# Identificación de candidatas a estrellas Be utilizando redes neuronales

Y. Aidelman[1,2], C. Escudero[2], F. Ronchetti[3,4], F. Quiroga[3], A. Granada[5,6] & L. Lanzarini[3]

[1] *Departamento de Espectroscopía, Facultad de Ciencias Astronómicas y Geofísicas, UNLP, Argentina*
[2] *Instituto de Astrofísica de La Plata, CONICET–UNLP, Argentina*
[3] *Instituto de Investigación en Informática LIDI, CIC–UNLP, Argentina*
[4] *Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina*
[5] *Universidad Nacional de Río Negro, Argentina*
[6] *Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina*

*Contacto* / aidelman@fcaglp.unlp.edu.ar

**Resumen** / Las bases de datos astronómicas proporcionan actualmente grandes volúmenes de información espectroscópica y fotométrica. En particular, los datos fotométricos resultan relativamente más fáciles de obtener debido al menor tiempo de uso del telescopio, con lo cual existe una creciente necesidad de utilizarlos para identificar automáticamente objetos específicos y luego estudiarlos en detalle. En este trabajo, nos centramos en la identificación fotométrica de estrellas Be, objetos tempranos que presentan la línea H$\alpha$ en emisión. Este tipo de objeto es de interés para el entendimiento de la evolución de estrellas en alta rotación, y también para el estudio de la física de discos circunestelares. Para su identificación, utilizamos datos fotométricos (VPHAS+, 2MASS y AllWISE) y espectroscópicos (LAMOST), junto con técnicas de aprendizaje automático, como las redes neuronales. Nuestros resultados muestran que utilizar los índices $Q$ libres de enrojecimiento como descriptores, proporcionan una mejora significativa en la identificación fotométrica de estrellas Be.

**Abstract** / Astronomical databases currently provide large volumes of spectroscopic and photometric information. In particular, as photometric data is relatively easier to obtain due to the shorter use time of the telescope, there is an increasing need to use those data in order to automatically identify specific objects and study them in detail afterwards. In this work, we focus on the photometric identification of Be stars, early-type stars with H$\alpha$ line in emission. These kind of objects are very interest for understanding the evolution of fast rotating stars, and also for the study of the physics of circumstellar disks. For their identification, we use photometric (VPHAS+, 2MASS, AlWISE) and spectroscopic (LAMOST) databases, together with machine learning techniques, such as neural networks. Our results show that using the reddening-free $Q$ indices as features provides a significant improvement in the photometric identification of Be stars.

*Keywords* / methods: data analysis — stars: emission-line, Be — surveys

## 1. Introducción

En el último tiempo, se han incrementado apreciablemente las bases de datos y los relevamientos del cielo. La gran cantidad de datos disponibles hace necesario implementar nuevas estrategias para poder analizarlos y estudiarlos. Por ello, las técnicas de aprendizaje automático resultan una herramienta fundamental para la búsqueda de objetos particulares, y así comprender de manera más acabada sus características y los procesos astrofísicos que están en juego. En este contexto, el análisis desarrollado en este trabajo, el cual pretende identificar estrellas Be mediante el uso de redes neuronales entrenadas a partir de datos fotométricos, resulta una continuación del trabajo Aidelman et al. (2020, Aidelman20 en adelante).

Una estrella Be es una estrella de tipo espectral B no supergigante, que generalmente se encuentra rodeada por una envoltura circunestelar en forma de disco y en rotación kepleriana (Porter & Rivinius, 2003; Rivinius et al., 2013). Estos objetos tienen la particularidad de presentar alta rotación, lo que los convierte en laboratorios únicos para la evaluación de nuestro conocimiento de la evolución de estrellas con esta característica.

La presencia del disco se manifiesta en el espectro, principalmente por la emisión de la línea H$\alpha$ (Jaschek et al., 1981; Collins, 1987) como se muestra en panel medio de la Fig. 1. Esta característica permite distinguirlas fotométricamente de las estrellas B normales utilizando filtros centrados en esta línea particular.

En Aidelman20, se mostró que los mejores resultados en la identificación de estrellas Be se obtuvieron utilizando una red neuronal entrenada con los índices $Q$ libres de enrojecimiento (Johnson & Morgan, 1955). Estos índices se contruyeron utilizando magnitudes VPHAS+, IPHAS, SDSS y 2MASS (Drew et al., 2014; Barentsen et al., 2014; Alam et al., 2015; Skrutskie et al., 2006, respectivamente). Continuando en esta línea de trabajo, el análisis presentado aquí, tiene por objetivo analizar y mostrar cuáles serían los descriptores* más adecuados para entrenar la red neuronal. Además, mos-

*Utilizamos la palabra "descriptor" como traducción de *feature*.

