Universidad Nacional de La Plata
Facultad de Ciencias Astronómicas y Geofísicas

Tesis para obtener el grado académico de
Licenciado en Astronomía

Algoritmos de aprendizaje automático para el estudio de efectos sistemáticos en las velocidades radiales obtenidas para la búsqueda de planetas extrasolares

Juan R. Serrano Bell

Director: Dr. Rodrigo F. Díaz
Co-Director: Dr. Octavio M. Guilera

La Plata, Argentina
- Marzo de 2022 -

Prefacio

Esta Tesis es presentada como parte de los requisitos para obtener el grado académico de Licenciado en Astronomía de la Universidad Nacional de La Plata. La misma contiene los resultados de las investigaciones desarrolladas bajo la dirección del Dr. Rodrigo F. Díaz y la codirección del Dr. Octavio M. Guilera.

Juan R. Serrano Bell.
e-mail: serranojuan@fcaglp.unlp.edu.ar
La Plata, Marzo de 2023.

Resumen

SOPHIE es un espectrógrafo echelle ubicado en el Observatorio de Haute-Provence, Francia. Mediante calibración simultánea de la longitud de onda puede alcanzar precisiones de velocidad radial (RV) cercanas a 1 $m\ s^{-1}$. Sin embargo, el punto cero del instrumento presenta derivas a baja frecuencia de algunos $m\ s^{-1}$ que deben ser corregidas para alcanzar la alta precisión que requieren los programas actuales de búsqueda de exoplanetas. Con este fin se monitorean regularmente estrellas estándar de velocidad radial constante, y se usan estas mediciones para corregir las velocidades observadas por la variación del punto cero. En este trabajo, nos proponemos lograr una nueva forma de realizar la corrección de punto cero de instrumentos como SOPHIE. Usamos técnicas de aprendizaje automático supervisado para predecir los cambios de punto cero a partir de numerosas variables ambientales, instrumentales, y observacionales. Una vez entrenados, los algoritmos de aprendizaje automático tienen el potencial para permitirnos realizar la corrección sin necesidad de observar estrellas estándar de RV y de obtener conocimiento sobre el instrumento que permita mejorar su estabilidad y precisión. Exploramos distintos algoritmos hasta dar con un buen modelo y luego realizamos predicciones de la deriva del punto cero para un grupo de estrellas. Finalmente comparamos la corrección obtenida por nuestro modelo con la corrección realizada con observación de estrellas estándar.

Abstract

SOPHIE is an echelle spectrograph located at the Haute-Provence Observatory, France. By simultaneous wavelength calibration, it can reach radial velocity (RV) accuracies close to 1 $m\ s^{-1}$. However, the zero point of the instrument exhibits a low frequency drift of a few $m\ s^{-1}$ which must be corrected to achieve the high precision required by current exoplanet search programs. To this end, standard stars of constant radial velocity are regularly monitored, and these measurements are used to correct the observed velocities for zero-point variation. In this work, we aim to achieve a new way to perform zero-point correction of instruments such as SOPHIE. We use supervised machine learning techniques to predict the zero-point drift from numerous environmental, instrumental, and observational variables. Once trained, the machine learning algorithms have the potential to allow us to perform the correction without the need to observe standard VR stars and to gain knowledge about the instrument to improve its stability and accuracy. We explored different algorithms until we found a good model and then made predictions of the zero-point drift for a group of stars. Finally, we compared the correction obtained by our model with the correction obtained by observing standard stars.

Agradecimientos

A mis padres. Por ayudarme siempre, por el apoyo incondicional, por ser las mejores personas que conozco.

A Rodrigo. Por la confianza, la dedicación y los infinitos consejos.

A mis hermanos. Sil, Pato, Mari, Miguel y Valen. Por estar siempre unidos, son mi constante.

A mis amigos, los del barrio y los que hice en estos años: Agus, Kala, Lean, Mati, Lucho, Juli, Chaco, Nati. Gracias por hacer este camino más fácil, los quiero. A Azu, por aparecer en los meses más desafiantes de mi vida y hacer todo más lindo.

Al Observatorio de La Plata, mi lugar favorito en el mundo. Y a toda su comunidad. A la gente de Planetario y Museo, lugares en los que me siento privilegiado de haber trabajado y donde he aprendido tanto, me llevo montones de hermosas experiencias.

A la Universidad Pública Argentina.

Índice general

Acrónimos

Lista de acrónimos utilizados en esta tesis (notar que las siglas usualmente corresponden a las utilizadas en el idioma inglés):

- RV: Velocidad radial (*radial velocity*)
- SNR: Relación Señal a Ruido (*signal-to-noise ratio*)
- IP: Perfil instrumental (*instrumental profile*)
- CCF: Función de correlación Cruzada (*cross-correlation function*)
- RMS: Media cuadrática (*root mean square*)
- GBMs: Máquinas de gradient boosting (*gradient boosting machines*)
- GB: Incremento por gradiente (*gradient boosting*)
- GBDT: Árboles de decisión incrementados por gradiente (*gradient boosted decission tree*)
- RMSE: Raíz del error cuadrático medio (*root mean squared error*)
- WRMSE: Raíz del error cuadrático medio pesado (*weighted root mean squared error*)
- BJD: Día juliano baricéntrico (*baricentric julian day*)
- CM: Centro de masa.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

> *Tras cada hombre viviente se encuentran treinta fantasmas, pues tal es la proporción numérica con que los muertos superan a los vivos. Desde el alba de los tiempos, aproximadamente cien mil millones de seres humanos han transitado por el planeta Tierra.*
>
> *Y es en verdad un número interesante, pues por curiosa coincidencia hay aproximadamente cien mill millones de estrellas en nuestro universo local, la Vía Láctea. Así, por cada hombre que jamás ha vivido, luce una estrella en ese Universo.*
>
> *Pero, cada una de esas estrellas es un sol, a menudo mucho más brillante y magnífico que la pequeña y cercana a la que denominamos el Sol. Y muchos -quizá la mayoría- de esos soles lejanos tienen planetas circundándolos. Así, casi con seguridad hay suelo suficiente en el firmamento para ofrecer a cada miembro de las especies humanas, desde el primer hombre-mono, su propio mundo particular: cielo... o infierno.*
>
> *—2001 Una Odisea Espacial. Arthur C. Clarke.*

Estas líneas fueron publicadas en 1968. Hoy, más de 50 años después podemos asegurar que por cada ser humano que alguna vez vivió además de una estrella en nuestra galaxia existe -como mínimo- una galaxia en el Universo observable. La idea de un Universo con incontables mundos que viene ya desde la antigua Grecia, no pertenece más a la especulación sino que es una realidad, y somos la primera generación que vive sabiendo esto con total certeza. Hoy conocemos más de cinco mil otros mundos fuera de nuestro sistema solar, y la misión Kepler nos reveló que existen más planetas que estrellas. La pregunta que queda por responder todavía, sin embargo, es qué tan frecuente es la vida en estos mundos. ¿Podremos alguna vez responderla? No sabemos. Pero hay razones para ser optimistas, siendo no más que una especie de primates que hace apenas trescientos mil años (un 0.006 % de la edad de la Tierra!) empezó a merodear un pequeño planeta en un confín de una galaxia entre miles de millones, lo que hemos logrado expandir la frontera de nuestro conocimiento sobre la naturaleza es muchísimo. Las preguntas más difíciles en la Ciencia requieren esfuerzos prolongados con pequeños y sucesivos avances técnicos y teóricos. La curiosidad es el gran

motor del conocimiento, y estoy seguro de que mientras existamos como especie intentaremos zanjar este interrogante. Esta tesis tiene que ver con aportar un granito de arena en esa dirección.

1.1. Contexto

El primer descubrimiento de un mundo extrasolar en una estrella de secuencia principal se dio en 1995. Mayor & Queloz (1995) detectaron el planeta 51 Peg b en una estrella similar al Sol utilizando el método de velocidades radiales (RVs). Con este método se revela indirectamente la presencia de un planeta a través del movimiento periódico que induce en una estrella a lo largo de su órbita alrededor del centro de masa del sistema estrella-planeta. Hablaremos en detalle de este método en el Cap. 2. Este hito inauguró una nueva rama de la astronomía y les valió a estos dos científicos el premio Nobel de física en 2019. En los siguientes 30 años el número de planetas que conocemos ha pasado de ocho a más de cinco mil, y el número crece cada vez más rápido.

51 Peg b y los planetas que le siguieron fueron detectados a través del mismo método de velocidades radiales. La velocidad radial no es más que la componente de la velocidad de la estrella en la dirección de nuestra línea de visión, y se obtiene a través de un instrumento llamado espectrógrafo, que separa la luz de la estrella según su longitud de onda, creando un *espectro* de la misma. En el Cap. 3 veremos cómo a partir de este se puede obtener la velocidad radial de una estrella. Si bien la espectroscopía estelar existe desde mediados del siglo XIX, el punto en que la técnica estuvo lo suficientemente desarrollada para poder medir el pequeño efecto causado por un planeta al orbitar una estrella recién se dio a fines del siglo XX. Mayor y Queloz usaron el entonces nuevo espectrógrafo ELODIE ubicado en el Observatorio de Haute-Provence para monitorear las velocidades radiales de una serie de estrellas. Con ELODIE era posible detectar movimientos causados por la presencia de planetas del tamaño de Júpiter o mayores, fue específicamente construído con este objetivo. Y así lo hicieron, encontraron a 51 Peg b, un tipo de planeta desconocido hasta ese momento: de una masa de alrededor de 0.46 veces la masa de Júpiter pero ubicado unas siete veces más cerca de su estrella que lo que está Mercurio del Sol. Luego resultó que estos planetas no son algo extraordinario, y además son los más fáciles de encontrar debido a la gran amplitud de los efectos que producen en sus estrellas. Se los denomina Júpiter calientes.

La aparición de los Júpiter calientes representó un desafío para los modelos de formación planetaria, en general no se creía posible la existencia de este tipo de planetas antes de ser descubiertos. Los dos escenarios que se consideran posibles para su formación son: el de *core acretion*, en el cual se requiere una formación rápida de un núcleo sólido que luego acreta gas del disco protoplanetario antes de que este se disipe (Pollack et al., 1996), y el de inestabilidad gravitacional del disco, en el cual el planeta se forma por contracción del gas en discos masivos (Boss, 1997). Sin embargo, sólo el primer escenario es compatible con la correlación observada entre ocurrencia de planetas gigantes y metalicidad estelar (Wang & Fischer, 2015).

La ciencia exoplanetaria ha avanzado mucho desde aquel puntapié inicial de Mayor y

Queloz, en los primeros años el método de RVs siguió siendo la única forma de encontrar otros mundos, hasta que en 1999 se detectó por el método de tránsitos el (ya conocido por RVs) exoplaneta HD 209458b (Charbonneau et al., 2000). El método de tránsitos es conceptual y técnicamente más simple que el de RVs. En éste, se monitorea el brillo de una estrella por un período de tiempo, si la estrella tiene algún planeta cuya geometría orbital está dispuesta de forma que en algún punto pasa por delante de la estrella (vista desde nuestra línea de visión), el pasaje va a generar una caída temporaria y periódica en el brillo que medimos de la estrella. Estas caídas son los tránsitos. Este método hoy en día es el que más exoplanetas ha aportado y lo trataremos más en detalle en la Sección 2.1. Estos dos métodos son los más importantes, pero hay muchos más, hoy incluso podemos ver exoplanetas de forma *directa* en ciertos casos.

Con la detección de exoplanetas convertida en algo ya rutinario dentro del área, la frontera del conocimiento se está inclinando hacia la caracterización en detalle de los candidatos más interesantes. Un número de nuevos telescopios e instrumentos poseen la capacidad de estudiar los elementos en las atmósferas de estos planetas, el primer paso para tratar de buscar condiciones favorables para la vida tal como la conocemos. Hoy se habla de detección de biomarcadores, mapas globales de temperatura, velocidades de vientos, cosas que hasta hace no muchos años eran imposibles de medir. Algunos de estos instrumentos ya están en uso y otros a pocos años de su primera luz, y necesitamos buenos exoplanetas candidatos para aprovecharlos al máximo.

A pesar de la gran cantidad de exoplanetas conocidos, los métodos con los que contamos no favorecen en general la detección de planetas similares a la Tierra. El método más apto para encontrar planetas rocosos de masas similares a la Tierra en zonas habitables de sus estrellas es el de velocidades radiales, pero se requiere de precisiones en las medidas de las RVs menores a 1 $m\ s^{-1}$ estables por largos períodos de tiempo. Para lograr esto, los espectrógrafos de RVs se colocan lejos del telescopio en habitaciones aisladas con control de vibraciones, temperatura y presión. La luz se traslada en fibras ópticas desde el telescopio al espectrógrafo y se realiza calibración simultanea de la longitud de onda mediante lámparas de referencia. Si bien la precisión de estos instrumentos va aumentando permanentemente, existen limitaciones a la precisión alcanzable debidas a errores sistemáticos difícil de corregir, a la actividad estelar intrínseca de las estrellas que se manifiesta en forma de ruido en las RVs y en última instancia al ruido fotónico inherente a la naturaleza de la luz.

1.2. Esta tesis

El trabajo presentado aquí combina métodos de la ciencia de datos con astronomía observacional para desarrollar una nueva forma de corregir los errores de punto cero en espectrógrafos de alta resolución con el objetivo de mejorar la precisión en la medición de las velocidades radiales. Explotando el efecto Doppler, los espectros de alta resolución de estrellas de tipo solar permiten obtener medidas de velocidades radiales con precisiones mejores que el metro por segundo. Así, es posible revelar sistemas con planetas de baja masa o en órbitas muy

alejadas de su estrella.

Nuestro trabajo se centra en el instrumento SOPHIE, un espectrógrafo *echelle* alimentado por fibra óptica que se encuentra en el telescopio de 1.93 m del Observatorio de Haute Provence, Francia. SOPHIE fue diseñado como el instrumento sucesor de ELODIE, para lograr precisiones de hasta 1 $m\ s^{-1}$ mediante el método de control del perfil instrumental (Figueira, 2018) y calibración simultánea de la longitud de onda. Si bien alcanza precisiones entre 1 y 2 $m\ s^{-1}$, lo que lo hace capaz de detectar planetas dentro del orden de las supertierras (Díaz et al., 2019), el punto cero del instrumento muestra una deriva a largo plazo. Actualmente este efecto se corrige mediante la observación regular de estrellas estándar de velocidad radial (Courcol et al., 2015).

La idea es aprovechar la gran cantidad de observaciones de estrellas constantes en la base de datos de SOPHIE para modelar los efectos sistemáticos del instrumento. Al ser estrellas brillantes y con una velocidad radial fija y bien conocida, la dispersión de las medidas está dominada por los errores sistemáticos del instrumento. Las mediciones de velocidad radial de SOPHIE vienen acompañadas con una enorme cantidad de variables ambientales, observacionales e instrumentales, que no siempre son fácilmente accesibles. Recogimos estas variables para un conjunto de estrellas constantes, bajo la hipótesis de que servirán como variables explicativas razonables de los efectos sistemáticos observados. Sin embargo, sospechamos que no debe existir una relación sencilla entre estas variables y los efectos sistemáticos en velocidad radial.

El uso de algoritmos de aprendizaje automático es ideal para el tipo de problema que queremos abordar, ya que permiten modelar un problema a partir de un gran número de variables sin tener que definir una forma funcional explícita. Además, en muchos casos es posible extraer de los algoritmos información valiosa sobre la importancia de estas variables en la predicción, es decir, en cierta forma podemos acceder a lo que "aprende" el algoritmo. Modelar los errores sistemáticos nos permitirá proporcionar un nuevo método de corrección y a su vez, identificando cuáles son las variables de mayor importancia en los algoritmos podemos llegar a comprender las causas de los mismos.

El abordaje propuesto presenta un grado de originalidad, ya que no hemos encontrado en la bibliografía trabajos que usen algoritmos de aprendizaje automático para modelar errores sistemáticos en este tipo de instrumentos. Todo el código se escribió en el lenguaje Python, utilizando el entorno de ejecución Google Colab. El código principal al finalizar contó con alrededor de ∼3300 líneas. La implementación de los algoritmos de aprendizaje automático se hizo con scikit-learn (Pedregosa et al., 2011).

