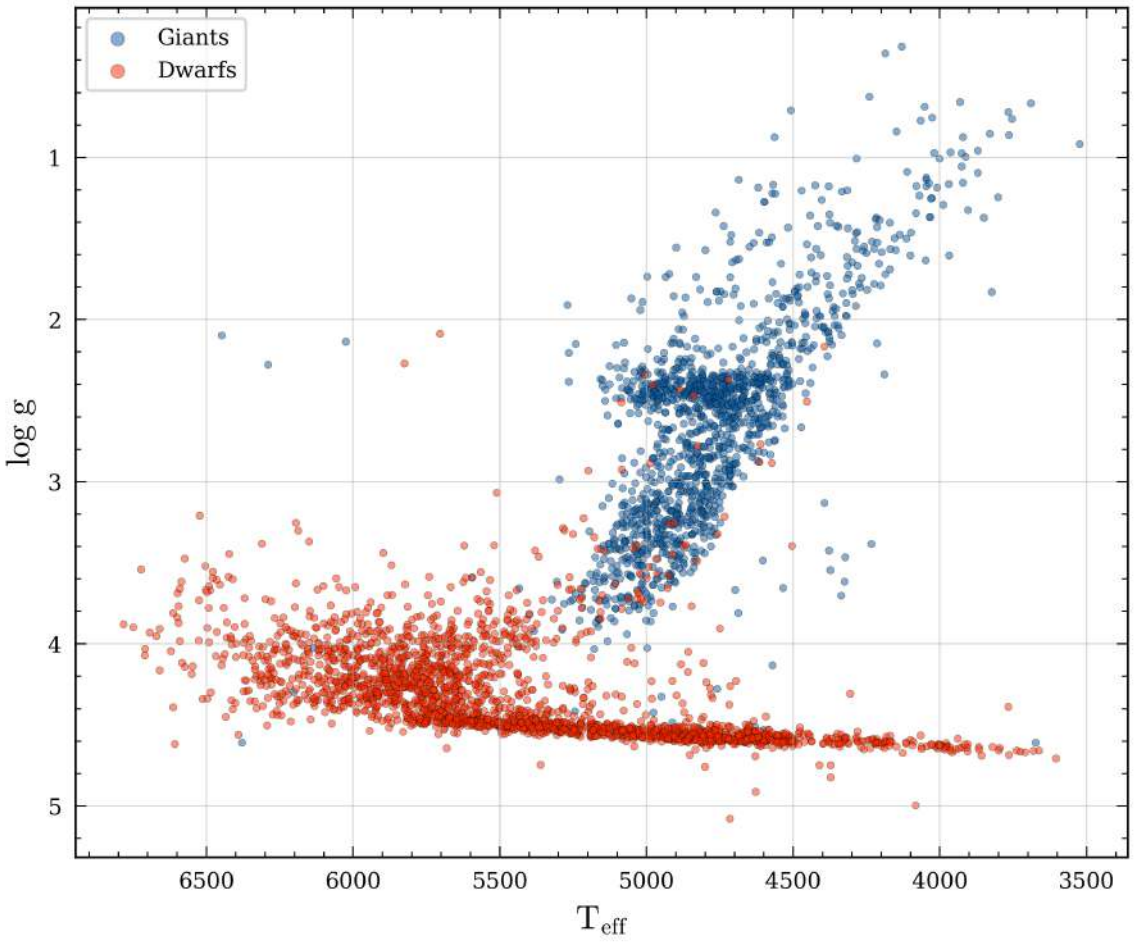


Figura 4.9: Diagrama de Kiel con las predicciones del modelo de *random forest* en el set de pruebas. Estrellas que el modelo clasificó como enanas se muestran en rojo y las que el modelo clasificó como gigantes en azul.

mediciones espúreas en las magnitudes, *i.e.* magnitudes muy grandes,quedando un catálogo con magnitud límite en r (mag) igual a 18. Con este set de datos se realizó un *cross-match* con el catálogo de distancias de Gaia EDR3 de Bailer-Jones et al. (2021). El radio de búsqueda se fijó en 1.0 arcosegundo, ya que dentro de este radio se obtuvieron coincidencias únicas para más del 90 % del catálogo. Se obtuvieron entonces las distancias de 812,378 estrellas de S-PLUS iDR3 n4. Estas estrellas fueron clasificadas entre gigantes o enanas por la red neuronal artificial ANN-C926 descrita en la Sección anterior, siguiendo el procedimiento detallado en la Sección 3.2.1.5. Los resultados de este proceso se muestran en la Tabla 4.4.

Tabla 4.4: Cantidad de estrellas de S-PLUS clasificadas como gigantes y enanas por la red neuronal ANN-C926.

	N° de estrellas	Fracción
Gigantes	130,172	0.16
Enanas	682,206	0.84
Total	812,378	1

El 84 % de las estrellas fueron clasificadas como enanas y el 16 % restante como gigantes. Este resultado es esperable, debido a que las estrellas enanas son más abundantes, considerando una magnitud límite de 18 en la banda r dadas las restricciones en señal ruido aplicadas.

Para inspeccionar los resultados más confiables se seleccionaron las estrellas cuya probabilidad de ser gigante o enana determinada por la ANN-C926 sea mayor a 0.95. La cantidad de estrellas se encuentra en la Tabla 4.5. En este conjunto de datos el 14 % de las estrellas fueron clasificadas como gigantes y el 86 % como enanas.

Tabla 4.5: Cantidad de estrellas de S-PLUS clasificadas como gigantes y enanas por la red neuronal ANN-C926 con probabilidad de pertenecer a la categoría > 0.95.

	N° de estrellas	Fracción
Gigantes	107,492	0.14
Enanas	638,695	0.86
Total	746,457	1

Con las distancias de Gaia EDR3 se determinaron las magnitudes absolutas en la

banda r y de esta forma se construyó el diagrama color-magnitud $M_{r,o}$ vs $(g - r)_o$ que se presenta en la Figura 4.10. En esta Figura se observa dónde quedan posicionadas las estrellas clasificadas como gigantes y enanas por la red ANN-C926. Se observan resultados esperables, en el sentido de que la gran parte de las estrellas gigantes se ubica en la zona superior del diagrama y las enanas están poblando la secuencia principal en la parte inferior del diagrama.

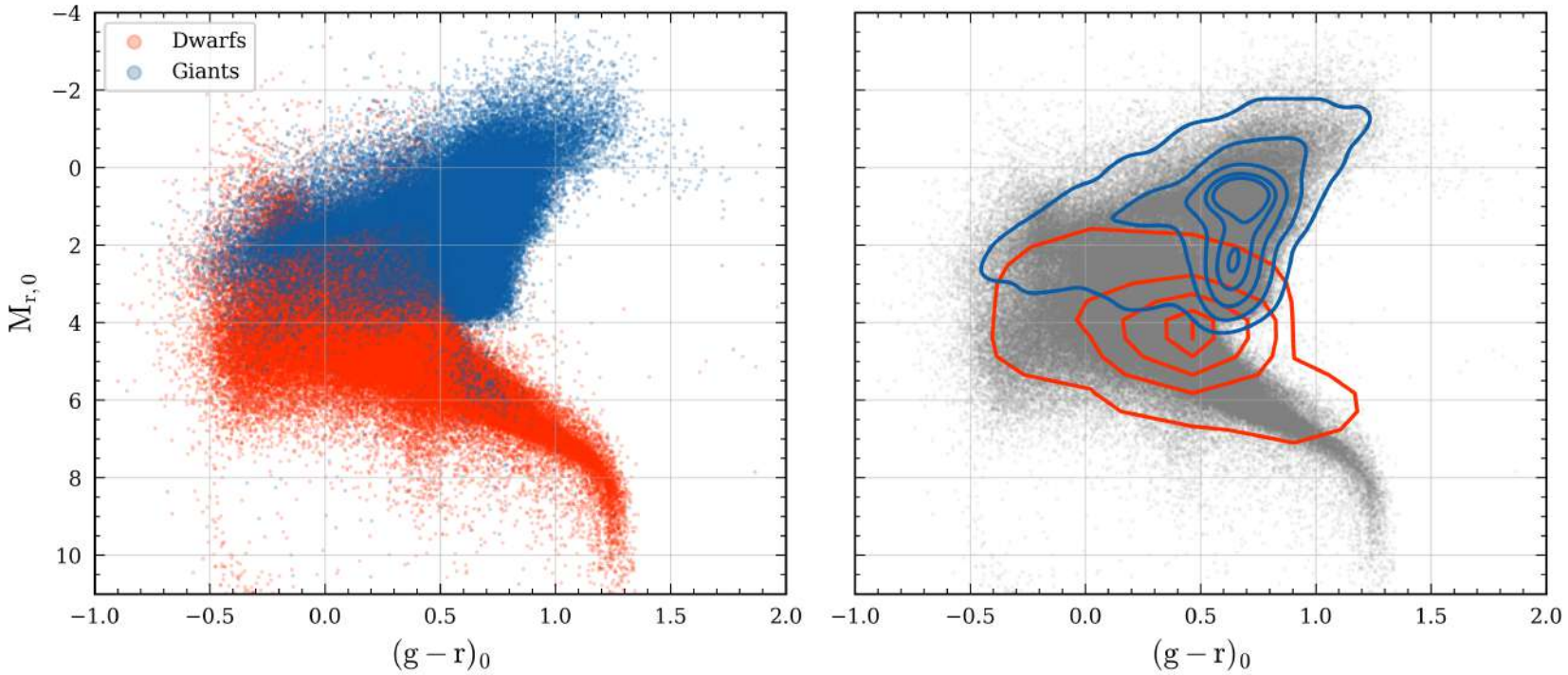


Figura 4.10: Diagrama color magnitud $M_{r,o}$ vs (g - r)$_o$ (g y r con las magnitudes de S-PLUS corregidas por extinción) de las 812,378 estrellas de la muestra de S-PLUS. Panel izquierdo: las 130,172 estrellas clasificadas como gigantes por la ANN-C926 se muestran en azul y las 682,206 estrellas clasificadas como enanas se muestran en rojo. Nótese que las estrellas gigantes se graficaron sobre las estrellas enanas. Panel derecho: las 812,378 estrellas de la muestra se presentan en gris y los contornos representan la densidad de objetos contenidos en ellos. Los contornos azules corresponden a las estrellas gigantes y los contornos rojos a las enanas.

4.4. Modelos para la determinación de metalicidades estelares

4.4.1. Catálogo de entrenamiento, de validación y de pruebas

Las ANNs construidas para determinar las metalicidades de todas las estrellas utilizaron los mismos catálogos de entrenamiento, de validación y de pruebas que los algoritmos de clasificación, pero con el valor de las metalicidades estelares de cada estrellas como salida o *label* en lugar de la categoría.

Por su parte, para las redes entrenadas sólo con estrellas gigantes, las 7,356 estrellas del catálogo descrito en la Sección 3.1.2 identificadas como gigantes por el algoritmo de *clustering*, fueron divididas entre los catálogos de entrenamiento, de validación y de pruebas de forma aleatoria en las proporciones indicadas en la Tabla 3.3. Lo mismo se hizo con las 10,706 estrellas restantes utilizadas en las ANNs que determinaron las metalicidades de estrellas enanas. El número de estrellas en cada uno de estos set utilizado en cada conjunto de estrellas se encuentra en la Tabla 4.6.

Tabla 4.6: Número de estrellas en cada set de entrenamiento, de validación y de pruebas utilizados en las redes neuronales de regresión.

	N° estrellas		
Catálogo	Sólo gigantes	Sólo enanas	Todas
Entrenamiento	5,296	7,708	13,004
Validación	589	857	1,446
Pruebas	1,471	2,141	3,612

4.4.2. Arquitectura de las redes neuronales

Para los tres conjuntos de redes neuronales, se realizó un ajuste de hiperparámetros según el procedimiento indicado en la Sección 3.3.1.2. Los hiperparámetros de las ANNs resultantes del *hypertuning* se muestran en las Tablas 4.7, 4.8 y 4.9, para los ajustes con la muestra completa, sólo las estrellas gigantes y sólo las enanas, respectivamente.

Al igual que para las redes de clasificación, en todas las redes de regresión se ajustó el *learning rate* en 0.001. Las redes ajustadas que utilizaron toda la muestra de estrellas para 5, 6, 7, y 8 colores son las mismas, esto es, tienen exactamente los mismos hiperparámetros. Este proceso se repitió para asegurarse de que estos resultados no fueran producto de algún error de digitación, sin embargo, se obtuvieron los mismos hiperparámetros. En las redes diseñadas para determinar las metalicidades de las estrellas gigantes ocurre esto mismo cuando se hizo el *hypertuning* con 5 y 6 colores de entrada. Además de esto, las redes son diferentes para los demás colores de entrada, con estructuras entre 3 capas ocultas (para 4 colores de entrada) y 5 (para 7 colores de entrada). En las redes para las estrellas enanas también se encuentran redes de estructura idéntica para 4 y 5 colores de entrada y el resto de las redes son diferentes. Estas coincidencias se atribuyen a la presencia de *callbacks*, de modo que el proceso de *hypertuning* se detuvo luego de no conseguir arquitecturas más óptimas sin seguir recorriendo el espacio de búsqueda asignado.