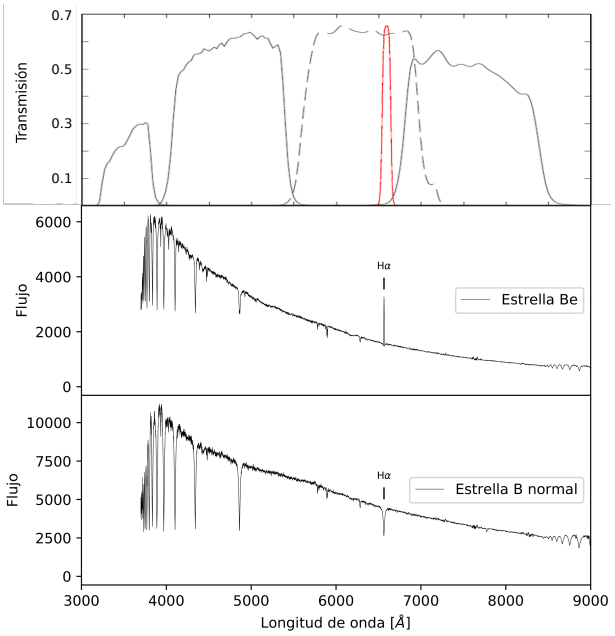


Figura 1: *Panel superior:* Perfiles de transmisión de los filtros $u, g, r, i$ utilizados por VPHAS+. El filtro de banda angosta centrado en H$\alpha$ se muestra en rojo. *Panel central:* Espectro típico de una estrella Be clásica donde se indica la línea H$\alpha$ en emisión. *Panel inferior:* Espectro típico de una estrella de tipo espectral B.

tramos cuál es el impacto de incorporar las magnitudes $W1$ y $W2$ (AllWISE, Cutri et al., 2013) en la capacidad del modelo para recuperar la mayor cantidad de estrellas Be.

## 2. Metodología

Las técnicas de aprendizaje automático supervisado consisten en deducir una función (modelo) a partir de un conjunto de datos "etiquetados", para luego utilizarlo sobre otro conjunto de datos sin etiquetar. Como se menciona en la sección § 2.1 de Aidelman20, utilizamos los datos fotométricos publicados por Mohr-Smith et al. (2017, Mohr17 en adelante) para entrenar y evaluar el modelo, el cual posteriormente se aplicó a los datos espectroscópicos de Liu et al. (2019, Liu19 en adelante). Para este trabajo se logró duplicar la muestra etiquetada utilizada en Aidelman20, obteniendo un total de 460 estrellas (Tab. 1).

De acuerdo a estos conjuntos de datos, los descriptores "originales" son las magnitudes aparentes. Sin embargo, como sabemos, estas se encuentran afectadas por la distancia y por el enrojecimiento interestelar que, a priori, no se conocen. Por otra parte sabemos que la manera usual de identificar estrellas con emisión en H$\alpha$ a partir de datos fotométricos, es a través de los diagramas color-color. En la Fig. 2 se observa cómo la mayoría de las estrellas brillantes en H$\alpha$ (cuadrados azules) se ubican por encima de la secuencia estelar de estrellas normales (círculos verdes). Si bien, este tipo de diagramas es independiente de la distancia, aún se encuentra afectado por el enrojecimiento. Por esto es necesario encontrar un descriptor más adecuado para entrenar un modelo que identifique estrellas Be en cualquier región del cielo.

Tabla 1: Distribución por clases de las muestras utilizadas en este trabajo. OB: estrellas de tipo espectral O y B normales. EM/Be: estrellas que presentan emisión en H$\alpha$.

| Datos | OB | EM/Be | Total |
|---|---|---|---|
| Mohr-Smith et al. (2017) | 5629 | 248 | 5877 |
| Liu et al. (2019) | 359 | 101 | 460 |

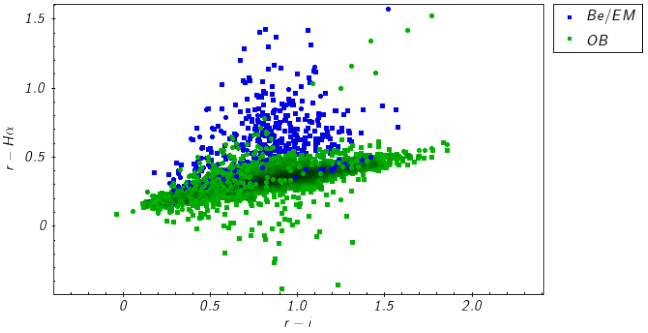


Figura 2: Diagrama color-color. Con símbolos cuadrados se indican los datos de Mohr-Smith et al. (2017) y con círculos los de Liu et al. (2019). En azul se indican las estrellas de ambas muestras etiquetadas como Be o EM, mientras que en verde se indican las OB normales.