La tesis está organizada de la siguiente manera:

- En el Cap. 2 se pone en contexto el estado del conocimiento sobre exoplanetas y se detallan las principales técnicas de detección.
- En el Cap. 3 se describe el método de velocidades radiales y se introduce el espectrógrafo SOPHIE.

- En el Cap. 4 se introducen definiciones y conceptos elementales sobre aprendizaje automático y se detallan los algoritmos a utilizar.
- En el Cap. 5 se describe la etapa de adquisición y preparación de los datos.
- En el Cap. 6 se detalla el entrenamiento, evaluación y selección de los algoritmos. Luego se aplica el modelo a nuevos datos y se comparan los resultados con el método tradicional de corrección.
- Finalmente en el Cap. 7 se establecen las conclusiones y se menciona el trabajo a futuro.

Capítulo 2

Exoplanetas

Estrictamente hablando, debemos mencionar que los primeros exoplanetas fueron encontrados alrededor de púlsares. Wolszczan & Frail (1992) y Wolszczan (1994) confirmaron un sistema de tres planetas orbitando el púlsar PSR B1257+12. Los detectaron mediante el método de variación del período de púlsares. Este método aprovecha el hecho de que los púlsares se comportan como relojes cósmicos muy estables, de manera que pequeñas variaciones en los tiempos de púlsación se pueden relacionar con un movimiento orbital del cuerpo central causado por la presencia de un planeta orbitándolo a una distancia cercana. Entonces, cuando nos referimos a 51 Peg b como el primer exoplaneta, en realidad estamos diciendo el primer exoplaneta en una estrella de secuencia principal; en general estos son nuestros principales objetos de interés.

Existen al menos once métodos para detectar exoplanetas. Describiremos en detalle en la siguiente sección los cuatro principales, que han aportado la inmensa mayoría de los exoplanetas conocidos (ver Figura 2.1). Estos son: el método de velocidades radiales, el de tránsitos, el de imagen directa y el de microlentes gravitacionales.

2.1. Métodos de detección

2.1.1. Velocidad Radial

Cuando decimos que, la Tierra por ejemplo, *orbita en torno al Sol*, nos estamos ubicando en un sistema de referencia con el Sol fijo. Pero físicamente, si consideramos sólo el sistema Tierra-Sol lo que sucede en realidad es que ambos orbitan alrededor del centro de masa del sistema (Figura 2.2). El centro de masa se encontrará más alejado del centro de la estrella cuanto más masivo sea el planeta y más amplia su órbita. Lo mismo sucede en todos los sistemas extrasolares, por lo que si detectamos este movimiento orbital en una estrella, es una forma indirecta de detectar la presencia de un exoplaneta.

En general, para la mayoría de las estrellas no es posible determinar su velocidad espacial, pero sí podemos mediante espectroscopía Doppler obtener medidas muy precisas de su velocidad radial, es decir la componente de la velocidad en la dirección de la visual. Si realizamos una serie temporal de medidas suficientemente precisas de esta velocidad para una estrella,

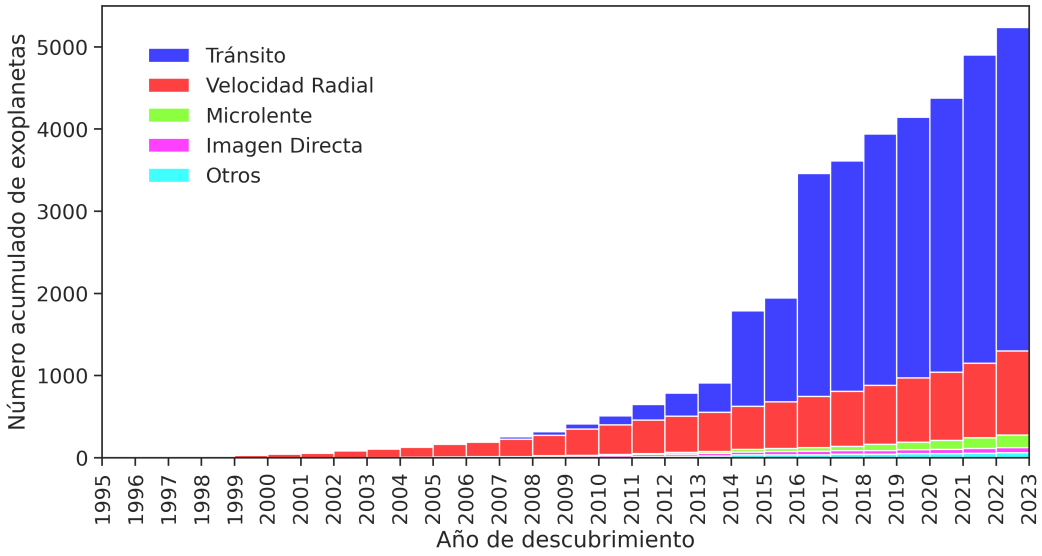

Figura 2.1. Histograma de número acumulado de planetas descubiertos por año desde 1995 hasta 2023. Los colores representan el método de descubrimiento.

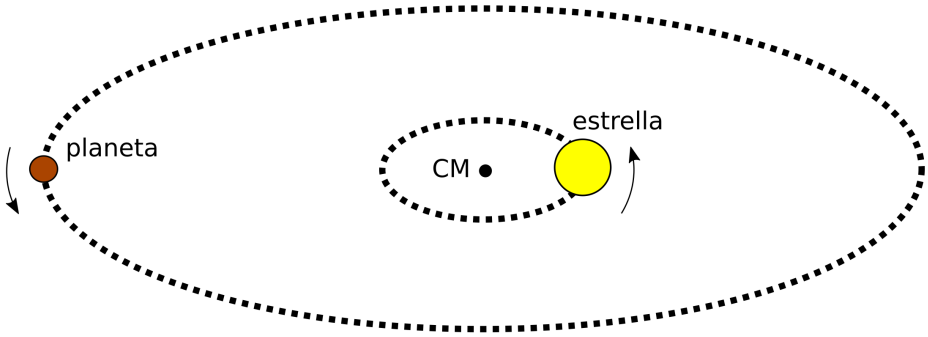

Figura 2.2. El esquema muestra las trayectorias orbitales de un planeta y una estrella alrededor de su centro de masa.

el efecto del movimiento alrededor del centro de masa del sistema se manifestará en forma de una señal periódica. Esto exceptuando, por supuesto, el caso en que el plano de la órbita sea perpendicular a la línea de visión. El método de velocidades radiales entonces, consiste en encontrar y modelar estas señales en datos de velocidades radiales. En la Figura 2.3, extraída del trabajo pionero de Mayor & Queloz (1995) se muestran curvas de velocidades radiales de 51 Peg b en función de la fase orbital para distintas épocas.

El estudio de estas curvas fue desarrollado originalmente para el estudio de estrellas binarias espectroscópicas. Se encuentra que los datos son bien descriptos por una curva Kepleriana:

$$V = V_0 + K[\cos(\nu + \omega) + e\cos\omega] \tag{2.1}$$

donde ν es la anomalía verdadera, ω es el argumento del periastro[(i)], e es la excentricidad y K es la semiamplitud. En el caso especial de un exoplaneta, donde la masa del mismo puede

[(i)] las definiciones precisas de la anomalía verdadera y el argumento del periastro se dan en la sección 3.1

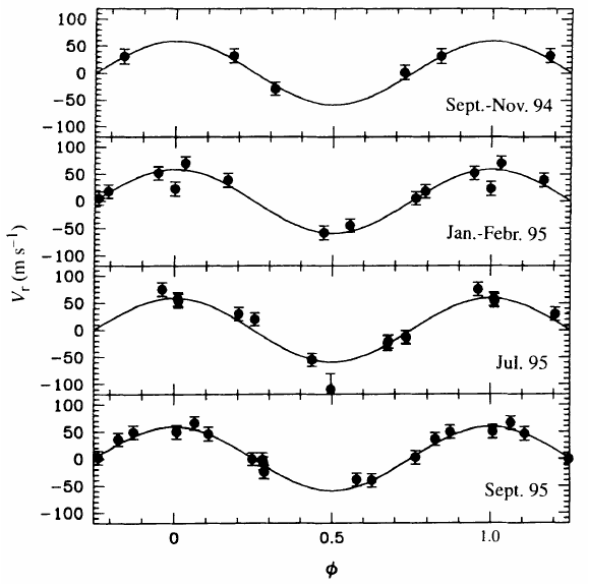

Figura 2.3. Movimiento orbital de 51 Peg b en cuatro épocas distintas. La línea sólida representa el modelo orbital y los puntos las velocidades radiales medidas. Mayor y Queloz, 1995.

despreciarse frente a la masa de la estrella, K se relaciona con los parámetros orbitales del planeta por:

$$K \approx \left(\frac{2\pi G}{P M_*^2}\right)^{\frac{1}{3}} \frac{M_p \sin i}{\sqrt{1-e^2}} \tag{2.2}$$

aquí P es el período, e es la excentricidad, i es la inclinación de la órbita y M_p y M_* son la masa planetaria y estelar respectivamente. Dado que cuanto mayor sea K será más fácil detectar el planeta, se deduce de la Ec. 2.2 que este método favorece la detección de planetas con períodos cortos, masas grandes y alrededor de estrellas de baja masa. Por otra parte, se requieren fuentes relativamente brillantes para poder hacer espectroscopía de alta resolución, limitando la detección a estrellas cercanas.

La importancia del método de velocidades radiales radica en que por sí solo nos permite obtener cotas inferiores para las masas de los planetas, y en combinación con el método de tránsitos, a partir del cuál obtenemos la inclinación orbital y el cociente de radios entre la estrella y el planeta, es posible derivar las masas absolutas y densidades de los planetas. Gracias a relevamientos fotométricos como el llevado a cabo con el telescopio espacial Kepler (Koch et al., 2010) y el más reciente TESS (Transiting Exoplanet Survey Satellite, Ricker et al. (2015)) se han detectado miles de candidatos de tránsitos exoplanetarios. Para confirmar y caracterizar estos sistemas se requieren observaciones adicionales de velocidades radiales precisas. Además, es el método apuntado como el más prometedor para encontrar planetas análogos a la Tierra en estrellas de tipo solar en el futuro cercano. La amplitud en velocidad radial que produciría tal planeta es de 9 cm s^{-1} con un período de 1 año (Figura 2.4). Si bien instrumentalmente ya es posible alcanzar este orden de precisión como se ha demostrado con el espectrógrafo ESPRESSO (Pepe et al., 2021), la posibilidad de detección

de este tipo de planetas actualmente está limitada por el efecto producido en las RVs por la variabilidad estelar intrínseca. Existen un número de fenómenos físicos que pueden producir variaciones aparentes en las RVs que contaminan las mediciones y son fuentes de falsos positivos, principalmente los relacionados con la actividad magnética como las manchas oscuras (Boisse et al., 2011) y los ciclos de actividad magnética (Díaz et al., 2016b), u otros como la granulación y las pulsaciones estelares (Santos et al., 2011).

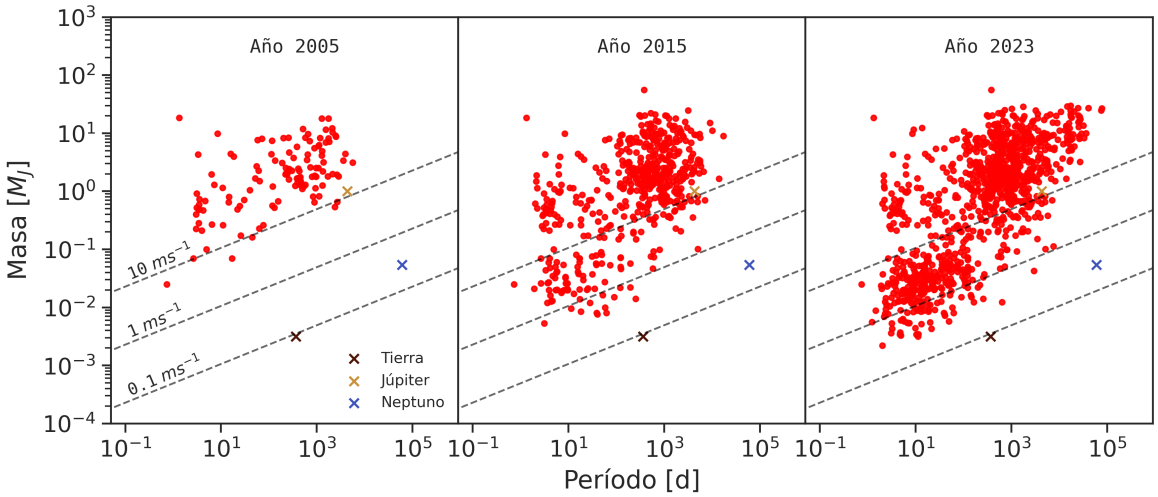

Figura 2.4. Planetas detectados mediante el método de velocidades radiales, acumulados hasta 2005, 2015 y 2023. En líneas punteadas se marca la amplitud en las curvas de velocidad radial producidas por los planetas (en una estrella de 1 $M_\odot$). Como referencia se ubica a Júpiter, Neptuno y la Tierra en el gráfico.

Por último, otra desventaja que tiene respecto al método de tránsitos es que no es posible hacerlo para muchas estrellas al mismo tiempo en una sola observación. Además de que para planetas en períodos largos hay que observar durante muchos años continuamente para poder hacer una detección.

2.1.2. Tránsitos

La detección de exoplanetas mediante tránsitos ha sido el método más eficiente hasta el momento. El fenómeno de tránsito planetario se conoce hace cientos de años, observándose para varios cuerpos en nuestro sistema solar. El primer registro que se tiene es de 1631 cuando Pierre Gassendi siguiendo predicciones de Kepler observó un tránsito de Mercurio por delante del Sol (Hortensius & Gassendi, 1633).

La situación física es simple, un tránsito se da cuando -desde el punto de vista del observador- un objeto menor pasa por delante de otro, bloqueando parte de su luz. En el sistema solar esta situación sólo puede darse para Venus y Mercurio si hablamos de tránsitos solares, o para lunas de los planetas exteriores, siendo el ejemplo más característico el de los tránsitos de las lunas galileanas de Júpiter. Tanto en el sistema solar como para planetas extrasolares, es claro que esta situación sólo se da para sistemas con una particular orientación geométrica según nuestra perspectiva (ver Figura 2.5).

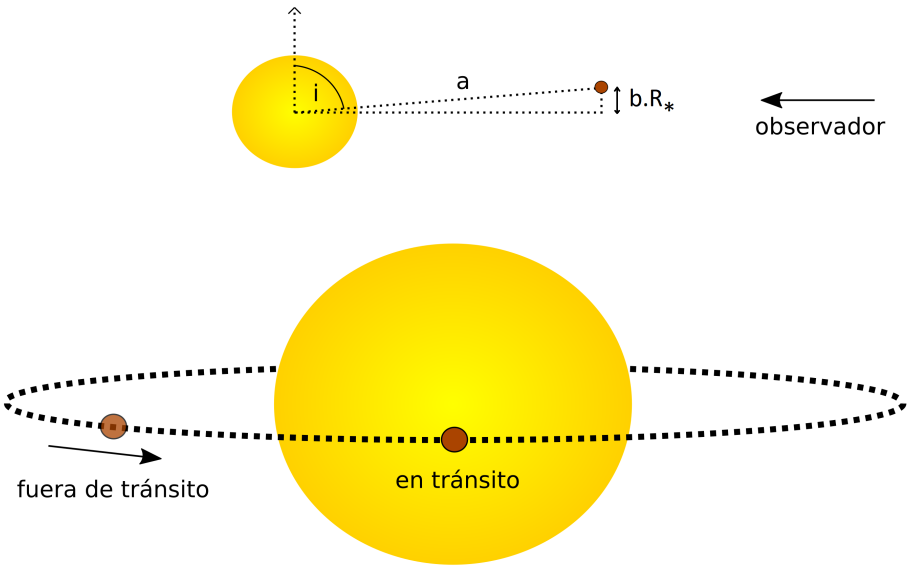

Figura 2.5. Situación esquemática en la que se observaría un tránsito. Donde a es el semieje mayor de la órbita, b es el parámetro de impacto, i la inclinación de la órbita y $R_\star$ es el radio de la estrella.