4.4.3. Determinación de las mejores redes neuronales

La función descrita en la Sección 3.3.1.3 generó un registro del desempeño de las ANNs entrenadas con las 990 combinaciones de colores para los 3 conjuntos de redes neuronales desarrolladas para la determinación de metalicidades estelares.

El desempeño de estos tres conjuntos, según el error absoluto medio o MAE, se muestra en la Figura 4.11. En esta Figura se observa que, en general, las redes entrenadas y validadas con estrellas enanas presentan errores menores, mientras que la distribución de errores en las redes entrenadas y validadas con estrellas gigantes

Tabla 4.7: Hiperparámetros determinados por KerasTuner para redes neuronales con las combinaciones de colores indicadas en la Tabla 3.4. En este caso utilizando datos de las estrellas gigantes y enanas.

Número de colores de entrada		4	5	6	7	8
	Capas ocultas	3	5	5	5	5
1° capa	Unidades	96	96	96	96	96
	Función de activación	relu	relu	relu	relu	relu
	Fracción de dropout	0.1	0.1	0.1	0.1	0.1
2° capa	Unidades	384	256	256	256	256
	Función de activación	relu	relu	relu	relu	relu
	Fracción de dropout	0.2	0.0	0.0	0.0	0.0
3° capa	Unidades	288	288	288	288	288
	Función de activación	sigmoid	relu	relu	relu	relu
	Fracción de dropout	0.0	0.1	0.1	0.1	0.1
4° capa	Unidades	-	352	352	352	352
	Función de activación	-	sigmoid	sigmoid	sigmoid	sigmoid
	Fracción de dropout	-	0.1	0.1	0.1	0.1
5° capa	Unidades	-	32	32	32	32
	Función de activación	-	relu	relu	relu	relu
	Fracción de dropout	-	0.0	0.0	0.0	0.0
	Tasa de aprendizaje	0.001	0.001	0.001	0.001	0.001

Tabla 4.8: Hiperparámetros determinados por KerasTuner para redes neuronales con las combinaciones de colores indicadas en la Tabla 3.4. En este caso utilizando datos de estrellas gigantes.

Número de colores de entrada		4	5	6	7	8
	Capas ocultas	3	4	4	5	4
1° capa	Unidades	96	192	192	96	224
	Función de activación	relu	relu	relu	relu	relu
	Fracción de dropout	0.1	0.1	0.1	0.1	0.0
2° capa	Unidades	384	64	64	256	128
	Función de activación	relu	tanh	tanh	relu	tanh
	Fracción de dropout	0.2	0.1	0.1	0.0	0.2
3° capa	Unidades	288	384	384	288	384
	Función de activación	sigmoid	sigmoid	sigmoid	relu	relu
	Fracción de dropout	0.0	0.2	0.2	0.1	0.1
4° capa	Unidades	-	128	128	352	352
	Función de activación	-	sigmoid	sigmoid	sigmoid	sigmoid
	Fracción de dropout	-	0.1	0.1	0.1	0.1
5° capa	Unidades	-	-	-	32	-
	Función de activación	-	-	-	relu	-
	Fracción de dropout	-	-	-	0.0	-
	Tasa de aprendizaje	0.001	0.001	0.001	0.001	0.001

Tabla 4.9: Hiperparámetros determinados por KerasTuner para redes neuronales con las combinaciones de colores indicadas en la Tabla 3.4. En este caso utilizando datos de estrellas enanas.

Número de colores de entrada		4	5	6	7	8
	Capas ocultas	4	4	4	3	5
1° capa	Unidades	192	192	224	96	96
	Función de activación	relu	relu	relu	tanh	relu
	Fracción de dropout	0.1	0.1	0.0	0.1	0.1
2° capa	Unidades	64	64	128	224	256
	Función de activación	tanh	tanh	tanh	tanh	relu
	Fracción de dropout	0.1	0.1	0.2	0.0	0.0
3° capa	Unidades	384	384	384	384	288
	Función de activación	sigmoid	sigmoid	relu	relu	relu
	Fracción de dropout	0.2	0.2	0.1	0.0	0.1
4° capa	Unidades	128	128	352	-	352
	Función de activación	sigmoid	sigmoid	sigmoid	-	sigmoid
	Fracción de dropout	0.1	0.1	0.1	-	0.1
5° capa	Unidades	-	-	-	-	32
	Función de activación	-	-	-	-	relu
	Fracción de dropout	-	-	-	-	0.0
	Tasa de aprendizaje	0.001	0.001	0.001	0.001	0.001

presentan mayor variación. Este diagrama también da una primera noción de que se estaría cumpliendo el supuesto que entrenar ANNs por separado para estrellas gigantes y enanas daría mejores resultados que utilizando todas las estrellas a la vez. Esto se debería a que se observan varias redes de estrellas enanas y algunas de estrellas gigantes con errores menores que las redes que trabajaron con todas las estrellas. Sin embargo, al igual que con la Figura 4.3 para el problema de clasificación, este diagrama no entrega por sí solo mayores detalles y es necesario revisar el registro tabular con las cifras de los errores, y así poder determinar cuáles redes resultan mejores para abordar este problema de regresión.

Analizando este registro se encontró que en 463 casos, correspondientes al 46.7 % del total, las redes que se entrenaron sólo con estrellas enanas y sólo con gigantes con tenían errores absolutos medios menores que las entrenadas con todas las estrellas. Adicionalmente, hubo 6 redes que obtuvieron errores medios menores a 0.1. Entre éstas, 5 pertenecían al conjunto que trabajó sólo con estrellas enanas y una al que trabajó sólo con estrellas gigantes. Esto último quiere decir que, efectivamente, se pueden obtener mejores resultados en la determinación de metalicidades trabajando ambas categorías por separado. Los desempeños de las redes con errores absolutos medios menores para cada conjunto se muestran en la Tabla 4.10, junto con los colores de entrada utilizados en ellas. Cabe recordar que se utilizó el error absoluto medio como función de pérdida. Notar que las redes con la combinación de colores de entrada 926: (F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z) y (g - z), los primeros seis correspondientes a las características espectrales [OII], Ca H+K, Hδ, banda G, triplete de Mg y triplete de Ca, respectivamente, es la que tiene mejor desempeño dentro del conjunto de redes que utilizaron sólo las estrellas enanas y del conjunto que usó todos los datos (y también es la combinación que da mejores resultados en el problema de clasificación). Esto no ocurre en el caso de las redes entrenadas y validadas sólo con estrellas gigantes, en donde la combinación de colores que derivó en un mejor desempeño para la determinación de metalicidades fue la 980: (F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F660 - r), (F861 - z) y (u - g), los primeros siete correspondientes a las características espectrales [OII], Ca H+K, Hδ, banda G, triplete de Mg, Hα y triplete de Ca, respectivamente. En ambos casos están presentes las características Ca H+K, triplete de Mg y triplete de Ca que son sensibles a la metalicidad. Con respecto a la Figura 4.11 es importante también destacar que aun utilizando solamente cuatro colores de entrada, lo que es

conveniente pensando en estrategias de observación, se pueden obtener predicciones con errores absolutos medios incluso menores a 0.12 dex, un valor bastante preciso para tratarse de un método fotométrico.

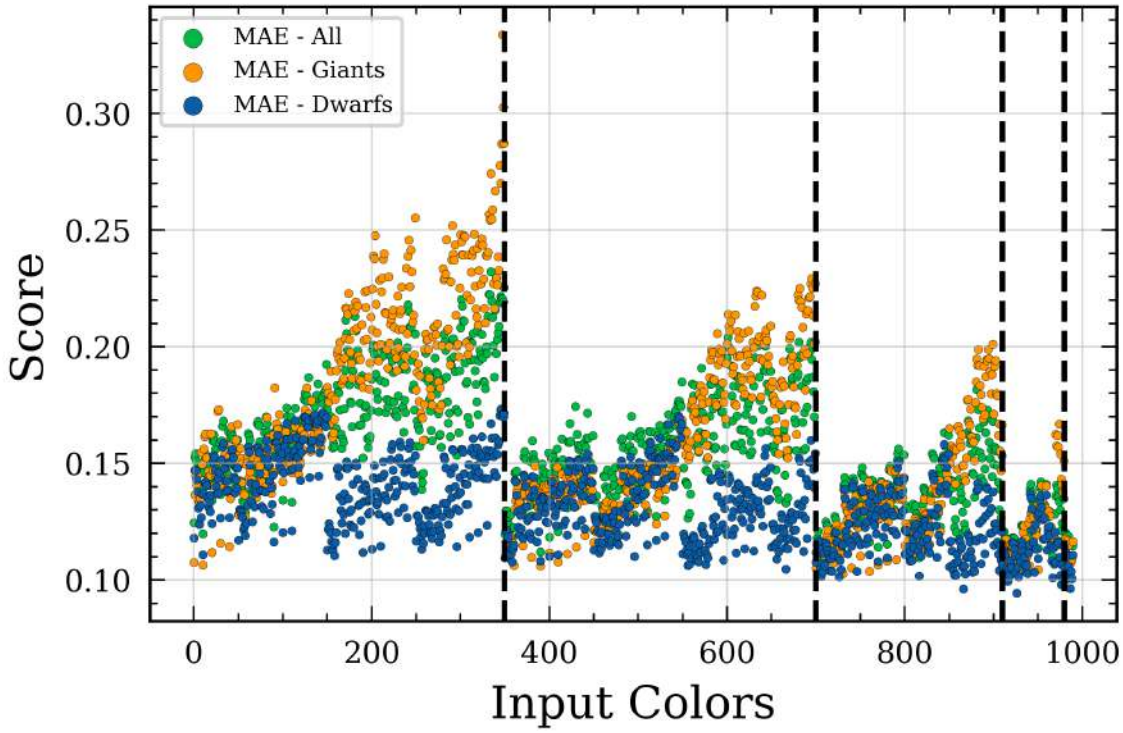


Figura 4.11: Comparación del error absoluto medio en las predicciones de metalicidades hechas por los tres conjuntos de redes neuronales desarrolladas para el problema de regresión: las que ocupan toda la muestra de estrellas (verde), sólo estrellas gigantes (naranjo) y sólo estrellas enanas (azul). El eje horizontal indica el número asociado a cada una de las combinaciones de colores, identificadas con números enteros entre 0 y 989. Las cuatro líneas verticales segmentadas separan las combinaciones conformadas por distintas cantidades de colores: los valores del eje horizontal entre 0 y 349 indican combinaciones compuestas por cuadro colores, entre 350 y 699 combinaciones con cinco colores, entre 700 y 909 combinaciones con seis colores, entre 910 y 979 combinaciones con siete colores y entre 980 y 989 combinaciones con ocho colores. Notar que, por lo tanto, las líneas segmentadas adicionalmente están dividiendo la gráfica según la arquitectura de la ANN utilizada para los entrenamientos y las pruebas.

Luego de seleccionar las redes con mejor desempeño, se derivaron las metalicidades del set de pruebas por separado para estrellas gigantes y enanas, para evaluar su desempeño en datos que las redes no han visto antes. Para ello, se usaron las redes que ocupan las combinaciones de colores 980 para las gigantes y 926 para las enanas (de aquí en adelante ANN-R980 y ANN-R926[3]). Las Figuras 4.12 y 4.13 muestran el

[3]Significado del código: ANN - Artificial Neural Network, R - regresión, 980 y 926 corresponden al número de la combinación de colores.

Tabla 4.10: Resultados de las redes neuronales con menores errores absolutos medios en las predicciones de metalicidades estelares entrenadas y evaluadas con estrellas gigantes, con estrellas enanas y con todo el set de datos.

Estrellas	N°	Colores	MAE	MSE
Todas	926	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - z)	0.102571	0.021753
Gigantes	980	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F660 - r), (F861 - z), (u - g)	0.099655	0.024228
Enanas	926	(F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - z)	0.094407	0.015514

resumen y esquema de los modelos ANN-R980 y ANN-R926, respectivamente. Estos diagramas complementan la información otorgada en las Tablas 4.8 y 4.9.