A fin de minimizar los efectos antes mencionados, utilizamos lo que en aprendizaje automático se denomina "transformación de descriptores". En este caso, el entrenamiento del modelo se llevó a cabo empleando los parámetros $Q$ libres de enrojecimiento e independientes de la distancia. Definidos por Johnson & Morgan (1955), estos parámetros se pueden construir a partir de las magnitudes aparentes:

$$Q_{1234} = (m_{\lambda_1} - m_{\lambda_2}) - \frac{r_{\lambda_1} - r_{\lambda_2}}{r_{\lambda_3} - r_{\lambda_4}} (m_{\lambda_3} - m_{\lambda_4})$$
$$= (m^0_{\lambda_1} - m^0_{\lambda_2}) - \frac{r_{\lambda_1} - r_{\lambda_2}}{r_{\lambda_3} - r_{\lambda_4}} (m^0_{\lambda_3} - m^0_{\lambda_4}).$$

siendo los términos $m_{\lambda_i}$ y $m^0_{\lambda_i}$ las magnitudes aparente e intrínseca en 4 filtros diferentes centrados en la longitud de onda $\lambda_i$. Por su parte, el término $r_{\lambda_i}$ se define como $r_{\lambda_i} = A_{m\lambda_i}/A_V$, siendo $A_{m\lambda_i}$ los coeficientes de extinción en la longitud de onda respectiva y $A_V$ el coeficiente de extinción para el filtro $V$ (Johnson). Los valores $r_{\lambda_i}$ adoptados en este trabajo fueron calculados por Yuan et al. (2013) para $R_V = 3.1$.

## 3. Resultados y discusión

En primera instancia, como ahora contamos con más datos de Liu19 etiquetados, repetimos el experimento realizado en Aidelman20. Para ello, entrenamos la red neuronal con los $Q_{123}$ (esto es tomando $m_{\lambda_2} = m_{\lambda_3}$) usando las magnitudes $u, g, r$, H$\alpha$, $i, J, H, K$, y calibramos la salida para obtener una pureza** determinada en

**Utilizamos "pureza" y "completitud" como traducción de *precision* y *recall*, respectivamente.

los resultados de entrenamiento. Esto nos permitió obtener una completitud del 28 % para la nueva muestra, manteniendo una pureza muy alta del 100 %.

Posteriormente, entrenamos la red utilizando distintos descriptores con el fin de evaluar cuál de ellos arroja una mayor completitud sobre la muestra de Liu19. En todas las pruebas también se evaluó el modelo sobre una submuestra de Mohr17, obteniendo valores de completitud similares en todos los casos (80 %).

En los siguientes experimentos se utilizaron las magnitudes $u, g, r$, H$\alpha, i, J, H, K, W1$ y $W2$.

- $(i)$ Al entrenar la red con las magnitudes aparentes se obtiene una completitud del 27 %.
- $(ii)$ Entrenando la red con los índices de color $(u-g), (u-r), (u-i), (g-r), (g-i), (r-$H$\alpha), (r-i), (J-H), (J-K), (H-K)$ y $(W1-W2)$, se obtiene una completitud del 10 %.
- $(iii)$ Entrenando con los índices $Q_{123}$, obtenemos una completitud del 32 %.
- $(iv)$ Entrenando con los índices $Q_{1234}$ obtenemos una completitud del 36 %.
- $(v)$ Finalmente, las últimas pruebas consistieron en eliminar magnitudes para reducir la cantidad de descriptores. De todas las pruebas realizadas, el mejor resultado se logró entrenando una red neuronal con los índices $Q_{1234}$ utilizando H$\alpha$, $J$, $H$, $K$, $W1$, $W2$. En este caso, se obtuvo una completitud del 33 %.

Al comparar el resultado del experimento (i) con el obtenido por Aidelman20 (completitud del 9.5 %) se observa que las magnitudes $W1$ y $W2$ aportan información valiosa para separar las estrellas B normales de las Be (en acuerdo con los resultados obtenidos por Granada et al., 2018). Por su parte, el experimento (ii) indica que el enrojecimiento interestelar afecta apreciablemente la capacidad del modelo para identificar estrellas Be en diferentes regiones del cielo (resultado esperable ya que la extinción interestelar no es homogénea).