El observable de este método es la disminución del flujo estelar que se da durante el tránsito. Si suponemos que ambos cuerpos son esféricos, que el planeta no aporta flujo y despreciamos el oscurecimiento al limbo, la caída de flujo se puede expresar como:

$$\Delta F \approx \left(\frac{R_P}{R_*}\right)^2 = k^2 \tag{2.3}$$

Donde R_P y R_* son los radios del planeta y la estrella respectivamente, y k es el cociente de radios. Además se puede obtener la duración del tránsito t_T y el período P, que es el tiempo entre dos tránsitos consecutivos. De la Ec. 2.3 se deduce que será más fácil detectar un tránsito cuanto mayor sea el radio del planeta y menor el de la estrella. Otra cosa a considerar es el parámetro de impacto b (ver Figura 2.5), que para órbitas circulares se define por:

$$b = \frac{a}{R_*}\cos i \tag{2.4}$$

De la ecuación anterior y la figura se deduce que un tránsito por el medio del disco estelar tendrá un $b = 0$ y el máximo valor de b para el cuál se puede dar un tránsito es $1 + R_P/R_*$, por tanto para que se dé el tránsito se debe cumplir:

$$\begin{aligned} b &\leq 1 + \frac{R_P}{R_*} \\ \frac{a}{R_*}\cos i &\leq 1 - \frac{R_P}{R_*} \\ a_{critico} &\leq \frac{R_* - R_P}{\cos i} \end{aligned}$$

Por tanto, cuanto menor sea el semieje mayor del planeta se tiene un mayor rango de inclinaciones para las cuales se puede detectar un tránsito, de manera que el método también

favorece a planetas con órbitas pequeñas. En la Figura 2.6 se muestran dos de los tránsitos usados en la primera observación por este método de un planeta extrasolar en el año 2000, el Júpiter caliente HD 209458b (Charbonneau et al., 2000).

El gran éxito de este método evidenciado en la Figura 2.1 se debe a que al estar basado en mediciones fotométricas es posible hacerlo para miles de estrellas de un campo dentro de la misma observación, es por eso que, incluso teniendo en cuenta los fuertes efectos de selección del método, igualmente se encuentran muchos planetas. Entre los distintos telescopios diseñados para sondear el cielo en busca de estos eventos de tránsitos se han medido más de ciento treinta millones de curvas de luz[(ii)].

Cuando se tienen muchos tránsitos de un mismo planeta, el estudio de las pequeñas variaciones en los tiempos de tránsito debido a perturbaciones de la órbita debido a la presencia de más planetas en el sistema, da lugar a otro método de detección llamado Variaciones de Tiempo de Tránsito (TTV).

Como todo método, el método de tránsitos cuenta con algunas desventajas, una de ellas es que no permite obtener masas planetarias, sólo su radio (excepto cuando hay TTVs, en esos casos sí es posible obtener masas). Además, para descartar falsos positivos que pueden surgir por aparición de manchas estelares o estrellas binarias es necesario recurrir a otros métodos, usualmente el de velocidades radiales, que al requerir fuentes brillantes para poder aplicarlo, sólo posible para estrellas cercanas. Si bien también es posible validarlos mediante métodos estadísticos (Mistry et al., 2023; Christiansen et al., 2022), existen varios miles de candidatos a tránsitos detectados por Kepler y TESS que todavía no han podido ser confirmados.

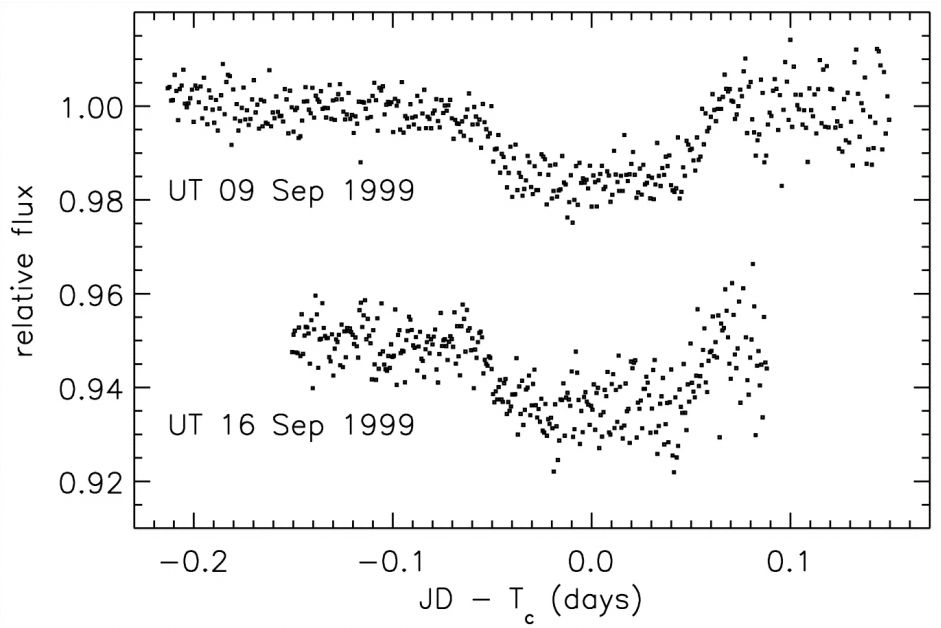

Figura 2.6. Tránsitos observados de HD 209458. El primer exoplaneta detectado mediante este método (Charbonneau et al., 2000).

[(ii)] fuente: https://exoplanetarchive.ipac.caltech.edu/

2.1.3. Imagen Directa

Es el único método directo, es decir donde se obtiene una *imagen* del exoplaneta propiamente dicho. En particular, se observa la luz infrarroja emitida por planetas gigantes jóvenes en órbitas muy amplias. Recién en 2004 se detectó el primer exoplaneta con este método: 2M1207b, de varias veces la masa de Júpiter y en una órbita de 40 AU alrededor de una enana marrón (Chauvin et al., 2004).

Para que sea posible obtener una imagen del planeta se necesita que esté separado por algunas unidades de resolución angular de la estrella, que para un sistema limitado por difracción escala con la longitud de onda y el tamaño del telescopio con λ/D. Para telescopios de aperturas $\geq 8\ m$ que es donde se instalan usualmente los instrumentos de imagen directa, se pueden detectar planetas a separaciones angulares de $\sim 100-300\ mas$, lo cual corresponde a $\sim 3-9\ AU$ para una fuente a 30 pc. A estas distancias orbitales la luz reflejada por los planetas es muy débil por lo que lo que la detección se da a través de la luz proveniente de su emisión térmica. En la Figura 2.7 se aprecia la gran diferencia en el cociente de flujos entre la estrella y el planeta para un sistema joven y uno maduro, y la disminución de este cociente con la longitud de onda. Para un análogo terrestre, a longitudes visibles el cociente de brillos está en el orden de 10^{10} mientras que en infrarrojo es del orden de $10^6 \sim 10^7$.

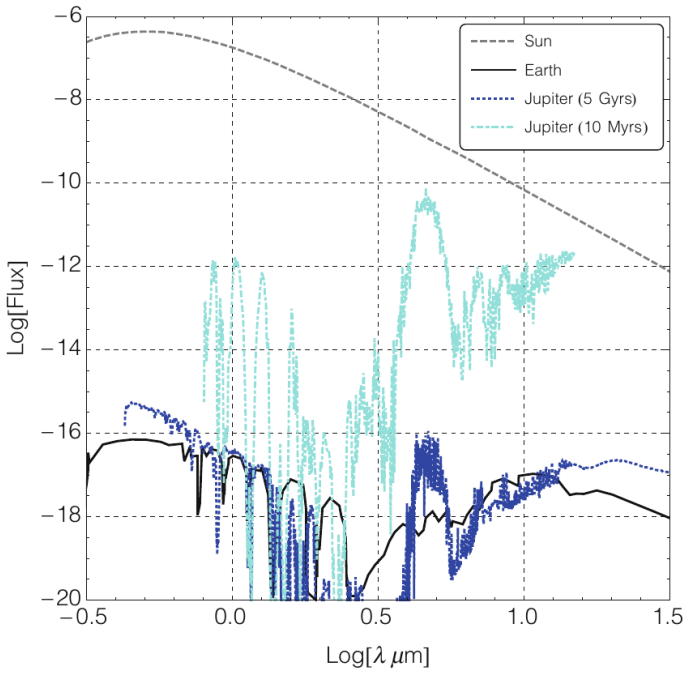

Figura 2.7. Comparación de brillos en función de la longitud de onda entre una estrella de tipo solar (representada por un cuerpo negro a 5870 K) y tres tipos de planetas: un Júpiter joven, un Júpiter maduro y un análogo terrestre. Se aprecia cómo varía el cociente de flujos entre el planeta y la estrella según la edad del sistema. (Laurent Pueyo, Direct Imaging as a Detection Technique for Exoplanets, Pueyo (2018).)

Por eso, se busca observar planetas jóvenes de manera que todavía se están enfriando, con temperaturas en el rango de $600-2000\ K$. Si se supone emisión de cuerpo negro, la longitud

de onda más favorable es en el infrarrojo, entre 1.4 y 4.8 μm. Los parámetros que se pueden medir mediante imagen directa son el radio orbital y la temperatura del planeta, en tanto que la masa puede estimarse sin mucha precisión a partir de la temperatura del planeta y la edad del sistema. En algunos casos se puede estimar el radio a partir de la temperatura, el brillo aparente y la distancia. Una gran ventaja de observar directamente el planeta es que se le puede sacar un espectro y estudiar su composición, además, debido a que se usan coronógrafos, para sacar el espectro no es necesario separar la luz del planeta de la de la estrella.

Un coronógrafo, cuyo nombre se debe a que originalmente fue diseñado para poder observar la corona del Sol, es un dispositivo que se introduce en el camino óptico del telescopio para bloquear la luz proveniente de la estrella central y dejar visible sólo al planeta, que de otra forma quedaría tapado por el gran brillo de la estrella. Debido a que la técnica requiere una alta estabilidad óptica, sólo es posible desde tierra mediante uso de sistemas modernos de óptica adaptativa. Los telescopios espaciales son ideales en ese sentido, al costo de menores aperturas que los de Tierra. En particular, el telescopio WFIRST a lanzarse en 2027 será capaz de detectar la luz reflejada por planetas análogos a Júpiter (Krist et al., 2016).

En la Figura 2.8 se muestran dos imágenes directas de exoplanetas, correspondientes a 2M1207b, el primer exoplaneta detectado con este método. Y uno más reciente, 51 Eridani b, el cuál fue detectado en 2015 con el instrumento Gemini Planet Imager (GPI).

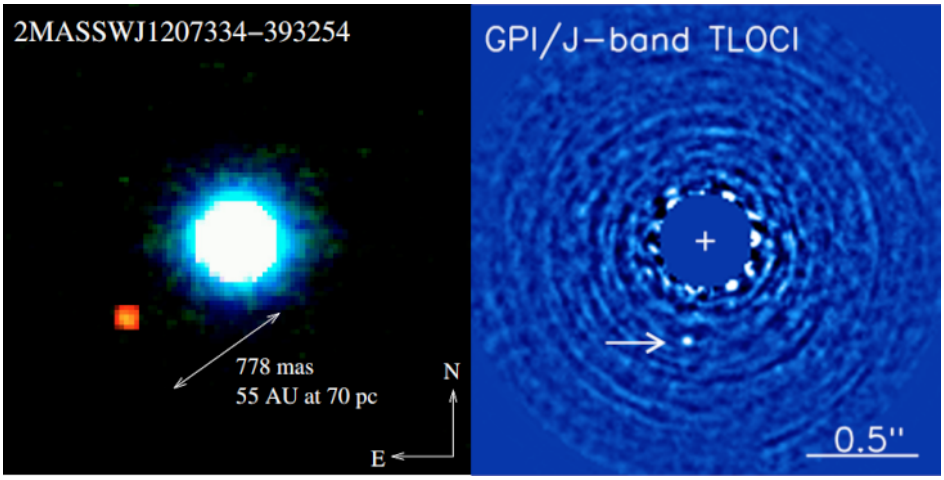

Figura 2.8. *Panel izquierdo:* Primera imagen directa de un exoplaneta. Obtenida con el instrumento NACO del telescopio VLT (Chauvin et al., 2004). / *Panel derecho:* Imagen del exoplaneta 51 Eridani b obtenida con el instrumento GPI, donde se aprecia el uso de un coronógrafo para tapar bloquear la luz de la estrella central (Macintosh et al., 2015).

2.1.4. Microlentes gravitacionales

El último método que veremos es otro indirecto y tiene como base las lentes gravitacionales. En este fenómeno que fue descripto por Albert Einstein en 1936, la luz de una fuente distante se curva al pasar cerca del campo gravitacional de un cuerpo masivo, de forma que

éste actúa como una lente. En las microlentes, el cuerpo que genera la curvatura de la luz es una estrella.

Este hecho puede usarse para la detección de exoplanetas de la siguiente forma: en la línea de visión del observador se tienen dos estrellas muy bien alineadas, una de fondo y una al frente (Figura 2.9). La estrella más cercana actúa de lente para la luz de la estrella de fondo, produciendo un aumento en el brillo de esta mientras dura el evento de microlente (la alineación). Si la estrella que está actuando de microlente tiene un planeta en órbita, este también actúa como una pequeña microlente adicional, generando un segundo aumento en el brillo de la estrella de fondo (Figura 2.10). El observable en este caso es la magnificación del brillo estelar, y los parámetros que permite deducir son la masa y el radio orbital.

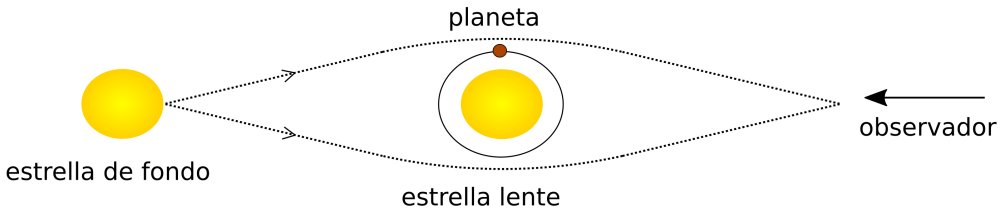

Figura 2.9. Situación esquemática de la geometría que se da en el método de microlentes gravitacionales. La líneas punteadas representan dos rayos de luz que son curvados por la estrella lente.

El método de microlentes gravitacionales tiene algunas ventajas únicas. Mientras otras técnicas son aplicables a enanas F, G, K y más recientemente se están extendiendo a M, éste es sensible a todo tipo de estrellas y remanentes. Sumado a que permite detectar planetas con un amplio rango de masas y a grandes distancias orbitales -incluídos análogos a la Tierra-, y que permite detectar planetas errantes, un sondeo de microlentes tal como el que llevará a cabo WFIRST (Spergel et al., 2015) permitirá obtener información demográfica independiente y complementaria a la de otros métodos, lo cuál tendrá un gran impacto en el entendimiento de los procesos de formación planetaria. Mientras que una de sus principales desventajas reside en la irrepetibilidad de los eventos y la nula posibilidad de observaciones de seguimiento con otros métodos dado que en general se encuentran en estrellas a varios kiloparsec de distancia o directamente no se conoce cuál es la estrella lente.