En la Figura 4.14 se presenta la función de pérdida, MAE, en los set de entrenamiento y validación en 50 épocas de la red ANN-R980 para las estrellas gigantes y, de manera similar, en la Figura 4.15 se muestra este diagrama para la red ANN-R926 de estrellas enanas. En este caso, el error absoluto medio está comparando los valores de metalicidad predichos por la red con las metalicidades espectroscópicas de APOGEE DR17. Notar que al realizar los entrenamientos en más épocas se producía sobreajuste. En la curva de aprendizaje del modelo ANN-R980 se observa un ajuste apropiado, donde las pérdidas del entrenamiento y la validación disminuyen hasta un punto estable con diferencias pequeñas entre las mismas. En cambio, para el caso de la ANN-R926 se observa una curva de entrenamiento que va disminuyendo hasta alcanzar una estabilidad y una curva de validación ruidosa. Este tipo de curvas se producen cuando el set de validación es poco representativo para poder evaluar la capacidad de generalización del modelo. Un set de validación poco representativo, por ejemplo, es aquel tiene pocos datos, que es la causa más probable en este caso. Esto a pesar de que las estrellas enanas son más que las estrellas gigantes, pues ambos conjuntos de datos son diferentes y con distribuciones de metalicidad diferentes como se observa en la Figura 4.2.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization (Normalizatio  (None, 7)                15
 n)

 dense (Dense)               (None, 96)                768

 dropout (Dropout)           (None, 96)                0

 dense_1 (Dense)             (None, 256)               24832

 dropout_1 (Dropout)         (None, 256)               0

 dense_2 (Dense)             (None, 288)               74016

 dropout_2 (Dropout)         (None, 288)               0

 dense_3 (Dense)             (None, 352)               101728

 dropout_3 (Dropout)         (None, 352)               0

 dense_4 (Dense)             (None, 32)                11296

 dropout_4 (Dropout)         (None, 32)                0

 dense_5 (Dense)             (None, 1)                 33

=================================================================
Total params: 212,688
Trainable params: 212,673
Non-trainable params: 15
_________________________________________________________________
```

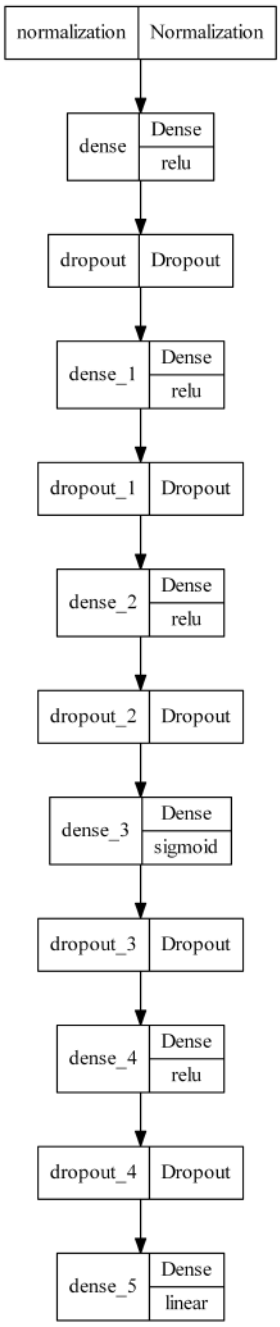


Figura 4.12: Estructura del modelo ANN-R980 para determinación de metalicidades de estrellas gigantes. Izquierda: resumen del modelo con la función `summary()`. Cada fila representa una capa con un nombre único. Cada capa tiene una salida (*output*) cuya forma se muestra en la columna "Output Shape", estas salidas son las entradas de las capas siguientes. La columna "Param #"muestra el número de parámetros entrenados en cada capa. Para las capas densas este número es igual a `output_channel_number` $\times$ `(input_channel_number + 1)`. El número total de parámetros es igual al número de parámetros entrenables y no entrenables. Los parámetros no entrenables en este modelo son los de la capa de normalización. Derecha: visualización del modelo ANN-R980 con `tf.keras.utils.plot_model`. Muestra las funciones de activación de cada capa densa.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization (Normalizatio  (None, 7)                15
 n)

 dense (Dense)               (None, 96)                768

 dropout (Dropout)           (None, 96)                0

 dense_1 (Dense)             (None, 224)               21728

 dropout_1 (Dropout)         (None, 224)               0

 dense_2 (Dense)             (None, 384)               86400

 dropout_2 (Dropout)         (None, 384)               0

 dense_3 (Dense)             (None, 1)                 385

=================================================================
Total params: 109,296
Trainable params: 109,281
Non-trainable params: 15
_________________________________________________________________
```

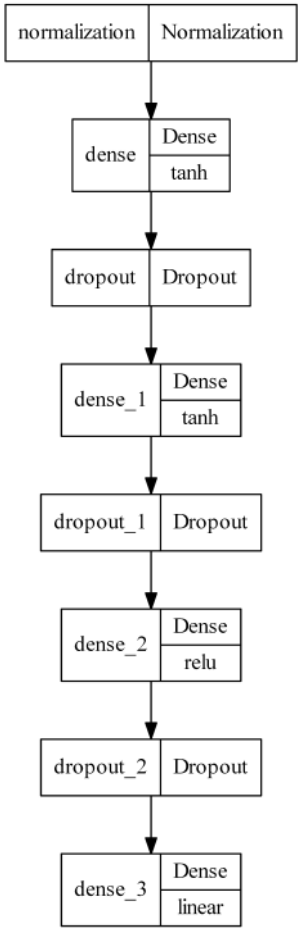


Figura 4.13: Estructura del modelo ANN-R926 para la determinación de metalicidades de estrellas enanas. Izquierda: representación resumida del modelo, generada por medio de la función `summary()`. Cada fila representa una capa con un nombre único. Cada capa tiene una salida (*output*) cuya forma se muestra en la columna "Output Shape", estas salidas son las entradas de las capas siguientes. La columna "Param #"muestra el número de parámetros entrenados en cada capa. Para las capas densas este número es igual a `output_channel_number` $\times$ `(input_channel_number + 1)`. El número total de parámetros es igual al número de parámetros entrenables y no entrenables. Los parámetros no entrenables en este modelo son los de la capa de normalización. Derecha: visualización del modelo ANN-R926 con `tf.keras.utils.plot_model`. Muestra las funciones de activación de cada capa densa.

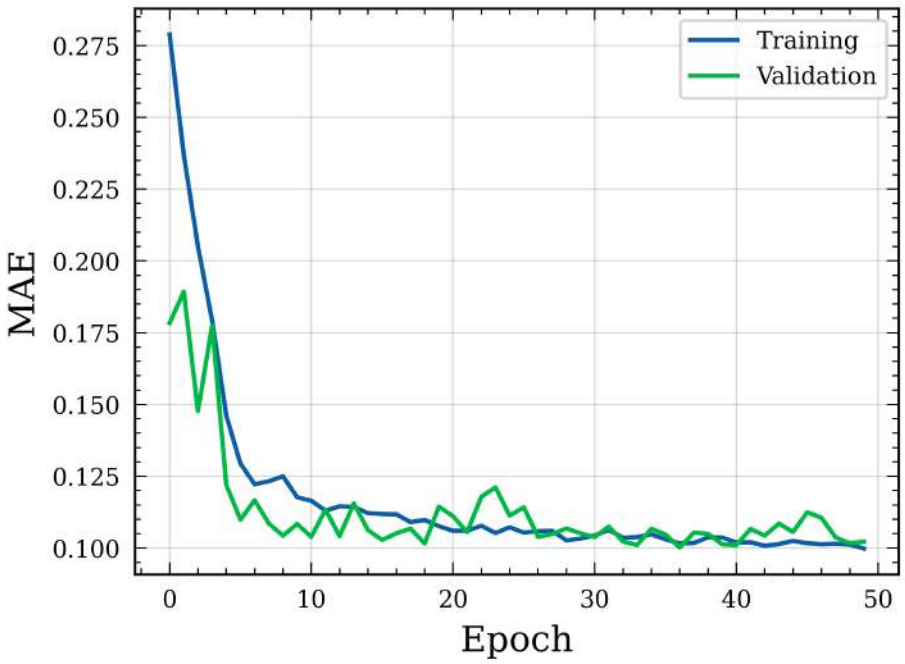


Figura 4.14: Curva de aprendizaje del modelo ANN-R980 para determinación de metalicidades de estrellas gigantes en 50 épocas. La función de pérdida utilizada es el error absoluto medio. Se muestra para el set de entrenamiento (azul) y para el set de validación (verde).

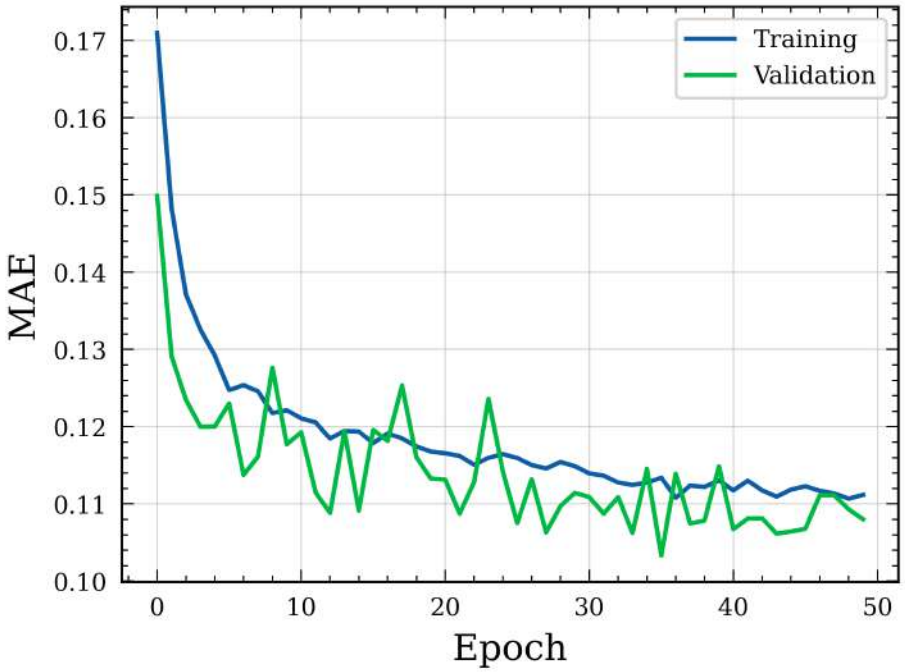


Figura 4.15: Curva de aprendizaje del modelo ANN-R926 para determinación de metalicidades de estrellas enanas en 50 épocas. La función de pérdida utilizada es el error absoluto medio. Se muestra para el set de entrenamiento (azul) y para el set de validación (verde).

Al realizar las predicciones sobre el set de pruebas se encontró que para la red ANN-R926 los errores se presentan con una inclinación, como se observa en la Figura 4.16. Es por esto que se realizó nuevamente el proceso de *hypertuning* para la misma combinación de colores establecida, pero esta vez sin el *callback*, esperando que el ajuste encontrara una arquitectura con mejores resultados dentro de todo el espacio de parámetros asignado. La arquitectura encontrada fue idéntica a de la red ANN-R980. A esta red de aquí en adelante se le denominará ANN-R926-2.

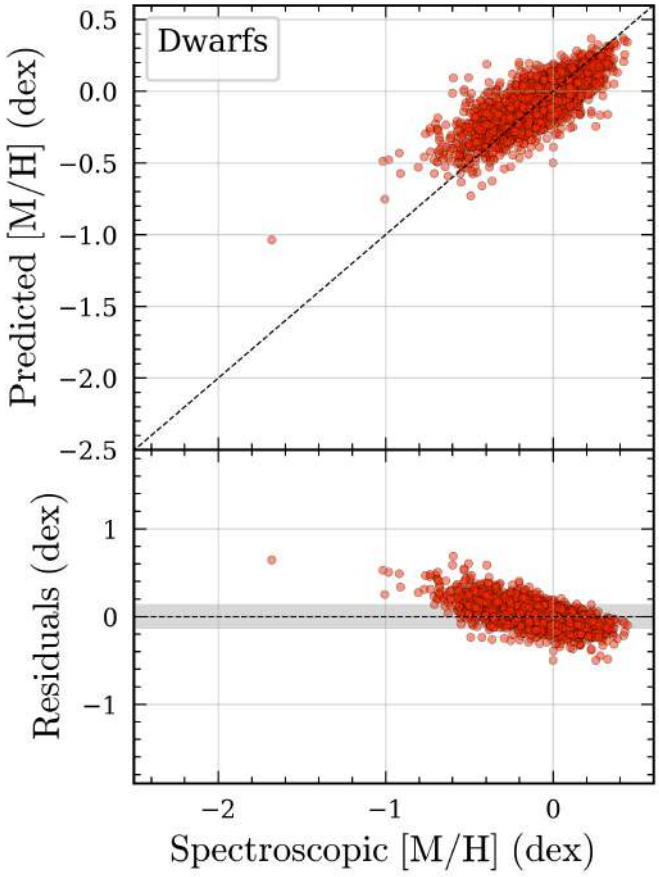


Figura 4.16: Resultados de la aplicación del modelo ANN-R926 para estrellas enanas en el set de pruebas. El panel superior muestra las predicciones versus las metalicidades espectroscópicas y el panel inferior los residuos y en gris el intervalo de ± 1 desviación estándar. Los residuos fueron calculados como las diferencias entre las predicciones y las metalicidades espectroscópicas. La desviación estándar es igual a 0.14 (dex).