Con respecto a los experimentos (iii), (iv) y (v), los resultados indican que los descriptores más adecuados para entrenar el modelo resultan ser los parámetros $Q$, ya que son independientes de la distancia y están prácticamente libres del enrojecimiento. Además, se obtienen los mejores resultados para la completitud de la muestra de Liu19. Adicionalmente, la ventaja de utilizar los $Q$ es que esta transformación de descriptores resulta más sencilla que la corrección por distancia y absorción interestelar propuesta en otros trabajos similares de la literatura (por ejemplo, Vioque et al., 2020).

## 4. Conclusiones

El objetivo principal de este trabajo se centró en analizar los distintos descriptores a fin de obtener el más adecuado para identificar estrellas Be de una muestra etiquetada. Para ello, es importante minimizar el efecto del enrojecimiento interestelar sobre el descriptor para poder aplicar el método a cualquier región del cielo.

De acuerdo a nuestros resultados, utilizar el parámetro $Q$ como descriptor mejora apreciablemente el valor de la completitud, en particular los construidos con 4 magnitudes distintas, alcanzando un 36 % de completitud en los datos de Liu19 con una pureza del $\sim 100$ %. Esto significa una mejora del 11 % respecto a nuestro trabajo anterior (Aidelman20).

Desde el punto de vista astronómico, el uso de los índices $Q$ resulta novedoso dado que es un parámetro muy poco utilizado en la literatura. Además, cabe señalar que los coeficientes $r_{\lambda_i}$ sólo dependen del filtro fotométrico utilizado, lo que facilita aún más el uso de los $Q$, ya que no es necesario saber a priori la absorción interestelar (ni la distancia) que afecta a cada estrella. Es interesante mencionar que desde el punto de vista del aprendizaje automático, el uso de los $Q$ para entrenar al modelo también es novedoso ya que es una transformación poco usual.

Otra particularidad a resaltar en este trabajo, es el hecho de haber fijado el valor de pureza cercano al 100 % lo que permite disminuir los casos falsos positivos que brinda el modelo. Esto significa que nuestra red neuronal encuentra el 36 % de las estrellas Be de la muestra de Liu19 sin equivocarse.

Finalmente, el uso de las magnitudes infrarrojas junto con la magnitud H$\alpha$, al ser trazadoras de la presencia de disco, resultan indispensables a fin de separar las estrellas B de las Be.

*Agradecimientos:* Agradecemos al referí por sus sugerencias y comentarios. Y.A. agradece el apoyo de la UNLP (I+D 11g162) y a la Agencia I+D+I (PICT-2016-1971). A.G. agradece el apoyo de la Agencia I+D+i (PICT2017-3790). Este proyecto ha recibido financiación dentro del marco del Programa de Investigación e Innovación Horizonte 2020 (2014-2020) de la Unión Europea en virtud del Acuerdo de subvención Marie Skłodowska-Curie No. 823734.

## Referencias

Aidelman Y., et al., 2020, E. Rucci, M. Naiouf, F. Chichizola, L. De Giusti (Eds.), *Cloud Computing, Big Data & Emerging Topics*, 111–123, Springer International Publishing, Cham
Alam S., et al., 2015, ApJS, 219, 12
Barentsen G., et al., 2014, MNRAS, 444, 3230
Collins II G.W., 1987, A. Slettebak, T.P. Snow (Eds.), *IAU Colloq. 92: Physics of Be Stars*, 3–19
Cutri R.M., et al., 2013, Explanatory Supplement to the AllWISE Data Release Products, Explanatory Supplement to the AllWISE Data Release Products
Drew J.E., et al., 2014, MNRAS, 440, 2036
Granada A., et al., 2018, AJ, 155, 50
Jaschek M., Slettebak A., Jaschek C., 1981, Be star terminology., Be Star Newsletter
Johnson H.L., Morgan W.W., 1955, ApJ, 122, 142
Liu Z., et al., 2019, ApJS, 241, 32
Mohr-Smith M., et al., 2017, MNRAS, 465, 1807
Porter J.M., Rivinius T., 2003, PASP, 115, 1153
Rivinius T., Carciofi A.C., Martayan C., 2013, A&A Rv, 21, 69
Skrutskie M.F., et al., 2006, AJ, 131, 1163
Vioque M., et al., 2020, A&A, 638, A21
Yuan H.B., Liu X.W., Xiang M.S., 2013, MNRAS, 430, 2188