2.2. Visión general

Con los cuatro métodos descriptos anteriormente se han encontrado casi el 99 % de los 5241 exoplanetas conocidos hasta la fecha (Enero de 2023[(i)]). En la Figura 2.11 se exhibe un gráfico de masas en función del período orbital, donde se pueden distinguir a grandes rasgos tres aglomeraciones de planetas. Arriba a la izquierda tenemos los del tipo "Júpiter

[(i)] fuente: https://exoplanetarchive.ipac.caltech.edu/

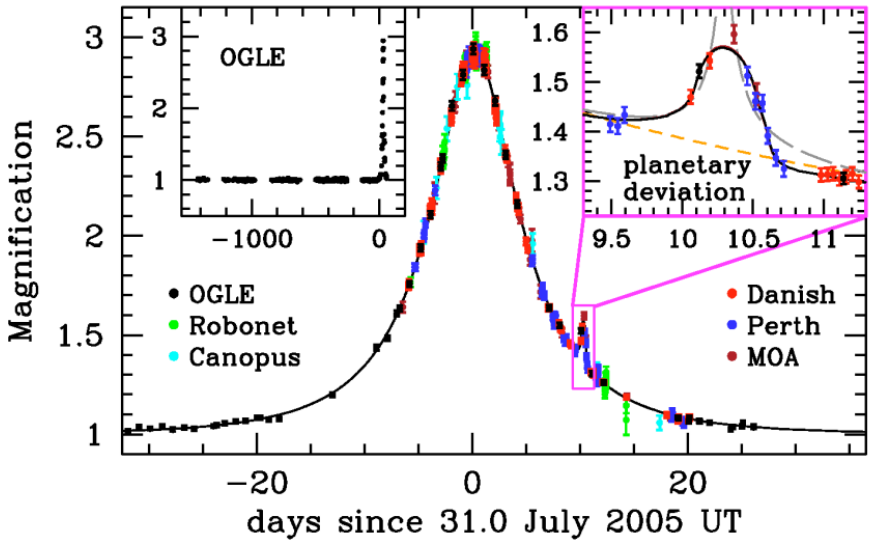

Figura 2.10. Curva de magnificación de la luz usada en la detección de la primer supertierra encontrada con este método, OGLE-2005-BLG-390Lb (Beaulieu et al., 2006).

caliente", planetas con masas del orden de Júpiter con periodos menores a 10 d. No sorprende que sean mayormente descubiertos por tránsitos, ya que son muy fácil de detectar en sondeos fotométricos incluso a distancias de un par de kiloparsecs. En cambio, los planetas detectados por velocidades radiales son todos en estrellas cercanas. A la derecha de estos, tenemos otro grupo de gigantes con períodos más largos, estos son más eficientemente detectados por el método de las velocidades radiales dado que un tránsito es menos probable. Hay una evidente sobrerrepresentación de planetas gigantes, Cumming et al. (2008) encuentra que sólo un $\sim 10\,\%$ de las estrellas de tipo solar tienen planetas con masas $0.3 < M_J < 10$ y períodos entre 2 y 2000 días. Y la ocurrencia es mucho menor para los Júpiter calientes, cercana al $1\,\%$ según Mayor et al. (2011). Es claro que esta muestra de la población exoplanetaria tiene algunos sesgos importantes, principalmente hacia planetas masivos y en órbitas pequeñas. Los planetas pequeños, ausentes prácticamente en la muestra actual, deberían ser frecuentes según modelos de formación planetaria (Miguel et al., 2011).

El otro grupo que se evidencia es el de las supertierras y Neptunos, donde hay tanto detectados por RVs como por tránsitos. Estos planetas son extremadamente comunes, se calcula que la mitad de las estrellas de tipo solar tienen al menos un planeta en este rango de masas con un período menor a 100 días (Mayor et al., 2011). Y aquí es interesante notar un vacío que aparece en períodos bajos tanto en masas como en radios (Figuras 2.11 y 2.12) para planetas de tipo Neptuno, esta región donde los métodos observacionales son eficientes pero sin embargo se encuentran pocos planetas se conoce como *desierto Neptuniano* (Mazeh et al., 2016). La explicación de este "faltante" de Neptunos calientes tiene que ver con que la fuerte radiación estelar produce fotoevaporación en las atmósferas de estos planetas, dejando solo un núcleo rocoso.

En cuanto a planetas de masa similar a la Tierra dentro de la zona habitable, es muy difícil estimar su ocurrencia dado que las técnicas de detección actuales son apenas sensibles

a este tipo. Por ejemplo mediante tránsitos, la caída de flujo que produce un análogo a la Tierra sería de un 0.008 %, algo detectable con instrumentos actuales, pero como vimos la probabilidad de tránsito disminuye con el período orbital y detectar varios tránsitos es aún más complicado ya que requiere observar por un largo período de tiempo. Este fue uno de los objetivos científicos detrás de la misión Kepler, que monitoreó la misma región del cielo durante casi 5 años. Kepler sólo detectó alrededor de 10 de estos planetas (Winn, 2018). Las mejores estimaciones basadas en los datos de la misión Kepler ubican la ocurrencia de análogos terrestres entre 0.05 y 11.5 % (Winn, 2018).

Respecto a microlentes gravitacionales, con el futuro telescopio espacial WFIRST a lanzarse en 2027 se estima que podrían detectarse unos 180 exoplanetas de masa terrestre con $a > 1$ AU (Penny et al., 2019) pero en estrellas muy lejanas. Otra opción a futuro es con astrometría relativa, la cuál requiere poder medir una señal de 0.3 μas, algo que hoy está lejos y que sólo será posible desde el espacio. Es generalmente aceptado que el camino más factible en el futuro cercano para la detección desde tierra de análogos terrestres orbitando estrellas cercanas de tipos F, G y K es con mediciones de velocidad radial de precisión extrema (< 10 cm s^{-1}), algo a lo que poco a poco nos vamos acercando, como se muestra en la Figura 2.4.

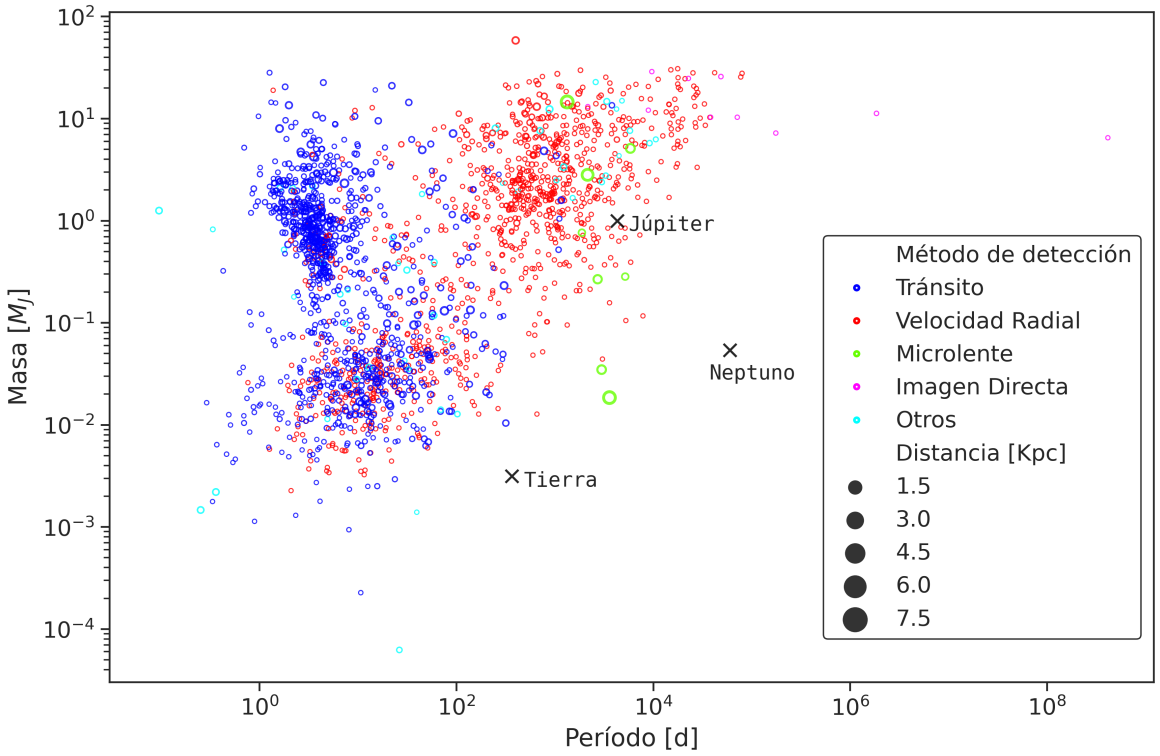

Figura 2.11. Masas en función de período para la población de exoplanetas conocidos hasta la fecha. Se diferencia con colores según método de descubrimiento y con tamaño de los puntos según la distancia al sistema planetario. Para referencia se marca la ubicación de Júpiter, Neptuno y la Tierra.

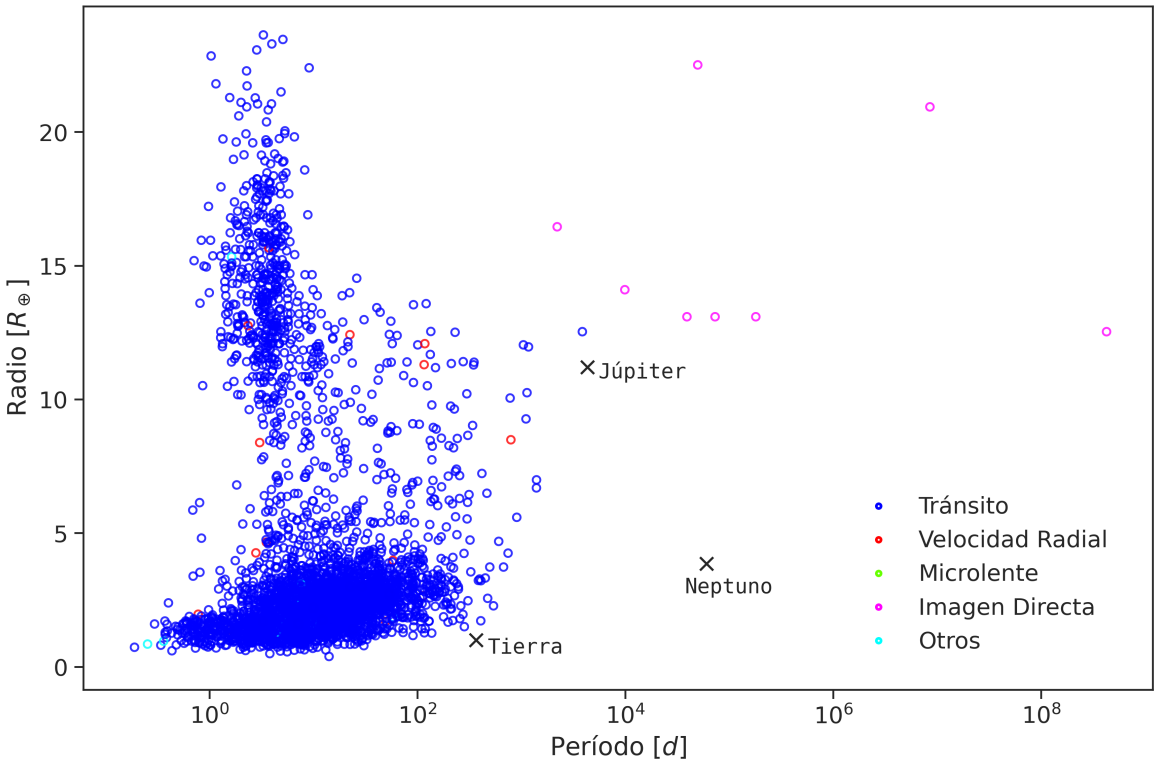

Figura 2.12. Radios en función de período para la población de exoplanetas conocidos hasta la fecha. Se diferencia con colores según método de descubrimiento. Para referencia se marca la ubicación de Júpiter, Neptuno y la Tierra.

Capítulo 3

Velocidades Radiales

Quizá lo que más diferencia la Astronomía entre las demás ciencias naturales es que la gran mayoría de las veces la única fuente de información que tenemos sobre los objetos de estudio es la luz que nos llega de ellos. Estos fotones representan el estado del objeto en el momento en que fueron emitidos y deben atravesar distancias inabarcables para nuestra imaginación antes de llegar a los telescopios. El astrónomo entonces, se ha especializado en extraer toda la información posible que puedan contener, y una de las herramientas más fructíferas ha sido el análisis espectroscópico de la luz. Con el desarrollo de la espectroscopía, hemos aprendido que una propiedad particular del objeto emisor de la luz viene codificada en ella; hablamos de la velocidad radial. Resulta interesante pensar la cantidad de cosas que este simple dato -a qué velocidad la fuente se aleja o se acerca- ha ayudado a descubrir: la existencia de planetas invisibles para nosotros, la materia oscura, y nada menos que la expansión acelerada del Universo.

3.1. Modelo físico

Se describe el modelo físico que reproduce el cambio en el tiempo de la velocidad radial de una estrella producido por la interacción gravitacional con un compañero planetario[(i)]. Antes de comenzar haremos las siguientes suposiciones: 1) Que los cuerpos son objetos puntuales sin dimensión. 2) Que siguen las leyes de la dinámica Newtoniana. 3) Que no tienen otra interacción más que la gravitacional. Bajo estas hipótesis, el movimiento del planeta y la estrella se reduce al conocido problema de dos cuerpos. Nos interesa además las posiciones en función del tiempo (problema de Kepler). Consideremos las ecuaciones de movimiento de los objetos m_1 (estrella) y m_2 (planeta) bajo interacción gravitatoria, descriptas en un sistema de referencia inercial con origen en O:

$$\vec{F}_1 = m_1 \ddot{\vec{r}_1} = +G\frac{m_1 m_2}{r^3}\vec{r} \tag{3.1}$$

$$\vec{F}_2 = m_2 \ddot{\vec{r}_2} = -G\frac{m_1 m_2}{r^3}\vec{r} \tag{3.2}$$

[(i)] Desarrollo basado en el trabajo de Díaz (2018).

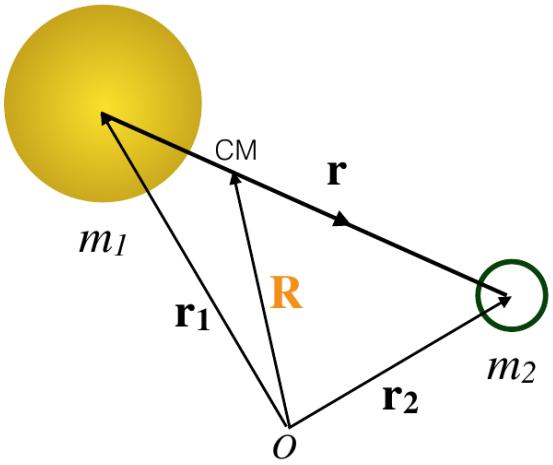

Figura 3.1. Esquema del problema de dos cuerpos. Díaz (2018).

donde G es la Constante de gravitación Universal, $\vec{r_1}$ y $\vec{r_2}$ son las posiciones de la estrella y el planeta respecto al sistema O y $\vec{r}$ es el vector de posición relativa que va desde la estrella al planeta (Figura 3.1). Y las posiciones de los cuerpos respecto del centro de masa (CM) del sistema son:

$$\vec{R_1} = -\frac{m_2}{m_1 + m_2}\vec{r} \tag{3.3}$$

$$\vec{R_2} = +\frac{m_1}{m_1 + m_2}\vec{r} \tag{3.4}$$

Si dividimos la Ec. 3.1 por m_1 y la 3.2 por m_2, y combinándolas se llega a la ecuación de movimiento relativo:

$$\ddot{\vec{r}} + G(m_1 + m_2)\frac{\vec{r}}{r^3} = 0 \tag{3.5}$$

Ahora tomando el producto vectorial de ambos lados de la Ec. 3.5 con el vector de posición relativa $\vec{r}$, y usando que $\vec{r} \times \vec{r} = 0$, se llega a $\vec{r} \times \ddot{\vec{r}} = 0$, integrando obtenemos la primera magnitud que se conserva en el problema de dos cuerpos:

$$\vec{r} \times \dot{\vec{r}} = \vec{h} \tag{3.6}$$

Que el producto vectorial entre la posición y la velocidad relativa sea un vector constante $\vec{h}$ nos indica que el movimiento del sistema se da en un plano perpendicular al mismo, que llamamos plano orbital. Luego, podemos reducir la descripción del problema a dos coordenadas. Es útil entonces usar coordenadas polares, con lo que: $\vec{r} = r\hat{r}$, $\dot{\vec{r}} = \dot{r}\hat{r} + r\dot{\theta}\hat{\theta}$. Luego, $\vec{h}$ queda;

$$\vec{h} = r^2\dot{\theta}\hat{z} \tag{3.7}$$

esta magnitud es una buena aproximación al momento angular del sistema para $m_1 \ll m_2$. Ahora, volvemos a la Ec. 3.5 y la reescribimos en coordenadas polares,

$$\ddot{r} - r\dot{\theta}^2 = -G\frac{m_1 m_2}{r^2} \tag{3.8}$$

Si sustituimos $u = 1/r$ y usamos que $h = r^2\dot{\theta}$, podemos reemplazar:

$$\dot{r} = -\frac{1}{u^2}\frac{du}{d\theta}\dot{\theta} = -h\frac{du}{d\theta} \tag{3.9}$$

$$\ddot{r} = -h\frac{d^2u}{d\theta^2}\dot{\theta} = -h^2u^2\frac{d^2u}{d\theta^2} \tag{3.10}$$

quedando entonces una ecuación diferencial de segundo orden:

$$\frac{d^2u}{d\theta^2} + u = G\frac{m_1 + m_2}{h^2} \tag{3.11}$$

cuya solución es:

$$u = G\frac{m_1 + m_2}{h^2}[1 + e\cos(\theta - \omega)] \tag{3.12}$$

volviendo atrás la sustitución, la solución para r queda:

$$r = \frac{p}{1 + e\cos(\theta - \omega)} \tag{3.13}$$

donde $p = h^2/(G(m_1+m_2))$ y e y ω son constantes de integración. Esta es la ecuación general de una cónica en coordenadas polares. Los casos de interés aquí se dan cuando $e = 0$ que la ecuación describe una órbita circular, y para $0 < e < 1$ que corresponde a una órbita elíptica con excentricidad e y semiejemayor a, con $p = a(1-e^2)$. Reemplazando para ese caso, queda:

$$r = \frac{a(1 - e^2)}{1 + e\cos(\theta - \omega)} \tag{3.14}$$

Las dos distancias extremas, llamadas *periapsis* y *apoapsis* ocurren para $\theta = \omega$ y $\theta = \pi + \omega$. Se puede definir un nuevo ángulo polar llamado la *anomalía verdadera* como $\nu = \theta - \omega$, finalmente la ecuación queda:

$$r = \frac{a(1 - e^2)}{1 + e\cos\nu} \tag{3.15}$$

Lo que nos interesa ahora es cómo se verá este movimiento orbital desde la Tierra. Empezamos definiendo la orientación de la órbita en el espacio. Para esto necesitamos tres ángulos. El ángulo entre el plano orbital y el plano del cielo será la inclinación i, y la intersección entre estos dos planos es la línea de los nodos. La longitud del nodo ascendente, Ω, es el ángulo entre una dirección de referencia en el plano del cielo y el radio vector en el nodo ascendente: que es el punto donde el planeta cruza el plano del cielo desde abajo hacia arriba. Por último, el argumento del periápsis, ω, es el ángulo entre el mismo radio vector y el periápsis medido en el plano orbital (ver Figura 3.2). Para obtener la velocidad radial de la estrella en el sistema, proyectamos la órbita a un sistema de coordenadas cartesianas centrado en el centro de masa del sistema y orientado de forma que el eje z apunte en la línea de visión del observador. Además, asumamos que la dirección de referencia definida por el eje x coincide con la línea de los nodos, de forma que $\Omega = 0$. En este sistema, las componentes cartesianas del movimiento referido al centro de masa (de las Ecs. 3.3 y 3.4) son:

$$x = R_1\cos(\omega + \nu) \tag{3.16}$$

$$y = R_1\sin(\omega + \nu)\cos i \tag{3.17}$$

$$z = R_1\sin(\omega + \nu)\sin i \tag{3.18}$$

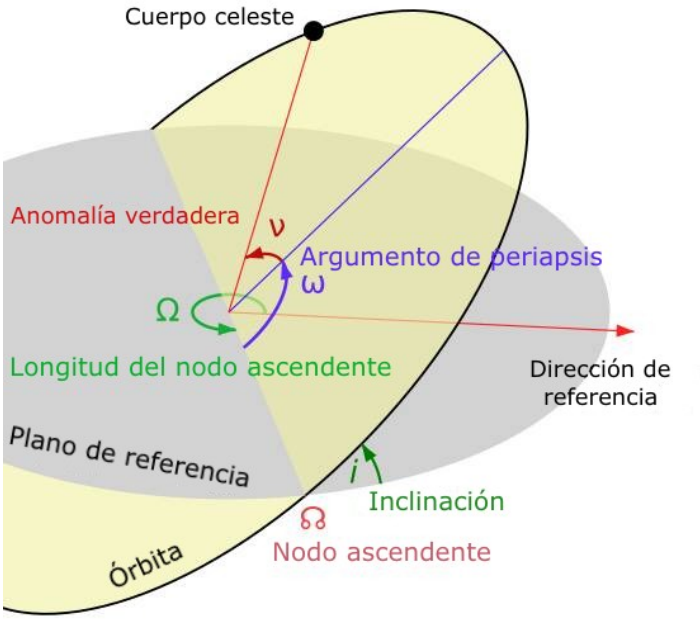

Figura 3.2. Elementos de la órbita en el espacio. Créditos: Wikipedia.

donde, de 3.3 y 3.15,

$$R_1 = a\frac{m_2}{m_1+m_2}\frac{(1-e^2)}{1+e\cos\nu} \tag{3.19}$$

es la distancia entre la estrella y el CM. Derivando la Ec. 3.18 con respecto a t y usando que:

$$\dot{\nu} = \dot{\theta} = h/r^2 = \frac{2\pi}{P}\frac{a^2\sqrt{1-e^2}}{r^2} \tag{3.20}$$

obtenemos la expresión de la velocidad radial de la estrella, $V = \dot{z}$, como función de la anomalía verdadera ν,

$$V = V_0 + K[\cos(\nu+\omega) + e\cos\omega] \tag{3.21}$$

con

$$K = \left(\frac{2\pi G}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}\frac{m_2\sin i}{(m_1+m_2)^{2/3}} \tag{3.22}$$

y donde V_0 es una constante que corresponde a la velocidad del CM respecto al observador desde la Tierra.

Lo último que nos falta hacer para relacionar el modelo con las observaciones es expresar la Ec. 3.21 en función del tiempo, es decir queremos encontrar $\nu(t)$, lo que constituye el problema de Kepler.

Daremos una descripción breve de la solución. Se definen dos nuevas variables auxiliares: la *anomalía excéntrica* ψ, y la *anomalía media* μ, que se definen por,

$$r = a(1-e\cos\psi) \tag{3.23}$$

$$\mu = \frac{2\pi}{P}(t-\tau) \tag{3.24}$$

donde τ es el tiempo de pasaje por el periápsis. Se puede mostrar que ψ y μ se relacionan por la ecuación de Kepler:

$$\mu = \psi - e \sin \psi \tag{3.25}$$

que es una ecuación trascendente, por lo cual se debe resolver mediante métodos numéricos. Luego invirtiendo la ecuación de Kepler, obtenemos ψ en función de t y por la Ec. 3.23 conocemos r en función de ψ, que si finalmente la combinamos con la Ec. 3.15 tenemos la relación entre la anomalía excéntrica y la verdadera:

$$\tan\left(\frac{\nu}{2}\right) = \sqrt{\frac{1+e}{1-e}} \tan\left(\frac{\psi}{2}\right) \tag{3.26}$$

Con esto, queda entonces expresada la dependencia temporal de la anomalía verdadera que necesitabamos para relacionar el modelo con las observaciones.

3.2. Espectroscopía Doppler

Un espectrógrafo nos permite analizar la intensidad de la luz en función de la longitud de onda λ, obteniendo un *espectro* de la misma. En estos aparecen líneas espectrales causadas por la absorción/emisión en longitudes de onda específicas por elementos químicos presentes en el camino de la luz. Estas líneas, para una fuente en reposo aparecen en una longitud de onda determinada, sin embargo, un movimiento relativo entre la fuente y el observador produce un cambio en la longitud de onda de las líneas. Si el movimiento entre ambos es de alejamiento, las líneas se desplazan hacia el rojo, mientras que si la fuente se acerca, el desplazamiento será hacia el azul. Este es el efecto Doppler.

Derivar velocidades radiales entonces implica obtener un espectro de la estrella y luego determinar el desplazamiento Doppler de las líneas. En su forma no relativista el desplazamiento se expresa como:

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c} \tag{3.27}$$

donde v es la velocidad radial de la fuente, c es la velocidad de la luz en el vacío, λ es la longitud de onda y $\Delta\lambda = \lambda - \lambda_0$ refiere a la diferencia entre la longitud de onda medida por el observador y la longitud de onda *en reposo* de la fuente. De la Ec. 3.27 vemos que el error en la velocidad radial va a estar relacionado con el ruido presente en el espectro y como éste afecta la incerteza en la determinación del desplazamiento de las líneas.

Ya hemos mencionado que la precisión en las medidas de velocidad radial limita la masa mínima de los planetas que podemos detectar. Aquí deberíamos aclarar a qué nos referimos con “precisión” y “exactitud” en las RVs. Precisión refiere a la incerteza estadística de cada medida asumiendo una estadística Gaussiana y que no hay correlación entre distintas observaciones. Es decir, la precisión nos dice la incerteza mínima e irreducible que podemos obtener para un dado sistema de medida. En cambio la exactitud tiene que ver con qué tan cerca está el valor medido de la ”verdad”, una medida puede ser precisa pero no exacta, o al revés. En el caso de RVs la inexactitud puede deberse a señales espurias que no se hayan considerado, como variabilidad estelar, contaminación telúrica o efectos instrumentales. Estos

efectos producen cambios en la forma o centroide de las líneas que no están relacionados con un verdadero movimiento de la fuente.

En RVs se suele usar el término *jitter* para designar al ruido no correlacionado que aparece debido a procesos físicos en las atmósferas estelares y que afecta las medidas de RV. Para mitigar el jitter, los sondeos se suelen concentrar en estrellas *tranquilas,* tipos F, G y K de secuencia principal. Pero a medida que se van minimizando los otros factores de error, la actividad estelar vuelve a ser un limitante a la precisión que un espectrógrafo puede alcanzar incluso con estas estrellas.

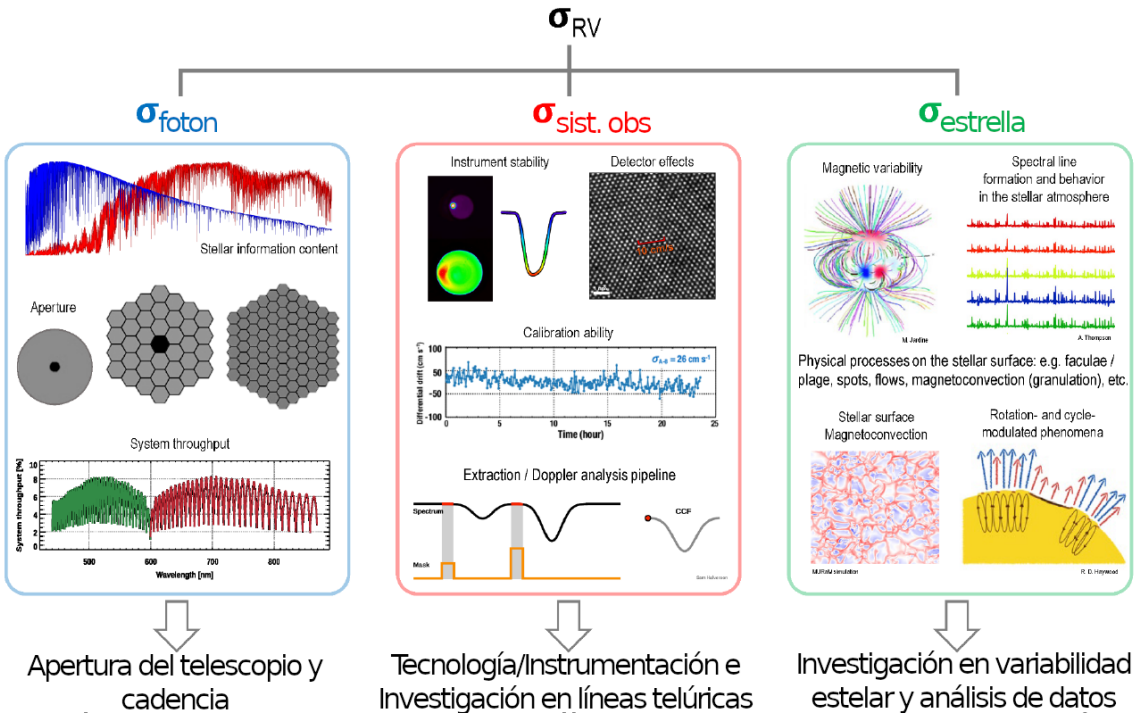

Figura 3.3. Distintas fuentes de incerteza en RVs, agrupadas en tres categorías: ruido fotónico, ruido por sistema observacional y ruido introducido por la estrella. Crédito: Sam Halverson. Crass et al. (2021).

Las fuentes de error en la determinación de RVs se pueden separar en tres amplios grupos (ver Figura 3.3). El error fundamental que marca el límite de precisión que se puede obtener es el dado por el ruido fotónico σ_{foton}, este depende de varias cantidades como la apertura, el rendimiento del instrumento, tiempo de exposición y la resolución espectral entre otros (Bouchy et al., 2001). Para espectros de estrellas brillantes cuya relación señal a ruido (SNR) es alta, este se reduce al ruido fotónico estelar (Figueira, 2018). El segundo tiene que ver con el instrumento en sí, la calibración, la reducción de los datos y la contaminación telúrica. Acá lo llamamos $\sigma_{sist.obs}$ y es el término que nos interesa en este trabajo. Por último está $\sigma_{estrella}$ que está relacionado con la variabilidad estelar y se origina por fenómenos magnéticos, de convección, pulsaciones, manchas, etc. que se dan en las atmósferas estelares. Si las últimas dos fuentes de error son mitigadas, se estará cerca del límite de precisión dado por el ruido fotónico.

El error relacionado con efectos instrumentales se ha ido reduciendo muchísimo generación tras generación de espectrógrafos de alta resolución. Se denomina *perfil instrumental* (IP) al perfil que representa el ensanchamiento introducido por el espectrógrafo en relación a un espectro teórico de resolución infinita emitido por una fuente (en un espectrógrafo ideal, el instrumento introduce un ensanchamiento sólo dependiente de su resolución). Todos los elementos del espectrógrafo tienen un impacto en el perfil instrumental, variaciones del IP en el tiempo tienen un efecto importante en las RVs ya que afectan el ancho y forma de las líneas, que a su vez modifican la medición de velocidad. Mantener un IP estable entonces es la clave para la estabilidad de las medidas de velocidad radial. Un paso fundamental en esta dirección se dio al empezar a desmontar los instrumentos de los telescopios, aislándolos en habitaciones controladas y alimentándolos a través de fibras ópticas que recogen la luz en el plano focal y la llevan al instrumento. Esto último se hace con el objetivo de reducir el movimiento del instrumento a cero y tener total control sobre las variables ambientales como temperatura, presión, humedad, etc. Además permite construir instrumentos de mayor tamaño y complejidad.

La calibración en longitud de onda también se mejoró notablemente al adicionar una fibra óptica con luz proveniente de una lámpara de Torio/Argón para calibración simultánea, mejorando el método de celda de absorción de Yodo usado, por ejemplo, en CORAVEL (Baranne et al., 1979). Estos grandes saltos se dieron en el diseño de ELODIE (Baranne et al., 1996) y luego se fueron mejorando cada vez más con sus sucesivas iteraciones: desde CORALIE (1998) a HARPS (Mayor et al., 2003) y SOPHIE (Perruchot et al., 2008), y el más preciso hasta el momento que es el espectrógrafo ESPRESSO (Pepe et al., 2021).

Como mencionamos, el cálculo de la RV se hace midiendo el desplazamiento Doppler de las líneas espectrales, pero ¿cuáles o cuántas líneas usar? En la práctica, se calcula una función de correlación cruzada (CCF) que es una convolución entre el espectro y una máscara que contiene las posiciones de miles de líneas espectrales para un tipo espectral similar al de la estrella. La CCF mide el parecido entre dos funciones en función del desplazamiento entre ellas, y tendrá un máximo al valor buscado de velocidad radial. Con este método se optimiza la SNR ya que se condensa la información de muchísimas líneas en una línea promedio representativa del espectro. Esta línea "promedio" es la CCF. Si se tiene un número grande de líneas, la forma de la CCF se ajusta a una Gaussiana cuyo centro dará la velocidad radial de la estrella y además los siguientes parámetros:

- FWHM: El ancho a mitad de altura (*full width at half maximum*).
- La altura de la gaussiana, que se denomina *contraste.*
- El *bisector span* (BIS), que se define como el cambio de velocidad en el bisector de la línea, y está relacionado con la asimetría de la CCF.

3.3. SOPHIE

SOPHIE es un espectrógrafo echelle alimentado por fibra óptica que se encuentra en operación desde Julio 2006 en el telescopio de 1.93 m del Observatorio de Haute Provence, Francia. Fue diseñado para reemplazar a ELODIE utilizando la experiencia y recursos adquiridos en el desarrollo de HARPS (Mayor et al., 2003), por lo que ambos comparten algunas características. La meta de diseño de SOPHIE fue la de mejorar por un factor 10 el rendimiento óptico y la precisión por un factor de 2 o 3, respecto de ELODIE.