En la Figura 4.17 se presentan los resultados de la aplicación de los modelos ANN-R980 y ANN-R926-2 para gigantes y enanas sobre el set de pruebas, respectivamente. En ambos paneles se configuraron los mismos límites en los ejes horizontal y vertical para tener una mejor comparación visual. Tal como se esperaba según el análisis

de las métricas en los set de validación, la ANN-R980 da errores mayores. En ambos casos, las predicciones de metalicidades de las estrellas más pobres en metales tienden a ser sobreestimadas y las más ricas en metales subestimadas. Las desviaciones estándar sobre las gigantes y enanas son $\sigma_{giants} \sim 0.15$ (dex) y $\sigma_{dwarfs} \sim 0.13$ (dex), respectivamente. Obsérvese que la mayoría de las estrellas enanas son ricas en metales, donde sólo cuatro estrellas tienen [M/H] < -1 (dex). Con respecto a las estrellas enanas, la nueva arquitectura atenúa el problema observado en la Figura 4.16 y también disminuye la desviación estándar.

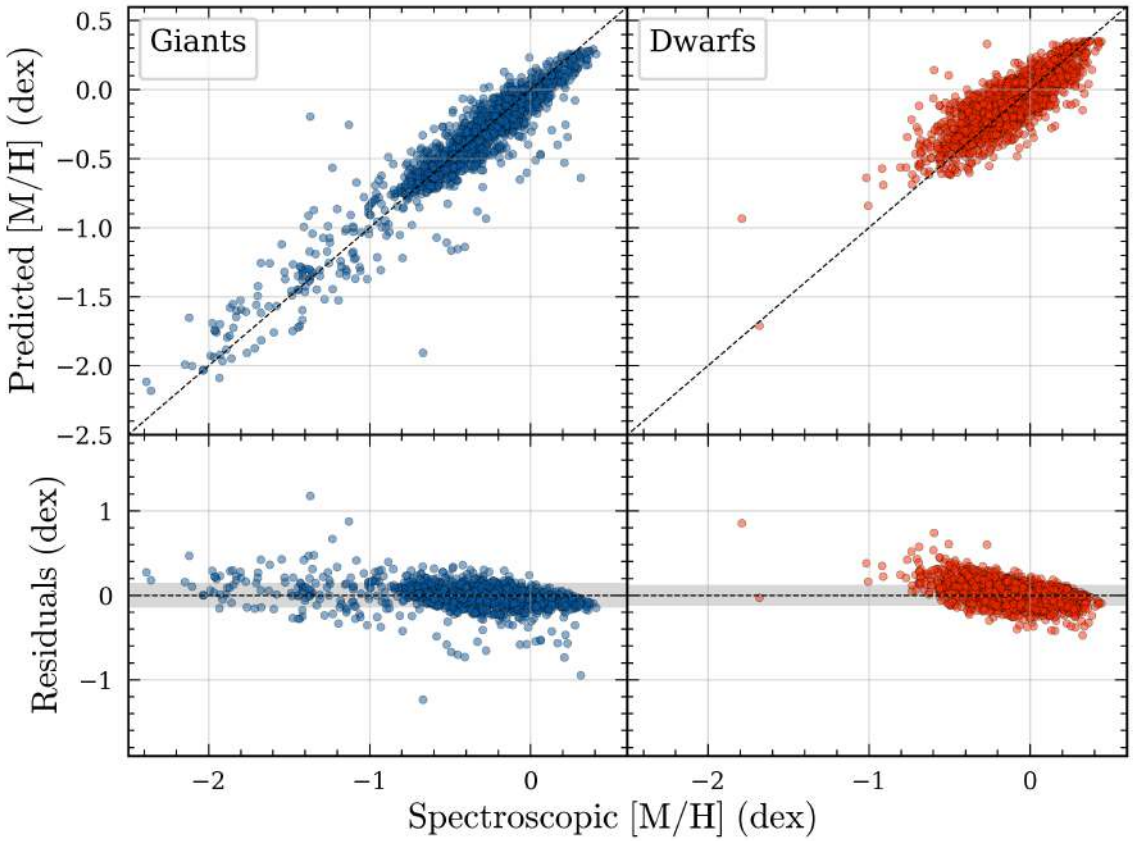


Figura 4.17: Resultados de la aplicación de los modelos ANN-R926-2 para estrellas enanas (derecha, en rojo) y ANN-R980 (izquierda, en azul) en el set de pruebas. Los paneles superiores muestran las predicciones versus las metalicidades espectroscópicas y los paneles inferiores los residuos y en gris el intervalo de ± 1 desviación estándar. Los residuos fueron calculados como las diferencias entre las predicciones y las metalicidades espectroscópicas. La desviación estándar es igual a 0.15 (dex) para las gigantes y 0.13 (dex) para las enanas.

4.5. Aplicación del modelo de regresión a estrellas de S-PLUS

Se determinaron las metalicidades fotométricas de las 812,378 estrellas de S-PLUS iDR3 n4, que fueron clasificadas entre gigantes y enanas por las redes neuronales artificiales (ver Sección 4.3). Para las 682,206 estrellas que fueron clasificadas como enanas se utilizó la ANN-R926-2 y para las 130,172 clasificadas como gigantes se utilizó la ANN-R980, ambas redes descritas en la Sección 4.4.3.

En la Figura 4.18 se muestran las densidades de probabilidad de metalicidad para estrellas gigantes y enanas. Se puede observar cómo las estrellas gigantes están aproximadamente distribuidas entre -2.2 y 0.3 dex, mientras que las enanas se distribuyen principalmente entre -0.9 y 0.4 dex. Tanto las estrellas gigantes como las enanas están ubicadas dentro de los rangos del set de entrenamiento (ver Figura 4.2). Las estrellas gigantes presentan un peak en $\sim$ -1.35 dex y un doble peak en $\sim$ -0.4 y -0.1 dex. Las enanas presentan un doble peak en $\sim$ -0.3 y -0.15 dex. En este último caso se hace evidente que la red no está extrapolando, es decir, no está identificando metalicidades estelares fuera del set que se utilizó para entrenarla. Esta es la razón más probable de por qué las estrellas enanas de S-PLUS se muestran tan ricas en metales, porque la red ANN-R926-2 no está identificando las estrellas pobres en metales. Hay que considerar también que un set de datos limitado en magnitud (como lo es el caso de los set de entrenamiento y del conjunto de datos utilizado en esta Sección) las estrellas gigantes serán detectadas a mayores distancias daba su mayor luminosidad, por lo que es esperable encontrar más gigantes pobres en metales del halo que enanas.

Dado que las incertidumbres (errores) de estas estimaciones son desconocidas (sólo se cuenta con los errores en el set de pruebas), se siguió un procedimiento que consiste en añadir ruido a las magnitudes teniendo en cuenta su incerteza para poder estimar los errores de la red sobre estas magnitudes artificiales. Se tomaron en total 1319 estrellas gigantes y 1472 estrellas enanas aleatorias distribuidas entre todo el rango de metalicidades de cada muestra. A cada una de las magnitudes de estas estrellas se les añadió ruido 20 veces de forma aleatoria pero de modo que siguieran una distribución normal para 10 escalas o desviaciones estándar distintas. El rango de valores cubiertos

para el ruido añadido a las magnitudes fotométricas fue hasta 0.2 mag. Este rango fue definido por el mayor valor encontrado al momento de inspeccionar la relación SNR vs σ_{mag}. En total para cada estrella se generaron 200 estrellas artificiales, es decir, con ruido añadido artificialmente en sus magnitudes. Cabe destacar que la escala del ruido añadido a las magnitudes de los doce filtros fue la misma para cada estrella artificial por simplicidad. Se calcularon los colores y se realizaron las predicciones de metalicidad con las redes neuronales ANN-R980 y ANN-R926-2, para los set de estrellas gigantes y enanas artificiales, respectivamente. Se calcula la desviación estándar de estas predicciones respecto a la metalicidad predicha originalmente por las ANNs. Las desviaciones estándar de las metalicidades, junto con la escala de ruido añadida a las magnitudes y la metalicidad predicha por la red para las estrellas originales se presentan en las Figuras 4.19 y 4.20, para los sets de estrellas gigantes y enanas, respectivamente. En cuanto a las estrellas enanas, se observa que los errores en los cálculos de metalicidad aumentan mientras mayores son los errores inducidos en las magnitudes y a menores metalicidades. En cuanto a las estrellas gigantes, los errores en las metalicidades aumentan mientras mayores son los errores inducidos a las magnitudes, particularmente en metalicidades entre -1.8 y -1.6 dex y entre -1.0 y -0.8 dex. Los valores típicos encontrados en las incertidumbres de metalicidades a partir de estas pruebas son 0.046 dex para las estrellas gigantes y 0.041 dex para las enanas.

En la siguiente Sección se muestran las distribuciones espaciales de las metalicidades; esto a su vez ayudaría a entender si las metalicidades determinadas por las ANNs tienen sentido astrofísico.

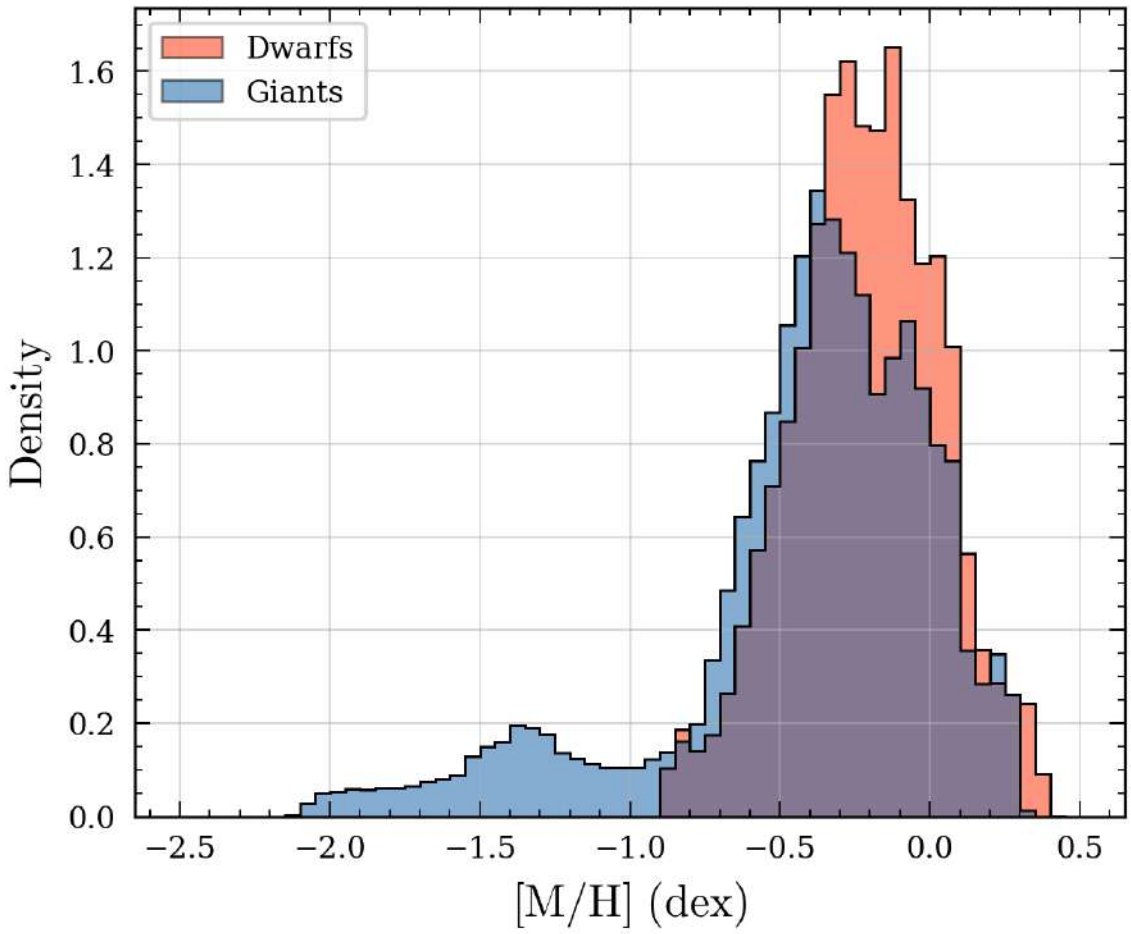


Figura 4.18: Distribución de metalicidades de las estrellas gigantes (azul) y enanas (rojo) de S-PLUS determinadas con las redes neuronales artificiales ANN-R980 y ANN-R926-2, respectivamente.