El instrumento se encuentra instalado en la sala Coudé del edificio donde se aloja el telescopio. En la Figura 3.4 se muestra un esquema del diseño. La luz llega por cuatro fibras ópticas, dos para cada modo: alta eficiencia (HE) y alta resolución (HR). Una trae la luz de la estrella y la otra luz de cielo o de lámparas de referencia para calibración simultanea (Torio, Tungsteno, LDSL y desde hace poco un interferómetro de Fabry-Perot). Las fibras tienen una longitud de 17 m y antes de entrar al instrumento pasan por un *scrambler* (sólo en modo HR) cuyo objetivo es "mezclar" la luz de manera que sea uniforme al ingresar al instrumento y la velocidad radial derivada sea insensible a la posición de la estrella en la apertura. Las fibras tienen una rendija de salida de 40 µm para lograr la resolución deseada en el modo HR. Al salir de las fibras la luz va hacia el espejo esférico que actúa de colimador, luego

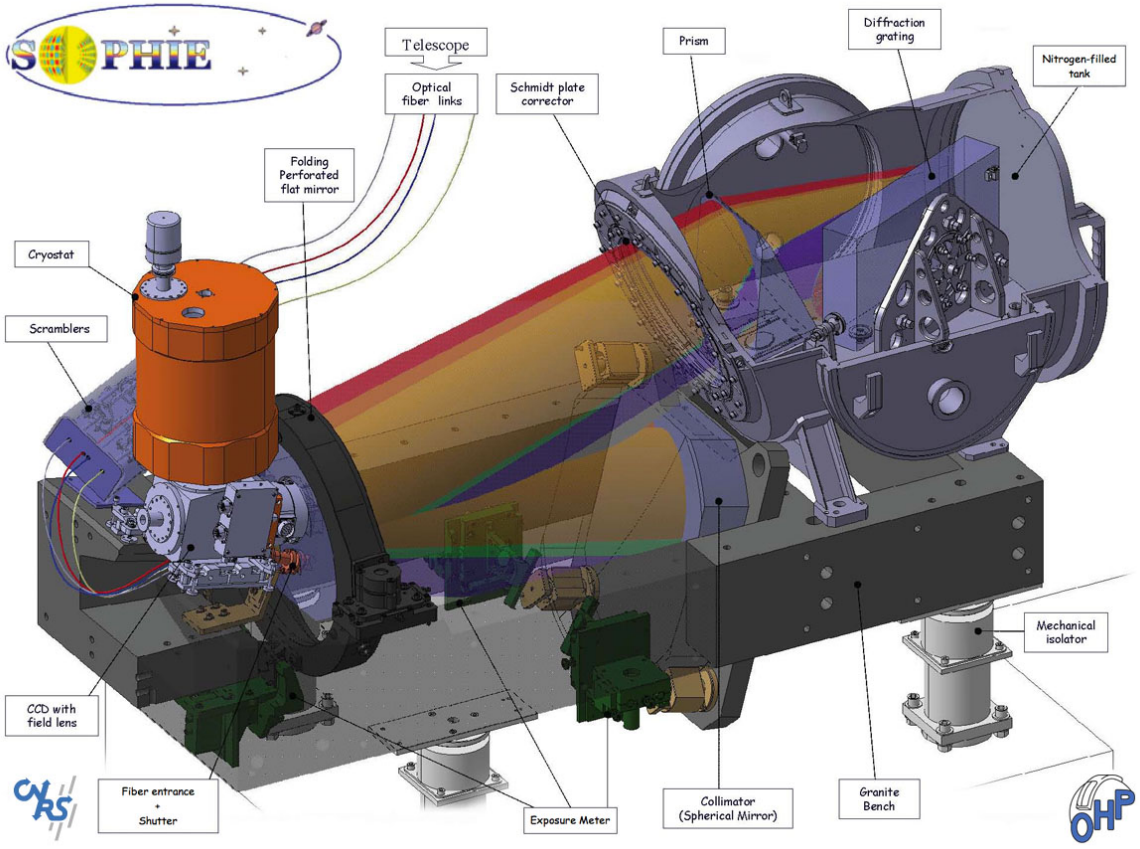

Figura 3.4. Esquema del espectrógrafo SOPHIE instalado en el telescopio de 1.93-m del Observatorio de Haute-Provence, Francia (Perruchot et al., 2008).

se refleja en el espejo plano y pasa por la cámara Schmidt hacia los elementos dispersivos. La dispersión principal se logra con la red de difracción mientras que la dispersión cruzada mediante una configuración de doble paso por el prisma. La luz vuelve a salir por la cámara Schmidt, se refleja en el espejo plano y finalmente es enfocada por el espejo esférico hacia la lente de campo por un agujero en el espejo plano, y esta forma la imagen final en el CCD. En la Tabla 3.1 se resumen las características principales del instrumento.

Todo el diseño está pensado para obtener estabilidad a largo plazo. El instrumento se apoya sobre soportes mecánicos anti-vibraciones, y la estabilidad térmica se logra mediante una triple protección: el instrumento está en una caja aislada con una estabilidad de 0.01 °C, que a su vez está dentro de una habitación a temperatura estable de 21°C, y además, las partes que potencialmente pueden introducir contaminación térmica como el CCD (enfriado a −100°C con N2), el obturador mecánico y el fotomultiplicador (que se usa para medir continuamente el flujo que entra al espectrógrafo) se encuentran fuera de la caja aislada. Los elementos dispersivos se encuentran encapsulados en un tanque a presión constante para estabilizar el índice de refracción del aire. Para referencia, una variación en la presión de 1 $mbar$ introduciría un cambio de 90 $m\ s^{-1}$ en la velocidad radial. Alrededor de 30 sensores de temperatura y presión se encuentran distribuidos por el instrumento para monitorear la estabilidad del instrumento.

La reducción de los espectros se hace de forma completamente automática con una pipeline adaptada a partir de la diseñada por el Observatorio de Geneva para HARPS. Incluye localización y extracción óptima de los órdenes, descarte de rayos cósmicos, calibración de la longitud de onda, corrección por flat y bias, división por la función de *blaze*. La correlación cruzada también se hace de forma automática para imágenes de ciencia, se calcula la CCF sobre un intervalo de $\pm 30\ km\ s^{-1}$ alrededor de la línea más profunda detectada y se ajusta un perfil Gaussiano. La velocidad radial derivada además viene con la corrección baricéntrica.

En su diseño inicial, la precisión lograda en estrellas estables estaba en el orden de los 5-6 $m\ s^{-1}$ (Perruchot et al., 2011), lo cual no es suficiente para caracterizar planetas del orden de las super-Tierras y Neptunos. La limitación principal estaba en la sensibilidad a cambios de iluminación en la pupila del instrumento causados por defectos de guiado, desenfoque, dispersión atmosférica y seeing (Perruchot et al., 2011). Para mejorar esto, en 2011 se implementó una modificación en las fibras ópticas del instrumento. Se agregaron partes de fibras de sección octogonal que presentan propiedades de scrambling muy superiores a las anteriores y los resultados fueron excelentes, llevando la precisión típica al rango de 1-2 $m\ s^{-1}$ para estrellas estandar de tipo solar. El instrumento fue renombrado como SOPHIE+ (Perruchot et al., 2011).

Desde sus inicios SOPHIE ha tenido un importante impacto en la detección de exoplanetas en el cielo Norte (ver Bouchy et al., Hébrard et al., Boisse et al., Díaz et al., Boisse et al., Moutou et al., Courcol et al., Bouchy et al., Wilson et al., Hébrard et al., Díaz et al., Rey et al., Hobson et al., Díaz et al., Hobson et al., Hara et al., Dalal et al., Demangeon et al.) como también en el seguimiento y confirmación por velocidades radiales de candidatos de surveys fotométricos como CoRot (Barge et al., 2008), SuperWASP (Collier Cameron et al.,

2007), Kepler (Santerne et al., 2011) y TESS (Figura 3.5).

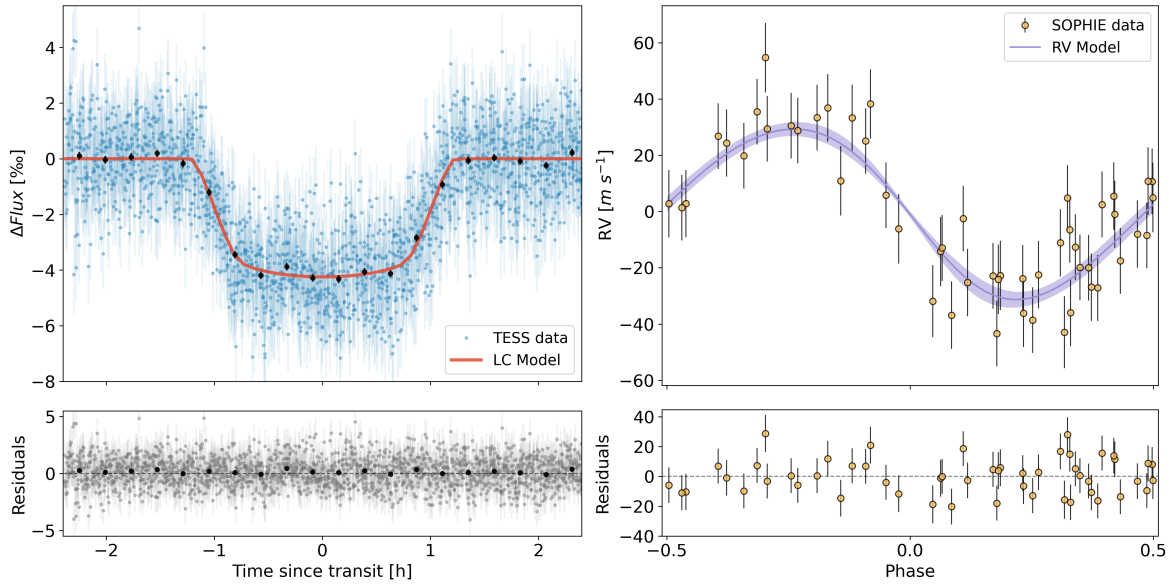

Figura 3.5. Detección y caracterización del Saturno caliente TOI-1199b con TESS y SOPHIE (Serrano Bell et al., in prep).

3.4. Corrección de punto cero de SOPHIE

A pesar de que con la mejora introducida por las fibras de sección octogonal SOPHIE alcanzó precisiones cercanas a 2 $m\ s^{-1}$ en escalas de tiempo prolongadas, se encuentra que el instrumento presenta una variación a largo plazo del punto cero de alrededor de $\sim 10\ m\ s^{-1}$ en 3.5 años (Figura 3.6, Courcol et al. (2015)). A algunos saltos o caídas en el punto cero que se ven en la Figura es posible relacionarlos con cambios puntuales que sufrió el instrumento: como la implementación de secciones adicionales de fibras octogonales en 2012, el aluminizado del espejo secundario (que provocó un cambio en el balance de flujo a través del rango espectral), el deterioro en el tiempo de la lámpara de Torio-Argón, y también variaciones de escala del orden de algunas semanas que correlacionan con cambios bruscos de la temperatura exterior que se propagan en el orden de décimas de grados a través del pilar del telescopio hacia el trípode del instrumento (Courcol et al., 2015).

Para mantener un registro y poder corregir la deriva del punto cero se monitorean un conjunto de cuatro estrellas estándares de RV (esto es, estrellas brillantes y de velocidad radial constante): HD 185144, HD 9407, HD 221354 y HD 89269. Se eligieron estas porque anteriormente habían sido monitoreadas con HIRES en el telescopio Keck de 10.2 m mostrando dispersiones muy bajas de 2.0, 1.7, 1.9 y 2.0 $m\ s^{-1}$, respectivamente, a lo largo de varios años. Además de estas super-constantes, en el procedimiento definido por Courcol et al. para corregir la deriva se tomaron 51 estrellas con baja dispersión en RVs de los programas de SOPHIE. A estas primero se les sustrajo una corrección hecha con las 4 iniciales y luego se tomó las que mostraron una dispersión menor a 3 $m\ s^{-1}$ para construir una nueva corrección

SOPHIE	
Parámetro instrumental	Valor
Campo	3 arcsec
Resolución	39000 en modo HE
	75000 en modo HR
Resolución intrínseca	200000
Rango espectral	387-694 nm
Número de ordenes	39
Detector	1 CCD 2048x4102 (61x31 mm)
	Tamaño de pizel 14 µm
SNR por pixel a 550 nm	SNR = 100 en 1 hora para m=11 (HE)
	SNR = 100 en 1 hora para m=10 (HR)
Red de difracción	Red R2 (ángulo de blaze $\theta = 65°$)
Diámetro de pupila	200 mm
Diámetro de espejo esférico	540 mm
Diámetro de espejo plano	440 mm
Placa correctora Schmidt	Diametro 320 mm
	Espesor 25 mm
	Perfil Kerber
Prisma (dispersor cruzado)	angulo 31°
	280x220 mm, 30 kg
Lente de campo	plano-convexa, diametro 90 mm

Tabla 3.1. Parámetros y características instrumentales del espectrógrafo SOPHIE.

agregando éstas. Así recursivamente hasta que ya no se pudieran agregar más estrellas. Al converger el proceso, se terminaron agregando 19 estrellas (todas con al menos 10 mediciones). Con este set de 23 estrellas se construyó una serie temporal de constantes de RV que representa una medida empírica del corrimiento del punto cero en el tiempo. A esta serie se le hace un filtrado promediando el valor de la corrección cada 15 medidas, de manera que la resolución temporal de la corrección depende de la frecuencia en la observación de constantes de RV. Con este paso, la escala temporal típica de la corrección es de 9 días. Por último, se hace un interpolado de la serie para poder aplicar la corrección a cualquier tiempo dentro del rango de la serie maestra. Para tener una idea del efecto de esta corrección, al aplicarla a la estándar HD 185144 se reduce la media cuadrática (o RMS, del inglés *Root Mean Square*) de 5.37 $m\ s^{-1}$a 1.61 $m\ s^{-1}$. En la Figura 3.7 se muestra la distribución de RMS en las 55 estrellas iniciales antes y después de la corrección de punto cero.

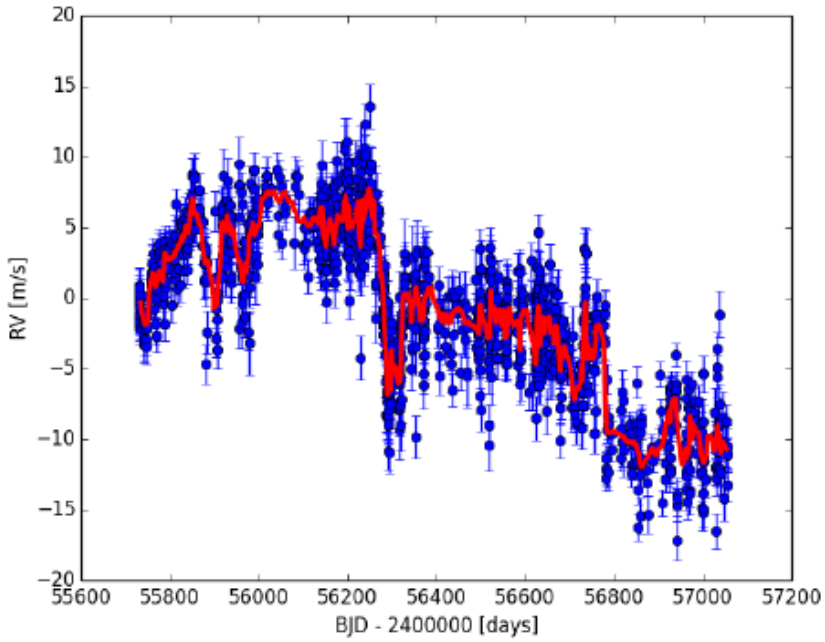

Figura 3.6. La línea roja marca la deriva del punto cero entre 2011 y 2015 obtenida a partir de promediar medidas de velocidades radiales de estrellas constantes (puntos azules). Courcol et al. (2015).

Este procedimiento y variaciones posteriores hoy en día es el estándar para la corrección del punto cero de SOPHIE en numerosos trabajos (ver Hobson et al. (2018), Hara et al. (2020), Heidari et al. (2022)).

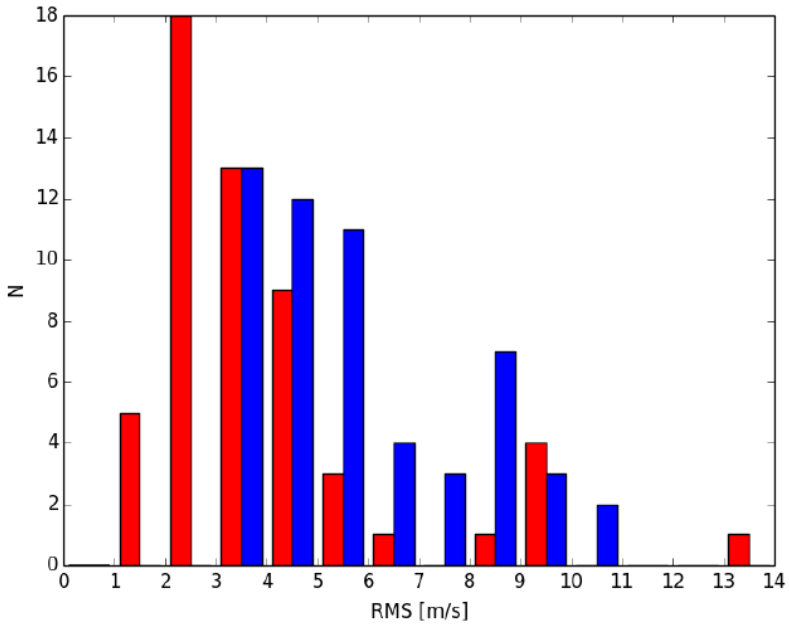

Figura 3.7. RMS iniciales (azul) y finales después de la corrección (rojo) de la muestra de 55 estrellas. Courcol et al. (2015).

Capítulo 4

Aprendizaje Automático

El término "aprendizaje automático" del inglés *machine learning* fue acuñado en 1959 por Arthur Lee Samuel, un pionero del campo de la inteligencia artificial y las ciencias de la computación. Hoy en día, es el nombre que lleva el campo de estudio de sistemas con la habilidad de *aprender* sin seguir instrucciones explícitas, a través de algoritmos diseñados para encontrar patrones en conjuntos de datos. Hay numerosos tipos de sistemas de aprendizaje automático que se pueden clasificar según ciertas categorías: por la forma de entrenamiento del algoritmo, en supervisado, no supervisado, semi-supervisado o aprendizaje por refuerzo; además el entrenamiento puede ser *en lote* (batch learning), es decir que aprende una sola vez sobre un conjunto grande de datos o *en linea* (online learning) donde el algoritmo va aprendiendo incrementalmente con nuevos datos. Y otra forma de categorización tiene que ver con la manera en que hacen las predicciones, si el algoritmo es basado en instancias o en modelos. Para ahondar en más detalles sobre estas clasificaciones se recomienda la siguiente bibliografía: Bishop (2006), Géron (2017), Müller & Guido (2016).

En un algoritmo supervisado, el aprendizaje se realiza sobre datos *etiquetados*. Es decir, se le da al algoritmo un conjunto de pares entrada-salida, donde la entrada son un conjunto de variables predictoras (o *features*) y la salida (etiquetas) son los valores a predecir, de manera que le estamos dando el valor real de la variable a predecir. De ahí el nombre supervisado, ya que estos algoritmos aprenden a minimizar una función de pérdida (o *loss function*) que mide qué tanto se equivocan en las predicciones. Por el contrario, en los algoritmos no supervisados sólo se busca encontrar patrones en datos *no etiquetados*.

Los algoritmos supervisados, que son los que nos interesan en este trabajo, pueden realizar dos tipos de tareas: clasificación o regresión. En la primera, se quiere determinar si algo pertenece o no a una dada clase (puede ser una o muchas) a partir de variables de entrada relacionadas con las características del objeto. Mientras que en la segunda, se quiere predecir un valor numérico continuo. Un ejemplo clásico de clasificación y que sirve para entender un poco mejor qué es aprendizaje automático supervisado es el reconocimiento de caras en una imagen. Hoy, cualquier cámara de un teléfono inteligente tiene la capacidad de reconocer caras, pero este problema es muy difícil de resolver sin algoritmos de aprendizaje automático. En un enfoque tradicional se debería programar a mano una lista extremadamente compleja

de reglas que tiene que cumplir una imagen digital para contener una cara, en cambio, si entrenamos un algoritmo de aprendizaje automático sobre un gran número de imágenes *que sabemos* que contienen caras, éste determinará de manera autónoma qué combinación de características son las necesarias para identificar una cara. Una vez entrenado, el modelo se puede usar para clasificar imágenes nuevas. El análogo para regresión es cuando conocemos el valor numérico de la variable que queremos predecir para un número de instancias de entrenamiento, por lo tanto el algoritmo puede aprender las relaciones entre las variables predictoras y la etiqueta (o *target*) durante el entrenamiento.

Un abordaje con aprendizaje automático es especialmente útil para problemas cuyos algoritmos requieren una larga lista de reglas y condiciones, o problemas donde un enfoque tradicional no tiene una buena solución. También para ambientes fluctuantes donde tiene la ventaja de ser adaptable en el tiempo, y cuando se quiere extraer información a partir de grandes cantidades de datos (minería de datos). En muchos algoritmos de aprendizaje automático se puede cuantificar la utilidad de cada variable predictora en el modelo, es decir que podemos aprender qué variables son más relevantes o contribuyen más en la predicción.

En la Figura 4.1 se muestra el flujo de trabajo para desarrollar un modelo con aprendizaje automático supervisado.

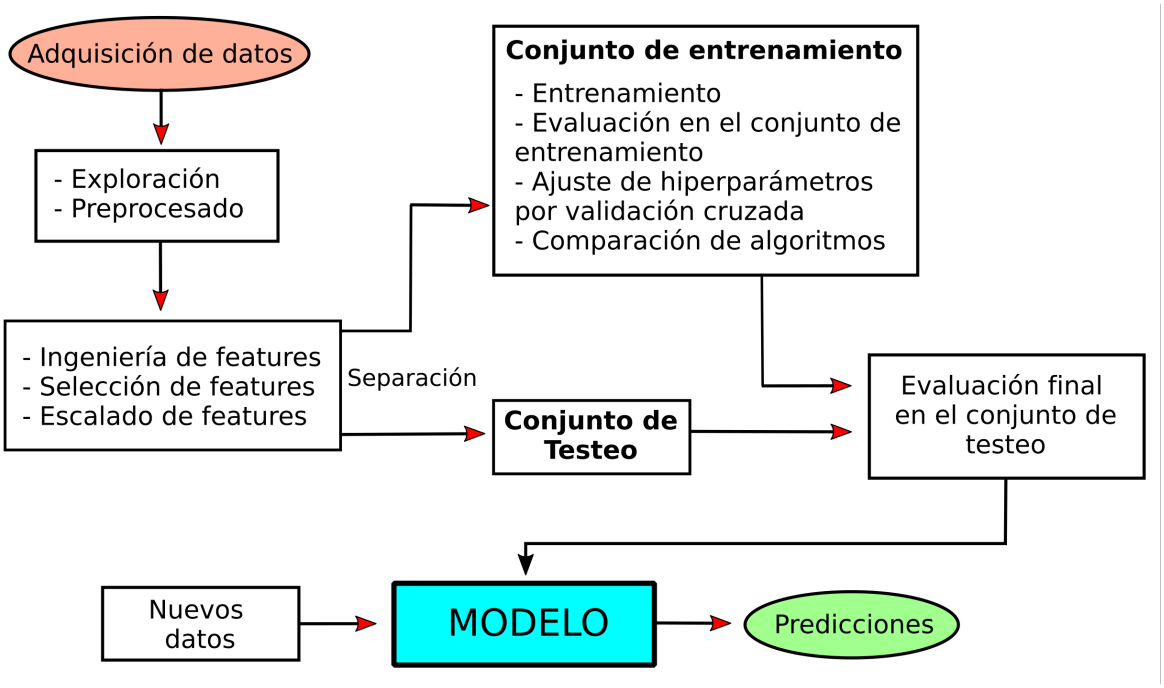

Figura 4.1. Flujo de trabajo usual en el desarrollo de un modelo de aprendizaje automático.

4.1. Métricas de desempeño para regresión

Antes de comenzar a entrenar y comparar algoritmos se debe definir con qué métricas se va a cuantificar el desempeño de un dado modelo en la tarea de predecir la variable de interés. Vamos a describir las dos que usamos durante este trabajo que son la Raíz del Error

Medio Cuadrático Pesado o WRMSE (Weighted Root Mean Squared Error) y el coeficiente de determinación R^2.

Se define el WRMSE de la siguiente forma:

$$\text{WRMSE} = \sum_{i=0}^{n-1} \frac{w_i(y_i - \hat{y}_i)^2}{wn} \tag{4.1}$$

donde n es el número de muestras, $\hat{y}_i$ es la predicción i-ésima del modelo para la muestra i e y_i es el valor verdadero correspondiente, w_i es el *peso* del dato i y w es la suma de todos los pesos.

Por otra parte, el coeficiente de determinación R^2 se define como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.2}$$

donde nuevamente $\hat{y}_i$ es la predicción i-ésima del modelo para la muestra i e y_i es el valor verdadero, mientras que $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. El mejor valor posible es 1 y se da en el caso en que el modelo ajusta perfectamente todos los datos. Mientras que un modelo que predice siempre el valor medio de y tendrá un R^2 de 0. Para valores intermedios se puede interpretar el coeficiente como el porcentaje de la varianza de y que ha sido explicada por las variables independientes del modelo.

4.2. Entrenamiento, validación y testeo

El conjunto de entrenamiento es aquella porción de datos que usaremos para entrenar el algoritmo. Luego del entrenamiento, nos interesa probarlo en algún conjunto de datos nuevos para saber qué tan bien aprendió la tarea para la cual entrenamos. Aquí es donde cabe distinguir entre dos conjuntos, el de validación y el de testeo, ya que tienen objetivos distintos. Con testeo, nos referimos a someter un modelo entrenado a predecir datos *nuevos*, es decir, con instancias de las variables predictoras que no hayan sido usadas en el entrenamiento y para las cuales conocemos el valor verdadero de la variable objetivo. De esta manera, se pueden comparar las métricas de la predicción obtenidas durante el entrenamiento con las obtenidas en el conjunto de testeo. Esta comparación nos permite determinar que tan bien generaliza el modelo, si el desempeño es parecido en ambos conjuntos entonces tenemos un modelo con buena generalización. Por el contrario, si las métricas de entrenamiento son muy superiores a las de testeo, es muy posible que nuestro modelo esté sobreajustado a los datos de entrenamiento y no reproduzca bien la relación real entre las variables predictoras y la variable objetivo. Ya que el sobreajuste hace que los modelos reproduzcan el *ruido* de los datos de entrenamiento.

Por otra parte, el conjunto de validación tiene como objetivo ayudarnos a elegir el mejor modelo posible antes de realizar el testeo. Supongamos la siguiente situación: entrenamos un modelo que funciona muy bien en el entrenamiento, lo aplicamos al conjunto de testeo y no tenemos buenos resultados. Volvemos atrás y modificamos los parámetros del modelo, ahora lo testeamos y obtenemos un resultado un poco mejor en el test. Seguimos así

hasta obtener un resultado satisfactorio. Si hiciéramos esto, estaríamos entonces metiendo información del conjunto de testeo en nuestro modelo, es decir estaríamos sobreajustando al conjunto de testeo, que va a depender de la separación particular de datos que hicimos entre entrenamiento-testeo. Esto no es una buena práctica. En cambio, una buena forma de encontrar los mejores parámetros para un modelo es separar otro conjunto más dentro de los datos de entrenamiento, a este conjunto se lo llama de *validación.* La desventaja es que reducimos aún más la cantidad de datos que tenemos para entrenamiento, y nuevamente la separación particular que hagamos para el conjunto de validación será influyente en el resultado final. Sin embargo, hay un método con el cuál se puede contrarrestar estas dos cosas, es el de *validación cruzada.* En la validación cruzada se separa el conjunto de entrenamiento en K subconjuntos de igual tamaño, luego se hacen K entrenamientos y evaluaciones usando en cada iteración un conjunto distinto para validación. El puntaje (valor de la métrica elegida) del modelo se obtiene promediando los K puntajes de la validación cruzada. Luego se re-entrena el mejor modelo usando el conjunto completo de entrenamiento. Se ilustra este proceso en la Figura 4.2.

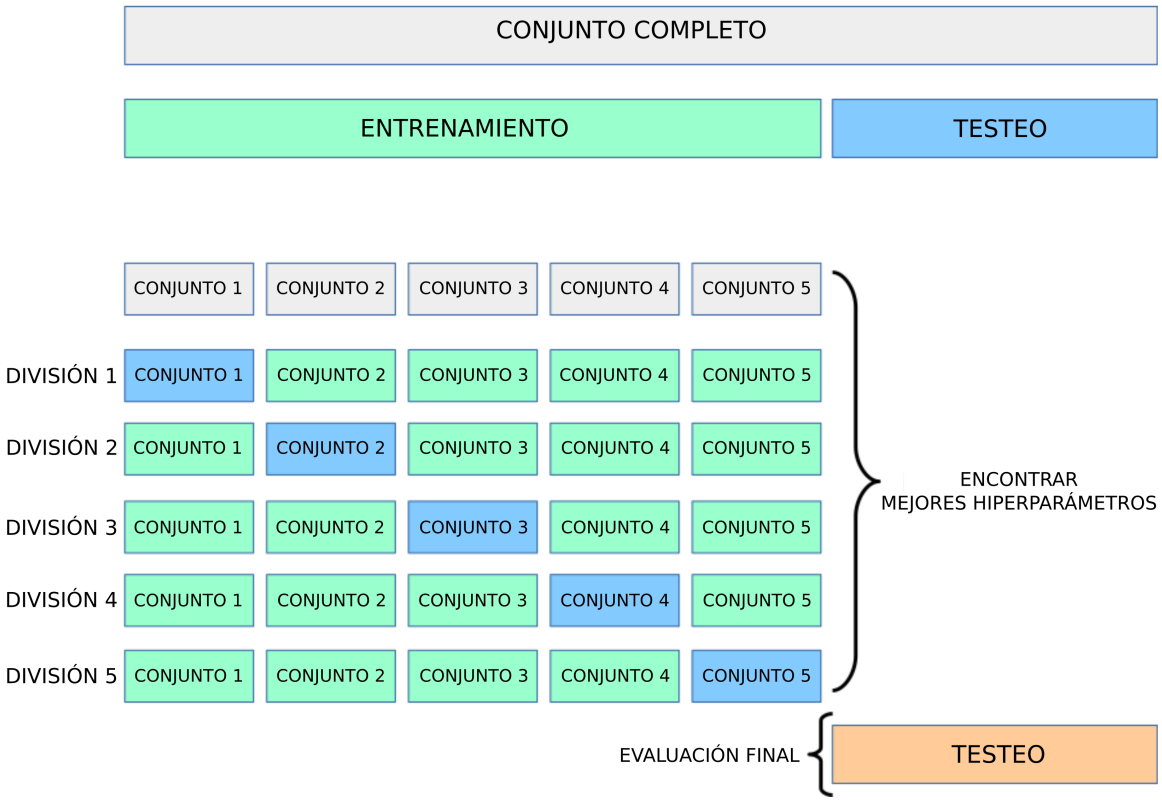

Figura 4.2. Esquema del método de validación cruzada para el caso de $K = 5$.

4.3. Algoritmos

Hay numerosos algoritmos de aprendizaje automático para regresión. Una parte importante en la implementación de un modelo de aprendizaje automático tiene que ver con la elección del algoritmo. Generalmente a priori no es claro cuál puede dar los mejores resultados. Por lo tanto la elección implica probar varios y comparar su rendimiento. Durante este trabajo se utilizaron los siguientes algoritmos: Regresión Lineal, Regresión Ridge, Regresión LassoLars, SVM (*Support Vector Machines*), Árboles de decisión, AdaBoost (*adaptive boosting*), XGB (*extreme gradient boosting*), Random Forest, Extra Trees (*extremely randomized trees*) y Gradient Boosting Machines. Detallaremos en las siguientes secciones dos de los algoritmos que fueron más importantes en nuestro trabajo, LassoLars y Gradient Boosting. Para más detalle de la implementación y funcionamiento de los demás algoritmos se recomienda la bibliografía citada en la introducción al Cap. 4 y la documentación del paquete `scikit-learn`[(i)] de Python.