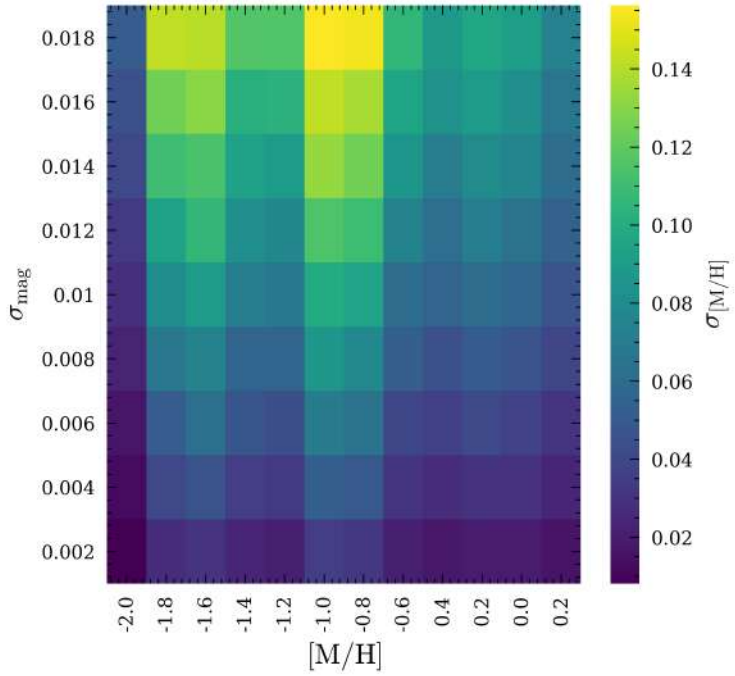


Figura 4.19: Errores en la predicción de metalicidades del set de estrellas gigantes artificiales.

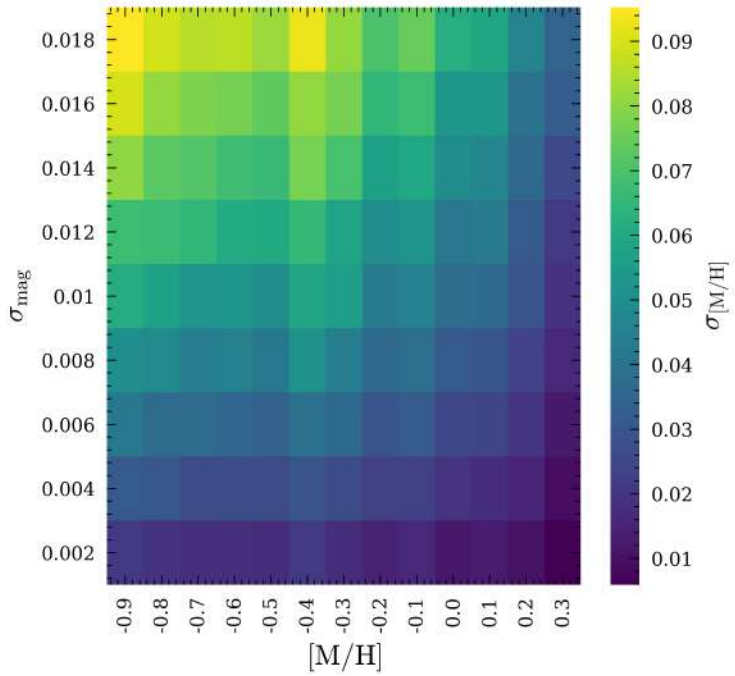


Figura 4.20: Errores en la predicción de metalicidades del set de estrellas enanas artificiales.

4.6. Distribución espacial de metalicidades de estrellas gigantes y enanas de S-PLUS

En esta sección se presenta la distribución espacial de metalicidad de las estrellas de S-PLUS iDR3 n4, determinadas por medio de redes neuronales artificiales.

Luego de obtener las metalicidades de las 130,172 estrellas gigantes y de las 682,206 estrellas enanas, a partir de las coordenadas Galácticas l y b y las distancias de Bailer-Jones et al. (2021), se pudieron obtener las coordenadas X, Y, Z y R, indicadas en las Ecuaciones 3.5. Esta información permite obtener la distribución espacial de metalicidad de las estrellas de S-PLUS en la Galaxia.

En las Figuras 4.21, 4.22 y 4.23 se muestra la distribución de las estrellas gigantes y enanas de S-PLUS en los planos (X,Z), (Y,Z) y (X,Y), respectivamente, junto con sus metalicidades en el mapa de color. En primer lugar, se observa que la distribución en el espacio de las estrellas gigantes es más amplia que la de las estrellas enanas. Esto es, las enanas se encuentran sobre todo a unos 4 kpc alrededor del Sol, aunque se ven algunas incluso a las distancias más altas, pero estas son pocas en relación a la cantidad de estrellas gigantes más lejanas. Esto se explica porque la muestra es limitada en magnitud y por ende las estrellas gigantes son detectables a mayores distancias, dada su mayor luminosidad.

En cuanto a las metalicidades, se observa que tanto estrellas enanas como gigantes presentan metalicidades similares cerca del Sol. En ambos casos, las estrellas se vuelven más pobres en metales al alejarse del Sol, sobre todo al aumentar |Z| (ver Figuras 4.21 y 4.22). Esto se observa más evidentemente en las estrellas gigantes. Considerando que los discos fino y grueso de la Vía Láctea tienen una altura de escala de $h_Z \sim 300$ pc y ~ 900 pc (Jurić et al., 2008), respectivamente, en las Figuras 4.21 y 4.22 las estrellas más pobres en metales están ubicadas en el halo de la Galaxia.

La Figura 4.24 es complementaria a las Figuras anteriores, mostrando las densidades de probabilidad de las abundancias en distintos rangos de R y Z. En cuanto a las estrellas gigantes, en los paneles inferiores, $|Z| < 4$ kpc y $R < 10$ kpc tienden a ser más ricas en metales con peaks alrededor de $\sim$ -0.4 dex, sin embargo, en los paneles

entre los rangos $2 < |Z| < 4$ kpc presentan un peak adicional en $\sim$ -1.4 dex menos dominante. Esto indicaría que en la cercanía del Sol (paneles $|Z| < 2$ kpc y $6 < R < 10$ kpc) la mayor parte de las estrellas observadas pertecen al disco Galáctico, mientras que en los paneles lejanos se observan más estrellas pertenecientes al Halo Galáctico. Esto se observa de forma más evidente en los paneles en $|Z| < 4$ kpc y $R > 10$ kpc, en los cuales ambos peaks, pertenecientes a estrellas del disco y del halo, tienen una cantidad de estrellas similar. Este doble peak se sigue observando en $4 < |Z| < 6$ kpc, pero el peak en $\sim$ -1.4 dex es el dominante, es decir, hay mayor contribución de estrellas pobres en metales del halo. En $|Z| > 6$ kpc este doble peak no se observa, sino que la gran parte de las estrellas se encuentra distribuida en la región pobre en metales con peak en $\sim$ -1.4 dex, aunque para $R > 8$ se observa otro peak en $\sim$ -1.8 dex. En cuanto a las estrellas enanas, éstas se concentran en regiones más ricas en metales (a diferencia de la distribución más extensa de las estrellas gigantes), presentando un único peak en $\sim$ -0.5 dex en todos los paneles. Sólo se destacan los paneles cercanos al Sol ($|Z| < 4$ kpc y $6 < R < 10$ kpc) en donde la distribución no muestra el peak tan marcado como en los demás paneles. Cabe destacar que el hecho de que este peak en $\sim$ -0.5 dex en la Figura 4.18 es por la normalización de los gráficos. Esto es, los paneles cercanos al Sol son los que contienen la mayor parte de las estrellas enanas de la muestra, estas estrellas son más ricas en metales y corresponden a los peaks en la Figura 4.18. Finalmente, es necesario tener en cuenta que las incertidumbres en las distancias de las fuentes más distantes y débiles pueden llegar a ser considerables.

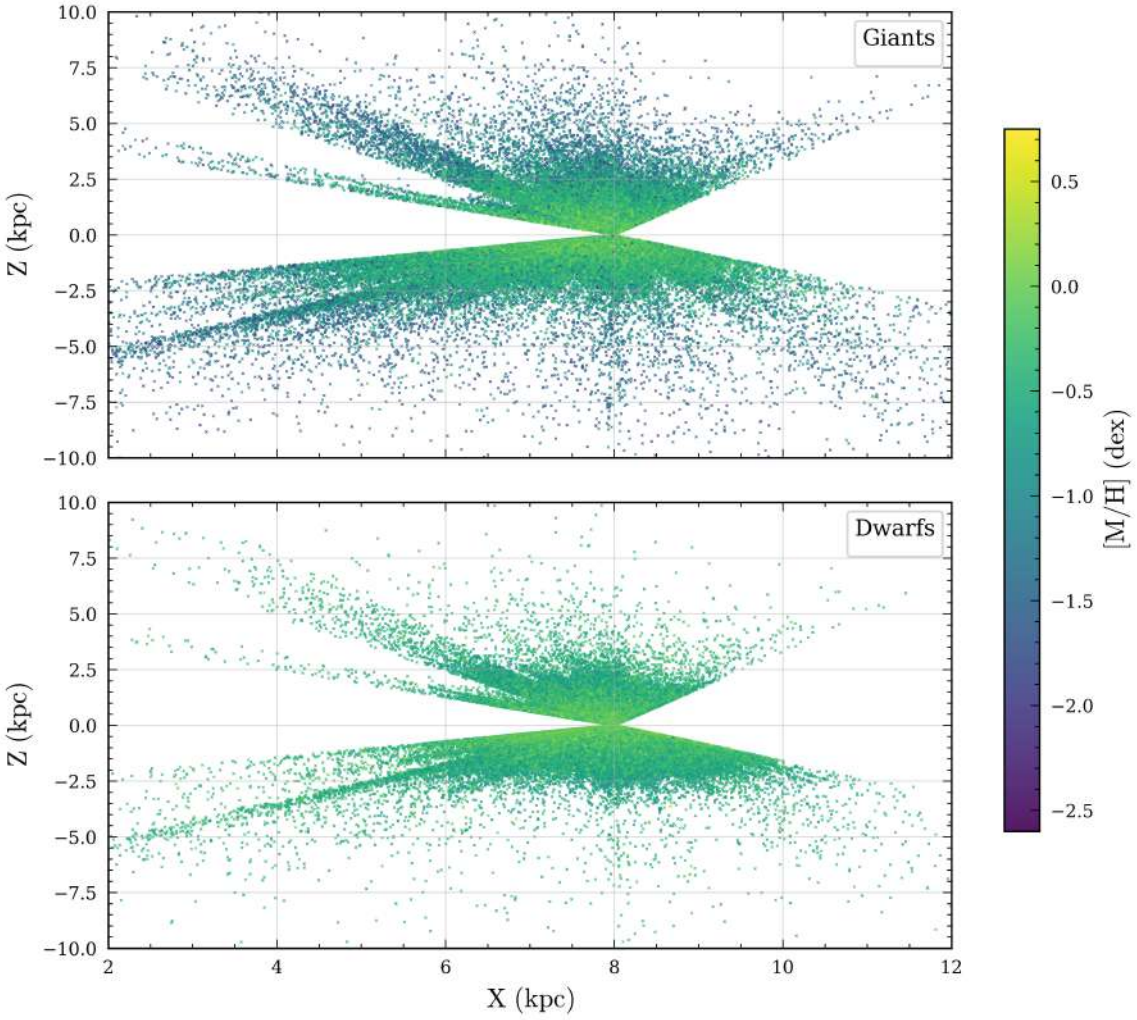


Figura 4.21: Distribución de metalicidad de estrellas gigantes (panel superior) y enanas (panel inferior) de S-PLUS en el plano (X,Z). El mapa de color indica las metalicidades. El Sol está ubicado en el punto (X,Z) = (8,0).

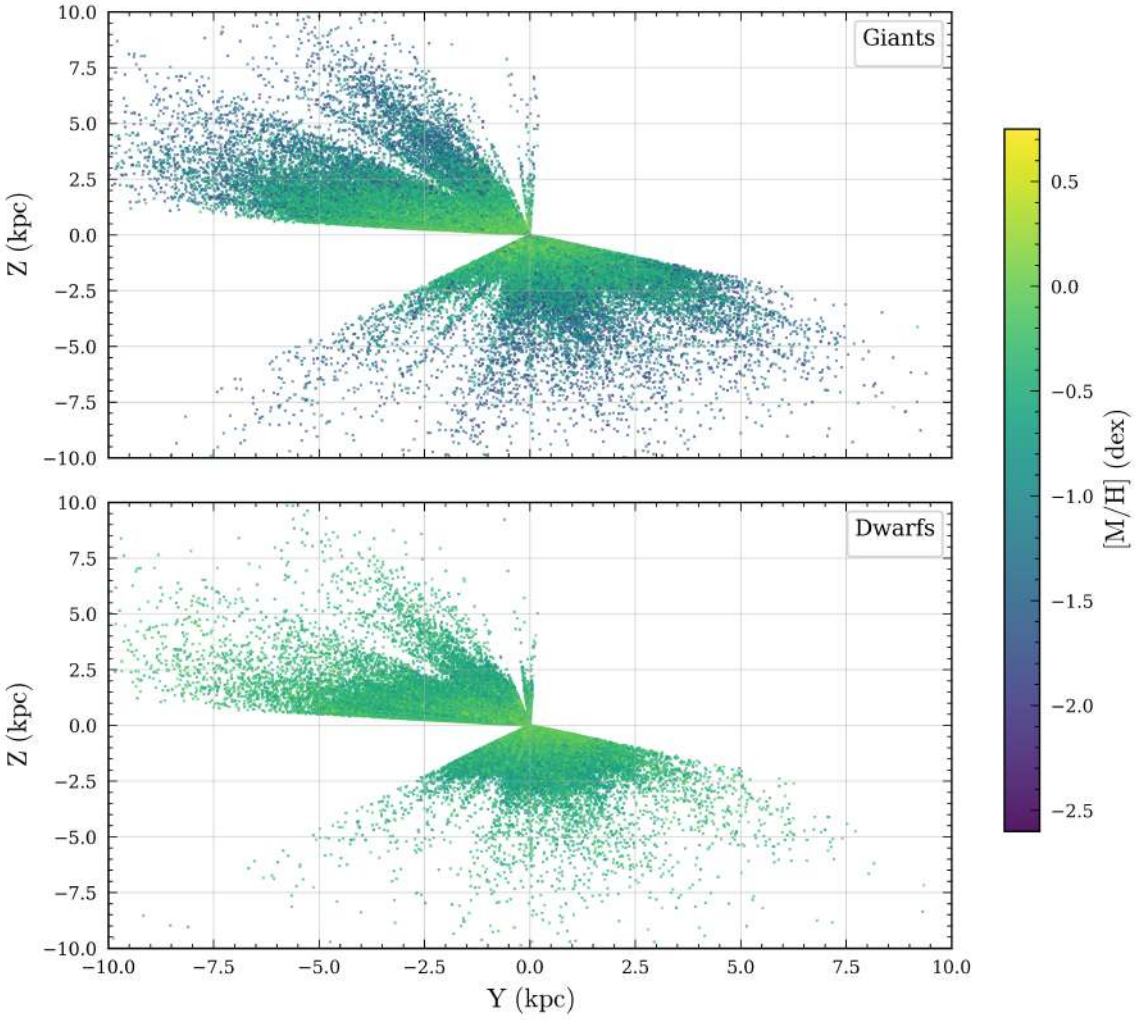


Figura 4.22: Distribución de metalicidad de estrellas gigantes (panel superior) y enanas (panel inferior) de S-PLUS en el plano (Y,Z). El mapa de color indica las metalicidades. El Sol está ubicado en el punto (Y,Z) = (0,0).

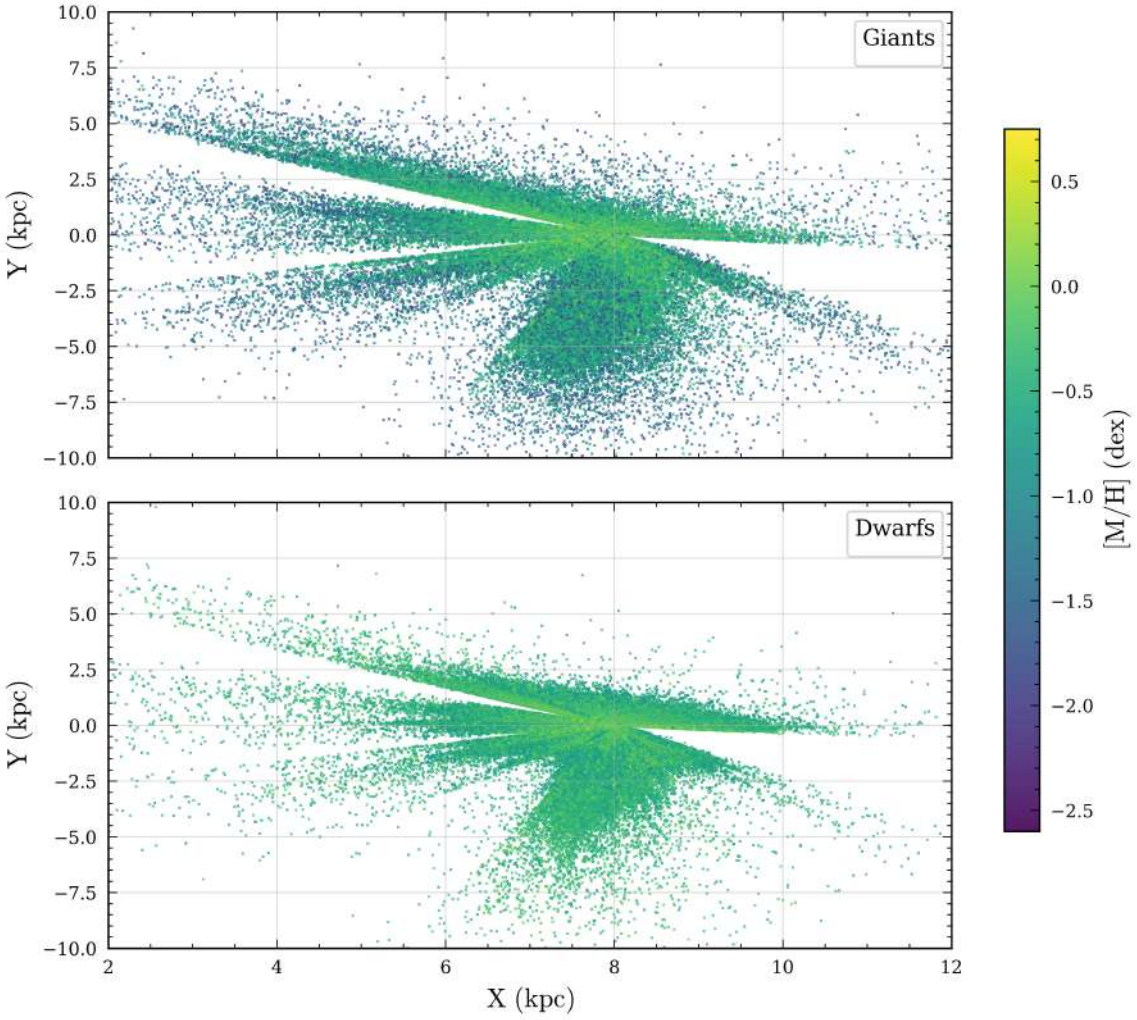


Figura 4.23: Distribución de metalicidad de estrellas gigantes (panel superior) y enanas (panel inferior) de S-PLUS en el plano (X,Y). El mapa de color indica las metalicidades. El Sol está ubicado en el punto (X,Y) = (0,8).

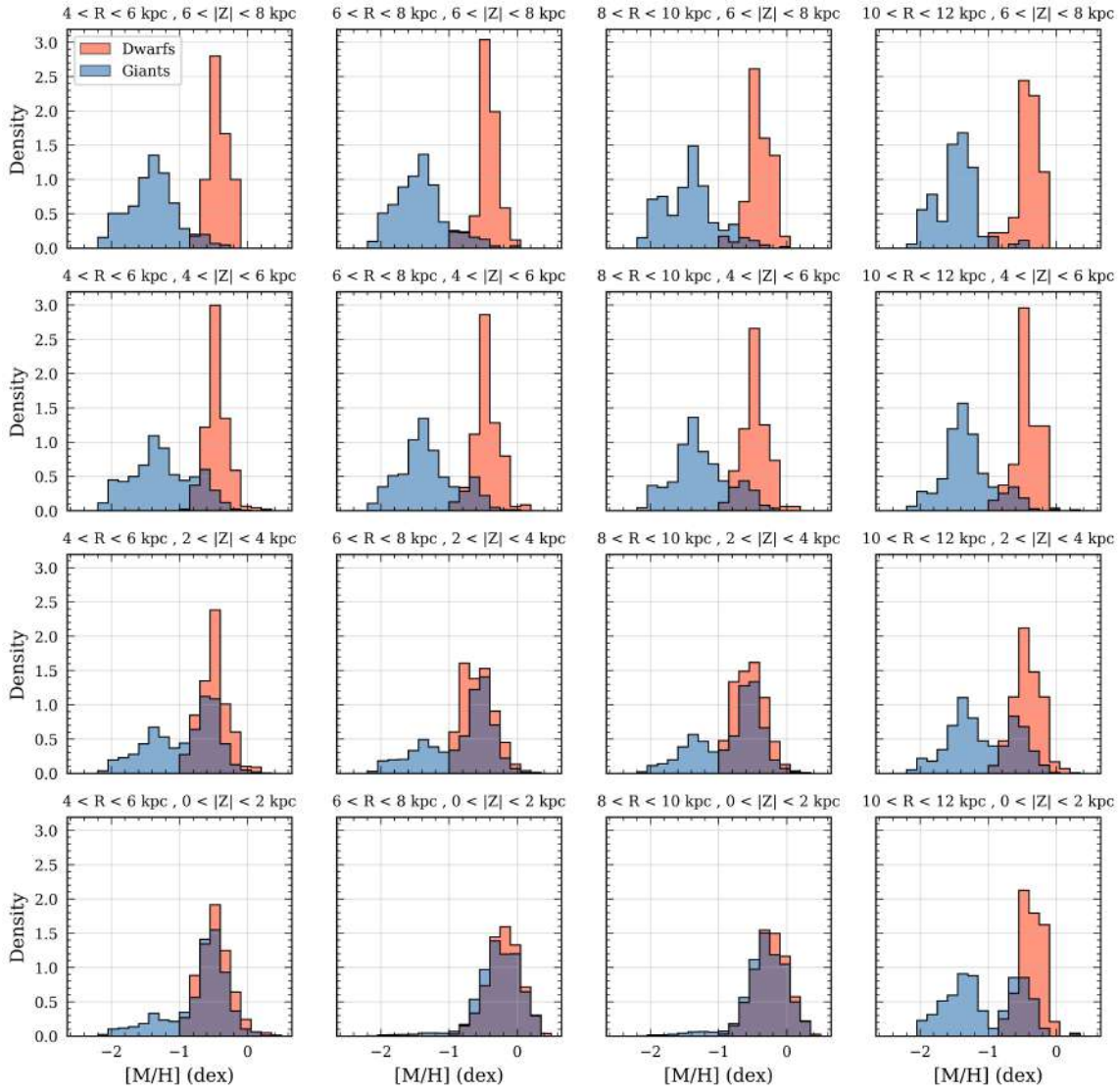


Figura 4.24: Densidad de probabilidad de metalicidades de las estrellas gigantes (azul) y enanas (rojo) de S-PLUS en distintos rangos de R y Z normalizada para el número de estrellas gigantes y enanas en cada panel.

Capítulo 5

Conclusiones

Como resultado de este trabajo de tesis se comprobó que la aplicación de métodos de *machine learning* al problema de derivación de metalicidades estelares puede alcanzar precisiones comparables con espectroscopia de resolución media pero a un costo de observación más bajo. Además, la clasificación de estrellas entre gigantes y enanas por medio de redes neuronales dio resultado satisfactorio.

La red neuronal de clasificación que presentó mejores resultados en la clasificación del set de pruebas cuenta con 4 capas ocultas y toma como características de entrada los colores fotométricos (F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - z), las primeras seis correspondientes a las características espectrales [OII], Ca H+K, Hδ, banda G, triplete de Mg y triplete de Ca, respectivamente. Esta red logró identificar correctamente el 97.1 % y el 97.9 % de las estrellas gigantes y enanas del set de pruebas, respectivamente. Entre las estrellas que este modelo clasificó como gigantes el 97.0 % estaban correctamente clasificadas y entre las estrellas que clasificó como enanas, el 98.0 % fueron correctamente clasificadas. Adicionalmente, se desarrolló un algoritmo de *random forest* cuyo rendimiento fue ligeramente menor, pero que dio resultados de todas formas consistentes con la red neuronal como clasificador.

En cuanto a las redes neuronales para las determinaciones de metalicidad se encontró que los algoritmos de regresión desarrollados de manera específica para cada tipo de estrella, es decir, específicamente para gigantes o enanas, produce estimaciones más precisas. La red que presentó menor error absoluto medio sobre el set de pruebas al momento de determinar las metalicidades fotométricas de las estrellas gigantes,

utiliza como características de entrada los colores (F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F660 - r), (F861 - z), (u - g), las primeras seis correspondientes a las características espectrales [OII], Ca H+K, Hδ, banda G, Hα, triplete de Mg y triplete de Ca, respectivamente. Esta red estimó las metalicidades de las estrellas del conjunto de pruebas con $\sigma_{giants} \sim 0.15$ dex. El algoritmo que presentó menor error absoluto medio en el set de pruebas para determinar las metalicidades de las estrellas enanas utiliza como características de entrada los colores (F378 - u), (F395 - g), (F410 - g), (F430 - g), (F515 - g), (F861 - z), (g - z), las primeras seis correspondientes a las características espectrales [OII], Ca H+K, Hδ, banda G, triplete de Mg y triplete de Ca, respectivamente. Esta red estimó las metalicidades de las estrellas del set de pruebas con $\sigma_{dwarf} \sim 0.13$ dex, con respecto a los valores espectroscópicos.

Posteriormente, se aplicaron las tres redes mencionadas en el párrafo anterior para clasificar y determinar las metalicidades de 812,378 estrellas del catálogo de S-PLUS iDR3 n4. Los algoritmos identificaron 130,172 estrellas gigantes con metalicidades entre -2.2 y 0.3 dex con un peak en $\sim$ -1.35 dex y un doble peak en $\sim$ -0.4 y $\sim$ -0.1 dex, y 682,206 estrellas enanas con metalicidades entre -0.9 y 0.4 dex, con un doble peak en $\sim$ -0.3 y $\sim$ -0.15 dex.

Con las distancias de Gaia EDR3, se construyeron gráficas de la distribución espacial de metalicidad de estas estrellas en la Galaxia. Através de esta inspección se encontró que las estrellas más ricas en metales se encuentran en la zona del disco Galáctico, y estas a su vez son las más numerosas. Sin embargo, al inspeccionar las distribuciones en distintos rangos en las coordenadas R y Z, se encontró que en las zonas más lejanas del Sol se empieza a notar la contribución de las estrellas gigantes del halo, representadas como un peak de estrellas pobres en metales pobre en metales. Dicha contribución incluso se nota a la altura del plano Galáctico pero más lejos del Sol, ya que en la vecindad solar la cantidad de estrellas ricas en metales es mucho mayor. Con respecto a las enanas, no se encontró mayor variación de metalicidades dentro de la Galaxia, puesto que la red sólo es capaz de determinar metalicidades en un rango entre -0.9 y 0.4 dex.

Capítulo 6

Trabajo a futuro

En este capítulo se presentan las acciones posteriores que se podrían realizar a partir de este trabajo.

En este trabajo se presentó un sistema de redes neuronales que da resultados bastante precisos para la obtención de metalicidades a partir de información netamente fotométrica. A futuro, aumentar el tamaño del set de datos de entrenamiento podría otorgar mejores resultados. Para esto, más adelante se espera contar con el set de datos completo de S-PLUS que tendrá un mayor *overlap* con APOGEE.

Para el desarrollo de arquitecturas de redes neuronales a partir de un ajuste de hiperparámetros que podrían ser más precisos se requiere infraestructura computacional mayor para poder abarcar el espacio de parámetros completo.

Cabe destacar que el objetivo de este trabajo está en el desarrollo del método para determinar metalicidades de estrellas gigantes rojas, es por esto que se presentó una inspección preliminar de los resultados hallados para los datos de S-PLUS. Un análisis más detallado de los mismos es parte de los trabajos futuros.

A partir de la información generada por la herramienta creada en este trabajo, es posible generar un catálogo de metalicidades de estrellas gigantes rojas de S-PLUS para poner a disposición de la comunidad por medio de un Value Added Catalog de S-PLUS.

Finalmente, como trabajo en progreso, producto de este trabajo se realizará una publicación Molina-Jorquera et al. (in prep).

Bibliografía

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

Abdurro'uf, Accetta, K., Aerts, C., et al. 2021, arXiv e-prints, arXiv:2112.02026. https://arxiv.org/abs/2112.02026

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, Astrophys. J. Suppl., 249, 3, doi: 10.3847/1538-4365/ab929e

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, Astrophys. J. Suppl., 219, 12, doi: 10.1088/0067-0049/219/1/12

Almeida-Fernandes, F. 2020, The S-PLUS Calibration Pipeline And schedule data releases, https://sites.usp.br/splus/wp-content/uploads/sites/846/2020/12/14_T_13_almeida-fernandes.pdf

Almeida-Fernandes, F., Sampedro, L., Herpich, F. R., et al. 2021, arXiv e-prints, arXiv:2104.00020. https://arxiv.org/abs/2104.00020

Angeloni, R., Gonçalves, D. R., Akras, S., et al. 2019, Astron. J., 157, 156, doi: 10.3847/1538-3881/ab0cf7

Armandroff, T. E., & Da Costa, G. S. 1991, Astron. J., 101, 1329, doi: 10.1086/115769

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, Astron. Astrph., 558, A33, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, Astron. J., 156, 123, doi: 10.3847/1538-3881/aabc4f

Baade, W. 1944, Astrophys. J., 100, 137, doi: 10.1086/144650

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, Astron. J., 161, 147, doi: 10.3847/1538-3881/abd806

Balas, V., Roy, S., Sharma, D., & Samui, P. 2019, Handbook of Deep Learning Applications, Smart Innovation, Systems and Technologies (Springer International Publishing). https://books.google.cl/books?id=Ih2KDwAAQBAJ

Baron, D. 2019, arXiv e-prints, arXiv:1904.07248. https://arxiv.org/abs/1904.07248

Bechtol, K., Drlica-Wagner, A., Balbinot, E., et al. 2015, Astrophys. J., 807, 50, doi: 10.1088/0004-637X/807/1/50

Belokurov, V., Deason, A. J., Erkal, D., et al. 2019, Mon. Not. R. Astron. Soc., 488, L47, doi: 10.1093/mnrasl/slz101

Belokurov, V., Erkal, D., Evans, N. W., Koposov, S. E., & Deason, A. J. 2018, Mon. Not. R. Astron. Soc., 478, 611, doi: 10.1093/mnras/sty982

Belokurov, V., Zucker, D. B., Evans, N. W., et al. 2006a, Astrophys. J. Let., 647, L111, doi: 10.1086/507324

Belokurov, V., Zucker, D. B., Evans, N. W., et al. 2006b, Astrophys. J. Let., 642, L137, doi: 10.1086/504797

Belokurov, V., Evans, N. W., Irwin, M. J., et al. 2007a, Astrophys. J., 658, 337, doi: 10.1086/511302

Belokurov, V., Zucker, D. B., Evans, N. W., et al. 2007b, Astrophys. J., 654, 897, doi: 10.1086/509718

Bertin, E., & Arnouts, S. 1996, Astron. Astrophys. Suppl. Ser., 117, 393, doi: 10.1051/aas:1996164

Bhattacharjee, J. 2017, Some Key Machine Learning Definitions, https://medium.com/technology-nineleaps/some-key-machine-learning-definitions-b524eb6cb48

Bilicki, M., Hoekstra, H., Brown, M. J. I., et al. 2018, Astronomy & Astrophysics, 616, A69, doi: 10.1051/0004-6361/201731942

Binney, J., Michael, M., & Merrifield, M. 1998, Galactic Astronomy, Princeton Series

in Astrophysics (Princeton University Press). https://books.google.cl/books?id=uDHNDwAAQBAJ

Breiman, L. 2001, Machine learning, 45, 5

Camarillo, T., Mathur, V., Mitchell, T., & Ratra, B. 2018, Pub. Astron. Soc. Pacific, 130, 024101, doi: 10.1088/1538-3873/aa9b26

Campante, T., Santos, N., & Monteiro, M. 2017, Asteroseismology and Exoplanets: Listening to the Stars and Searching for New Worlds: IVth Azores International Advanced School in Space Sciences, Astrophysics and Space Science Proceedings (Springer International Publishing). https://books.google.cl/books?id=keM8DwAAQBAJ

Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, Astrophys. J., 345, 245, doi: 10.1086/167900

Carroll, B., & Ostlie, D. 2017, An Introduction to Modern Astrophysics (Cambridge University Press). https://books.google.cl/books?id=PY0wDwAAQBAJ

Caruana, R., Karampatziakis, N., & Yessenalina, A. 2008, in Proceedings of the 25th International Conference on Machine Learning, ICML '08 (New York, NY, USA: Association for Computing Machinery), 96–103, doi: 10.1145/1390156.1390169

Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, Astron. Astrph., 622, A176, doi: 10.1051/0004-6361/201833036

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints, arXiv:1612.05560. https://arxiv.org/abs/1612.05560

Chollet, F., & Omernick, M. 2020, Working with preprocessing layers, https://keras.io/guides/preprocessing_layers/

Cole, A. A., Smecker-Hane, T. A., Tolstoy, E., Bosler, T. L., & Gallagher, J. S. 2004, Mon. Not. R. Astron. Soc., 347, 367, doi: 10.1111/j.1365-2966.2004.07223.x

Dark Energy Survey Collaboration, Abbott, T., Abdalla, F. B., et al. 2016, Mon. Not. R. Astron. Soc., 460, 1270, doi: 10.1093/mnras/stw641

Das, P., & Sanders, J. L. 2019, Mon. Not. R. Astron. Soc., 484, 294, doi: 10.1093/mnras/sty2776

Deason, A. J., Erkal, D., Belokurov, V., et al. 2021, Mon. Not. R. Astron. Soc., 501, 5964, doi: 10.1093/mnras/staa3984

Dettmers, T. 2015, Deep Learning in a Nutshell: Core Concepts, https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts/#layer

Draine, B. T. 2003, Annual Review of Astronomy & Astrophysics, 41, 241, doi: 10.1146/annurev.astro.41.011802.094840

Dwek, E., Arendt, R. G., Hauser, M. G., et al. 1995, Astrophys. J., 445, 716, doi: 10.1086/175734

Ferguson, A. M. N., Irwin, M. J., Ibata, R. A., Lewis, G. F., & Tanvir, N. R. 2002, Astron. J., 124, 1452, doi: 10.1086/342019

Fluri, J., Kacprzak, T., Lucchi, A., et al. 2019, Physical Review D, 100, doi: 10.1103/physrevd.100.063514

Fukugita, M., Ichikawa, T., Gunn, J. E., et al. 1996, Astron. J., 111, 1748, doi: 10.1086/117915

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, Astron. Astrph., 595, A1, doi: 10.1051/0004-6361/201629272

Gaia Collaboration, Babusiaux, C., van Leeuwen, F., et al. 2018, Astron. Astrph., 616, A10, doi: 10.1051/0004-6361/201832843

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, Astron. Astrph., 649, A1, doi: 10.1051/0004-6361/202039657

García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, Astron. J., 151, 144, doi: 10.3847/0004-6256/151/6/144

Geisler, D. 1986, Pub. Astron. Soc. Pacific, 98, 762, doi: 10.1086/131822

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning, Adaptive Computation and Machine Learning series (MIT Press). https://books.google.cl/books?id=Np9SDQAAQBAJ

Gramfort, A., Blondel, M., Grisel, O., et al. 2019, sklearn preprocessing MinMaxScaler, https://github.com/scikit-learn/scikit-learn/blob/7e1e6d09b/sklearn/preprocessing

Green, G. 2018, The Journal of Open Source Software, 3, 695, doi: 10.21105/joss.00695

Gruel, N., Moles, M., Varela, J., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8448, Observatory Operations: Strategies, Processes, and Systems IV, ed. A. B. Peck, R. L. Seaman, & F. Comeron, 84481V, doi: 10.1117/12.925581

Grus, J. 2019, Data Science from Scratch: First Principles with Python (O'Reilly Media). https://books.google.cl/books?id=YBKSDwAAQBAJ

Géron, A. 2017, Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (O'Reilly Media). https://books.google.cl/books?id=I6qkDAEACAAJ

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer series in statistics (Springer). https://books.google.cl/books?id=eBSgoAEACAAJ

Haykin, S., Haykin, S., & HAYKIN, S. 1999, Neural Networks: A Comprehensive Foundation, International edition (Prentice Hall). https://books.google.cl/books?id=bX4pAQAAMAAJ

Hettiarachchi, P., Hall, M. J., & Minns, A. W. 2005, Journal of Hydroinformatics, 7, 291, doi: 10.2166/hydro.2005.0025

Ibata, R., Irwin, M., Lewis, G., Ferguson, A. M. N., & Tanvir, N. 2001, Nature, 412, 49. https://arxiv.org/abs/astro-ph/0107090

Ibata, R., Malhan, K., Martin, N., et al. 2021, Astrophys. J., 914, 123, doi: 10.3847/1538-4357/abfcc2

Ibata, R. A., Gilmore, G., & Irwin, M. J. 1995, Mon. Not. R. Astron. Soc., 277, 781, doi: 10.1093/mnras/277.3.781

Ibata, R. A., Wyse, R. F. G., Gilmore, G., Irwin, M. J., & Suntzeff, N. B. 1997, Astron. J., 113, 634, doi: 10.1086/118283

Ibata, R. A., Lewis, G. F., McConnachie, A. W., et al. 2014, Astrophys. J., 780, 128, doi: 10.1088/0004-637X/780/2/128

IBM Cloud Education. 2020, Deep Learning, https://www.ibm.com/cloud/learn/deep-learning

Invernizzi, L., Long, J., Chollet, F., O'Malley, T., & Jin, H. 2019, Getting started with KerasTuner, https://keras.io/guides/keras_tuner/getting_started/

Ivezić, Ž., Sesar, B., Jurić, M., et al. 2008, Astrophys. J., 684, 287, doi: 10.1086/589678

Janesh, W., Morrison, H. L., Ma, Z., et al. 2016, Astrophys. J., 816, 80, doi: 10.3847/0004-637X/816/2/80

Jönsson, H., Holtzman, J. A., Allende Prieto, C., et al. 2020, Astron. J., 160, 120, doi: 10.3847/1538-3881/aba592

Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, Astrophys. J., 673, 864, doi: 10.1086/523619

Keras. 2019, Hyperband Tuner, https://keras.io/api/keras_tuner/tuners/hyperband/

Keras. 2020, Callbacks API, https://keras.io/api/callbacks/

Keras. 2021a, Softmax layer, https://keras.io/api/layers/activation_layers/softmax/

Keras. 2021b, Probabilistic losses, https://keras.io/api/losses/probabilistic_losses/

Keras. 2021c, Model plotting utilities, https://keras.io/api/utils/model_plotting_utils/#plotmodel-function

Kingma, D. P., & Ba, J. 2017, Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980

LeCun, Y., Bengio, Y., & Hinton, G. 2015, nature, 521, 436

Li, J., Xue, X.-X., Liu, C., et al. 2021, Astrophys. J., 910, 46, doi: 10.3847/1538-4357/abd9bf

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. 2018, Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. https://arxiv.org/abs/1603.06560

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, The Astrophysical Journal Supplement Series, 225, 31, doi: 10.3847/0067-0049/225/2/31

López-Sanjuan, C., Varela, J., Cristóbal-Hornillos, D., et al. 2019, Astron. Astrph., 631, A119, doi: 10.1051/0004-6361/201936405

López-Sanjuan, C., Yuan, H., Vázquez Ramió, H., et al. 2021, arXiv e-prints, arXiv:2101.12407. https://arxiv.org/abs/2101.12407

Louppe, G., Holt, B., Arnaud, J., & Hedayati, F. 2019, sklearn ensemble RandomForestClassifier, https://github.com/scikit-learn/scikit-learn/tree/7e1e6d09bcc2eaeba98f7e737aac2ac782f0e5f1/sklearn/ensemble

Mahabal, A., Sheth, K., Gieseke, F., et al. 2017, in 2017 IEEE Symposium Series on Computational Intelligence (SSCI) (IEEE), doi: 10.1109/ssci.2017.8280984

Majewski, S. R. 2004, 21, 197, doi: 10.1071/AS04031

Majewski, S. R., Nidever, D. L., Smith, V. V., et al. 2012, Astrophys. J. Let., 747, L37, doi: 10.1088/2041-8205/747/2/L37

Majewski, S. R., Ostheimer, J. C., Kunkel, W. E., & Patterson, R. J. 2000, Astron. J., 120, 2550, doi: 10.1086/316836

Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., & Ostheimer, J. C. 2003, Astrophys. J., 599, 1082, doi: 10.1086/379504

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, Astron. J., 154, 94, doi: 10.3847/1538-3881/aa784d

Marín-Franch, A., Chueca, S., Moles, M., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8450, Modern Technologies in Space- and Ground-based Telescopes and Instrumentation II, ed. R. Navarro, C. R. Cunningham, & E. Prieto, 84503S, doi: 10.1117/12.925430

Matteucci, F. 2001, The Chemical Evolution of the Galaxy, Astrophysics and Space Science Library (Springer Netherlands). https://books.google.cl/books?id=PT7O1nS7CksC

McConnachie, A. W., Irwin, M. J., Ibata, R. A., et al. 2009, Nature, 461, 66, doi: 10.1038/nature08327

Mcculloch, W., & Pitts, W. 1943, Bulletin of Mathematical Biophysics, 5, 127

McMahon, R. G., Banerji, M., Gonzalez, E., et al. 2013, The Messenger, 154, 35

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., et al. 2019, Mon. Not. R. Astron. Soc., 489, 241, doi: 10.1093/mnras/stz1985

Michel, V., Thirion, B., Gramfort, A., & Varoquaux, G. 2019, sklearn cluster AgglomerativeClustering, https://github.com/scikit-learn/scikit-learn/blob/7e1e6d09b/sklearn/cluster

Miller, A. 2015, A Photometric Machine-Learning Method to Infer Stellar Metallicity, doi: 10.1007/978-3-319-16313-0_17

Mitchell, T. 1997, Machine Learning, McGraw-Hill International Editions (McGraw-Hill). https://books.google.cl/books?id=EoYBngEACAAJ

Mohajon, J. 2020, Confusion Matrix for Your Multi-Class Machine Learning Model, https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

NASA/JPL-Caltech/R. Hurt (SSC/Caltech). 2017, A Roadmap to the Milky Way, https://solarsystem.nasa.gov/resources/285/the-milky-way-galaxy/

Newberg, H. J., Yanny, B., Rockosi, C., et al. 2002, Astrophys. J., 569, 245, doi: 10.1086/338983

Nidever, D. L., Majewski, S. R., & Butler Burton, W. 2008, Astrophys. J., 679, 432, doi: 10.1086/587042

Öhman, Y. 1934, Astrophys. J., 80, 171, doi: 10.1086/143595

O'Malley, T., Bursztein, E., Long, J., et al. 2019, KerasTuner, https://github.com/keras-team/keras-tuner

Ortolani, S., Renzini, A., Gilmozzi, R., et al. 1995, Nature, 377, 701, doi: 10.1038/377701a0

Oswalt, T., & Gilmore, G. 2013, Planets, Stars and Stellar Systems: Volume 5: Galactic Structure and Stellar Populations, Planets, Stars and Stellar Systems (Springer Netherlands). https://books.google.cl/books?id=lgwwkgEACAAJ

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Prialnik, D. 2000, An Introduction to the Theory of Stellar Structure and Evolution (Cambridge University Press). https://books.google.cl/books?id=TGyzlVbgkiMC

Radečić, D. 2020, Softmax Activation Function Explained, https://towardsdatascience.com/softmax-activation-function-explained-a7e1bc3ad60

Reid, M. J. 1993, Annual Review of Astronomy & Astrophysics, 31, 345, doi: 10.1146/annurev.aa.31.090193.002021

Reimers, C., Runge, J., & Denzler, J. 2020, Determining the Relevance of Features for Deep Neural Networks, 330–346, doi: 10.1007/978-3-030-58574-7_20

Rocha-Pinto, H. J., Majewski, S. R., Skrutskie, M. F., & Crane, J. D. 2003, Astrophys. J. Let., 594, L115, doi: 10.1086/378668

Rosenblatt, F. 1957, The perceptron - A perceiving and recognizing automaton, Tech. Rep. 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1985, Learning internal representations by error propagation, Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science

S-PLUS. 2019, S-PLUS: Instrumentation, https://www.splus.iag.usp.br/instrumentation/

Sammut, C., & Webb, G. I., eds. 2010, Adaptive System (Boston, MA: Springer US), 35–35, doi: 10.1007/978-0-387-30164-8_12

Samuel, A. L. 2000, IBM Journal of Research and Development, 44, 206, doi: 10.1147/rd.441.0206

Sanderson, R. E., Secunda, A., Johnston, K. V., & Bochanski, J. J. 2017, Mon. Not. R. Astron. Soc., 470, 5014, doi: 10.1093/mnras/stx1614

Sarang, P. 2020, Artificial Neural Networks with TensorFlow 2: ANN Architecture Machine Learning Projects (Apress). https://books.google.cl/books?id=4yOVzQEACAAJ

Schlafly, E. F., & Finkbeiner, D. P. 2011, Astrophys. J., 737, 103, doi: 10.1088/0004-637X/737/2/103

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, The Astrophysical Journal, 500, 525, doi: 10.1086/305772

Schneider, P. 2006, Extragalactic Astronomy and Cosmology: An Introduction (Springer). https://books.google.cl/books?id=uP1Hz-6sHaMC

Schultz, G. V., & Wiemer, W. 1975, Astron. Astrph., 43, 133

Shipp, N., Drlica-Wagner, A., Balbinot, E., et al. 2018, Astrophys. J., 862, 114, doi: 10.3847/1538-4357/aacdab

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, Astron. J., 131, 1163, doi: 10.1086/498708

Soumagnac, M. T., Abdalla, F. B., Lahav, O., et al. 2015, Mon. Not. R. Astron. Soc., 450, 666, doi: 10.1093/mnras/stu1410

Sparke, L., & Gallagher, J. 2007, Galaxies in the Universe: An Introduction (Cambridge University Press). https://books.google.cl/books?id=N8Hngab5liQC

Stanek, K. Z., Mateo, M., Udalski, A., et al. 1994, Astrophys. J. Let., 429, L73, doi: 10.1086/187416

Stringer, K. M., Drlica-Wagner, A., Macri, L., et al. 2021, Astrophys. J., 911, 109, doi: 10.3847/1538-4357/abe873

TensorFlow. 2021, tf.keras.Model, https://www.tensorflow.org/api_docs/python/tf/keras/Model#compile

TensorFlow. 2022, Introduction to the Keras Tuner, https://www.tensorflow.org/tutorials/keras/keras_tuner

Thackeray, A. D. 1939, Mon. Not. R. Astron. Soc., 99, 492, doi: 10.1093/mnras/99.6.492

The R Bootcamp. 2019, Machine Learning with R, https://therbootcamp.github.io/ML_2019May/

Thomas, G. F., Annau, N., McConnachie, A., et al. 2019, Astrophys. J., 886, 10, doi: 10.3847/1538-4357/ab4a77

Unavane, M., Wyse, R. F. G., & Gilmore, G. 1996, Mon. Not. R. Astron. Soc., 278, 727, doi: 10.1093/mnras/278.3.727

Villard, R., Christensen, L. L., Noyola, E., & Gebhardt, K. 2008, Astronomers Find Suspected Medium-Size Black Hole in Omega Centauri, https://www.nasa.gov/mission_pages/hubble/science/hst_img_20080402.html

Walmsley, M., Smith, L., Lintott, C., et al. 2020, Mon. Not. R. Astron. Soc., 491, 1554, doi: 10.1093/mnras/stz2816

Wang, B., Hu, S. J., Sun, L., & Freiheit, T. 2020, Journal of Manufacturing Systems, 56, 373, doi: https://doi.org/10.1016/j.jmsy.2020.06.020

Wannier, P., & Wrixon, G. T. 1972, Astrophys. J. Let., 173, L119, doi: 10.1086/180930

Warren, S. R., & Cole, A. A. 2009, Mon. Not. R. Astron. Soc., 393, 272, doi: 10.1111/j.1365-2966.2008.14268.x

Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2019a, Astron. Astrph., 622, A182, doi: 10.1051/0004-6361/201833368

Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2019b, Astron. Astrph., 622, A182, doi: 10.1051/0004-6361/201833368

Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2021, Astrophys. J., 912, 147, doi: 10.3847/1538-4357/abee7e

Xu, Y., & Goodacre, R. 2018, Journal of Analysis and Testing, 2, doi: 10.1007/s41664-018-0068-2

Yang, C., Xue, X.-X., Li, J., et al. 2019, Astrophys. J., 880, 65, doi: 10.3847/1538-4357/ab2462

Yanny, B., Newberg, H. J., Kent, S., et al. 2000, Astrophys. J., 540, 825, doi: 10.1086/309386

Yip, K. H., Nikolaou, N., Coronica, P., et al. 2019, in AAS/Division for Extreme Solar Systems Abstracts, Vol. 51, AAS/Division for Extreme Solar Systems Abstracts, 305.04

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, Astron. J., 120, 1579, doi: 10.1086/301513

Yuan, H. B., Liu, X. W., & Xiang, M. S. 2013, Mon. Not. R. Astron. Soc., 430, 2188, doi: 10.1093/mnras/stt039

Zeilik, M., & Gregory, S. 1998, Introductory Astronomy & Astrophysics, Saunders golden sunburst series (Saunders College Pub.). https://books.google.cl/books?id=iH7vAAAAMAAJ

Zell, A. 1997, Simulation neuronaler Netze (Oldenbourg). https://books.google.cl/books?id=bACTSgAACAAJ

Zoccali, M. 2005, in Astrophysics and Space Science Library, Vol. 327, The Initial Mass Function 50 Years Later, ed. E. Corbelli, F. Palla, & H. Zinnecker, 95, doi: 10.1007/978-1-4020-3407-7_14