4.3.1. Gradient Boosting Machines

Las *Gradient Boosting Machines* (GBMs) forman parte del conjunto de métodos de ensamble *(ensemble learning)*, que a diferencia de otros algoritmos en los que se construye un solo modelo poderoso, utilizan una combinación de un gran número de modelos débiles o "base", que adquieren en conjunto un excelente poder de predicción. Algunos algoritmos de ensamble simplemente promedian las predicciones de muchos modelos, pero en las GBMs la combinación es distinta. Los modelos-base se van agregando secuencialmente, la idea es que cada nuevo modelo-base se entrena con respecto al error total del ensamble hasta el paso anterior. Más precisamente, el nuevo modelo-base se construye de manera que tenga la máxima correlación con el gradiente negativo de la función de pérdida. La elección del modelo-base y de la función de pérdida son arbitrarias, por lo que las GBMs son muy flexibles. Si se elige como función de pérdida el error cuadrático medio, el procedimiento implica un ajuste sucesivo del error. Los GBMs son ampliamente usados en la industria y suelen tener un gran desempeño en las competencias de aprendizaje automático (Müller & Guido, 2016).

4.3.1.1. Gradient Boosted Regression Trees

Cuando los modelos base son árboles de decisión (ver apéndice C), se les llama *Gradient Boosted Regression Trees* (GBRT). Un GBRT es un modelo aditivo que da una predicción $\hat{y}_i$ para una dada entrada x_i con la siguiente expresión:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i) \tag{4.3}$$

donde los h_m son los modelos base: árboles de decisión con una profundidad fija. La constante M representa el número de estimadores base o iteraciones. El modelo se va construyendo de

[(i)] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

forma aditiva:

$$F_m(x) = F_{m-1}(x) + h_m(x) \tag{4.4}$$

donde cada nuevo estimador $h_m(x)$ se ajusta para minimizar la pérdida del ensamble en el paso previo F_{m-1}:

$$h_m = \arg\min_h L_m = \arg\min_h \sum_{i-1}^{n} l(y_i, F_{m-1}(x_i) + h(x_i)) \tag{4.5}$$

donde $l(y_i, F(x_i))$ es la función de pérdida, que debe ser una función diferenciable. De manera que usando una aproximación de Taylor, el valor de l se puede aproximar por:

$$l(y_i, F_{m-1}(x_i) + h_m(x_i)) \approx l(y_i, F_{m-1}(x_i)) + h_m(x_i) \left[\frac{\partial l(y_i, F(x_i))}{\partial F(x_i)} \right]_{F=F_{m-1}} \tag{4.6}$$

El término $\left[\frac{\partial l(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}$ es la derivada de la función de pérdida con respecto a su segundo parámetro, evaluada en $F_{m-1}(x)$. La cual es fácil de obtener para cualquier $F_{m-1}(x_i)$ dado que l es diferenciable. Denotándola con g_i, y deshaciéndonos de los términos constantes, nos queda:

$$h_m \approx \arg\min_h \sum_{i=1}^{n} h(x_i) g_i \tag{4.7}$$

A cada iteración, para minimizar esto se ajusta el nuevo estimador h_m para predecir un valor que sea proporcional al gradiente negativo $-g_i$. Los gradientes se actualizan a cada iteración. Este procedimiento puede ser considerado como una forma de descenso por gradiente en un espacio de funciones.

4.3.1.2. Implementación

En la práctica, utilizamos la implementación de `scikit-learn`[(ii)]. Esta implementación soporta cuatro tipos de funciones de pérdida mediante el hiperparámetro `loss` (los *hiperparámetros* en el contexto de aprendizaje automático, son todos los parámetros cuyos valores controlan el proceso de aprendizaje). El número de modelos base o *estimadores* M del ensamble se define con "`n_estimators`". Otro hiperparámetro de gran importancia es la *tasa de aprendizaje* ("`learning_rate`" en el método) que regula la contribución de cada árbol al ensamble, multiplicando a cada árbol por un factor constante, cuyo valor óptimo debe obtenerse mediante validación cruzada. Un valor pequeño de la tasa de aprendizaje hará que se necesiten muchos más estimadores base para ajustar el ensamble al conjunto de entrenamiento, esta técnica denominada *shrinkage* produce usualmente una mejor generalización del modelo (es decir, previene el sobreajuste a los datos de entrenamiento). Estos son sólo algunos de los muchos hiperparámentros que permite modificar el algoritmo, la búsqueda de la mejor combinación de hiperparámetros para un conjunto de datos y una tarea en particular conforma una parte importante del aprendizaje automático. En la Figura 4.3 se muestra una secuencia de aprendizaje de un GBDT con un solo predictor (x_1) y tres modelos base.

[(ii)] https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting

Algo a destacar de esta implementación es que asigna un coeficiente de *importancia* a cada variable predictora, que es proporcional a la reducción total de la pérdida obtenida mediante esa variable particular. La lista de coeficientes se puede obtener mediante el atributo "`feature_importances_`". Es muy útil para la interpretabilidad del modelo y además sirve como herramienta para seleccionar las variables más relevantes en un conjunto de entrenamiento.

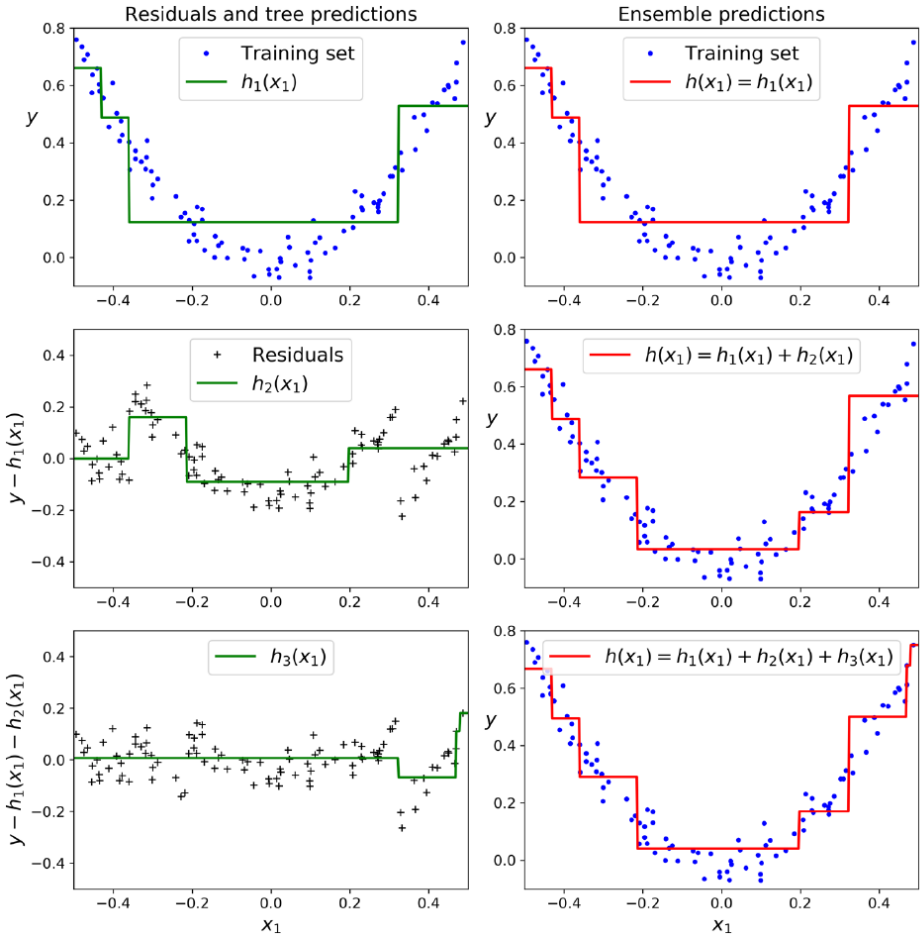

Figura 4.3. Secuencia de aprendizaje de Gradient Boosting. El primer predictor (arriba a la izquierda) se entrena normalmente, luego cada predictor consecutivo (columna izquierda) se entrena sobre los residuos del predictor previo. La columna derecha muestra como va quedando el ensamble de los predictores individuales. Fuente: *Hands on Machine Learning with Scikit-Learn, Keras & TensorFlow.* Aurelién Geron.

4.3.2. Lasso

Lasso (*Least Absolute Shrinkage and Selection Operator*) es un método de regresión lineal que utiliza regularización y selección automática de variables. La regularización es una técnica comúnmente usada para evitar el sobreajuste, controlando la complejidad de un modelo para reducir la varianza del mismo. Para esto, el método de Lasso agrega un término de

penalización a la función de pérdida de mínimos cuadrados:

$$\min_{w} \frac{1}{2n}||Xw - y||_2^2 + \alpha||w||_1 \tag{4.8}$$

Aquí n es el número de muestras, X es la matríz de variables independientes, y es el vector de la variable dependiente, $||w||_1$ es la norma L1 del vector de coeficientes y α es una constante que controla el término de penalización y que constituye un hiperparámetro del modelo a determinar (por validación cruzada por ejemplo). Al sumar la norma L1 del vector de coeficientes (regularización L1) el término de penalización previene que haya coeficientes muy grandes en el modelo, de hecho, hace que los coeficientes estén cerca del cero. Con lo cual un número de coeficientes serán efectivamente nulos (ver Figura 4.4), y de esa forma el método de Lasso construye modelos más simples e interpretables, funcionando como un selector de las variables predictoras más relevantes. Análogamente a este método, si se usa la norma L2 en el término de penalización, el método correspondiente se llama *Ridge*.

Según el algoritmo que se use para ajustar los coeficientes del método de Lasso se da lugar a distintas implementaciones. En la sección siguiente describiremos el algoritmo LARS que es el que usamos en este trabajo.

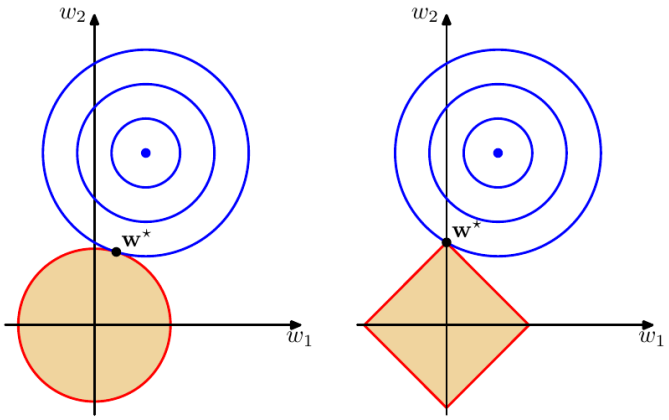

Figura 4.4. En azul se muestran curvas de nivel de la función de error sin regularizar, mientras que en rojo se ve a la derecha la region de restricción de la regularización L2 (ridge) y a la izquierda la correspondiente a la regularización L1 (lasso). El valor óptimo para el vector de coeficientes se denota con w^*. Lasso da una solución en la cual $w_1 = 0$. Fuente: *Pattern recognition and machine learning.* Bishop (2006).

4.3.3. LARS

El algoritmo LARS (*Least Angle Regression*) provee una forma de estimar, dado un conjunto de variables predictoras y una variable a predecir, cual es el mejor subgrupo de variables y sus respectivos coeficientes para lograr el mejor ajuste lineal. Usualmente se usa cuando se tienen conjuntos de alta dimensionalidad. El procedimiento básico del algoritmo es:

1. Empieza con todos los coeficientes igual a cero $w = 0$
2. Busca la variable predictora x_i más correlacionada con la variable objetivo y.

3. Incrementa el coeficiente w_j en la dirección del signo de su correlación con y. Calcula los residuos $r = y - \hat{y}$ en el camino. Frena cuando otro predictor x_k tiene la misma correlación con r que tiene x_j.

4. Incrementa (w_j, w_k) en la dirección *equiangular* (de aquí el nombre del algoritmo) entre los dos predictores, hasta que otro predictor x_m tenga la misma correlación con el residuo r a lo largo del camino.

5. Ahora incrementa (w_j, w_k, w_m) en la dirección equiangular conjunta hasta que otra variable x_n tenga la misma correlación con r.

6. Continúa hasta que todos los predictores estén en el modelo.

El resultado de LARS da la curva que denota la solución para cada valor de la norma L1 del parámetro de vectores (para más detalle, ver Efron et al. (2004)).

Capítulo 5

Planteo del problema y preparación de los datos

El problema que queremos abordar aquí es predecir el efecto sistemático en las velocidades radiales de SOPHIE que se manifiesta como una deriva a largo plazo del punto cero. Nuestra hipótesis es que podemos modelarlo a partir de alguna combinación de variables relacionadas a las condiciones observacionales, ambientales e instrumentales en el momento en que se hace la exposición. Construir un modelo físico explícito de este fenómeno resultaría imposible por la gran cantidad de variables que interfieren y cuya interacción con el efecto es a priori desconocida. En cambio, el enfoque con aprendizaje automático nos da la ventaja de no tener que definir una relación funcional entre las decenas de variables y la deriva del punto cero. Por otra parte, gracias a que la corrección se ha hecho durante muchos años mediante observaciones de estrellas constantes, contamos con un buen número de estas en la base de datos de SOPHIE. Además, la pipeline de reducción almacena muchísima información en los headers de las imágenes que nos pueden servir en principio como posibles predictores, y contamos con una treintena de sensores de temperaturas y presiones ubicados alrededor y en el instrumento cuyas medidas también son accesibles para cada observación. Todo esto nos permite armar un buen conjunto de datos para entrenar los algoritmos.

Este es un problema de regresión, los predictores a usar serán en principio todo tipo de variables que creamos que pueden llegar a tener algún efecto en la medición final de la RV. Esto es: variables observacionales (por ejemplo seeing, brillo del cielo, masa de aire, etc), ambientales (como temperatura exterior, humedad, presión atmosférica) e instrumentales (temperaturas y presiones en el instrumento, SNR de cada orden de dispersión, parámetros de la CCF, etc).

El entrenamiento es supervisado, es decir que se entrena el algoritmo en un conjunto de pares entrada-salida, donde la entrada son los predictores y la salida es la variable a predecir o *target*. La idea es usar como target todas las medidas de RV que tengamos de las cuatro super-constantes del trabajo de Courcol et al. (2015). Elegimos estas porque se han monitoreado regularmente por muchos años y la extrema estabilidad que tienen nos permite suponer que su variabilidad está directamente relacionada a los efectos sistemáticos del instrumento.

5.1. Adquisición

Para entrenar un modelo de aprendizaje automático supervisado se necesita un conjunto ordenado y tabulado de datos. Cada fila representa un dato, y las columnas son todas las variables predictoras y una correspondiente a la variable a predecir (el *target*). En líneas generales, cuantos más datos tengamos para entrenar se esperan mejores resultados, por lo cual se trata de recolectar la mayor cantidad posible de observaciones que sirvan a nuestro fin.

Un punto muy importante a tener en cuenta al construir un conjunto de entrenamiento tiene que ver con la uniformidad en la producción de los datos, y dado que los últimas modificaciones en el instrumento se hicieron en diciembre de 2017, restringimos el rango temporal del conjunto de entrenamiento a los años 2018, 2019 y 2020. Lo primero que se hizo entonces fue una búsqueda en la base de datos en línea de SOPHIE[(i)] de todas las observaciones disponibles durante el período del 1 de Enero de 2018 al 31 de Diciembre de 2020 de las cuatro super-constantes, a saber: HD 185144, HD 89269, HD 9407 y HD 221354. En esta búsqueda constatamos que HD 221354 solo contaba con tres observaciones en ese período, por lo cual la tuvimos que descartar.

5.1.1. Datos de los headers

Se extrajo información de dos tipos de archivos generados por el pipeline de reducción de SOPHIE: los "e2ds", que contienen el espectro bidimensional, orden por orden, corregido por bias, flat-field y rayos cósmicos, junto con información sobre la observación, y los "ccf", que contienen los datos del análisis de correlación cruzada. Examinando los encabezados de estos archivos se hizo una preselección de los parámetros que pudieran llegar a servir como posibles variables predictoras y posteriormente se escribió un script que recorre la totalidad de la base de SOPHIE y que extrae 77 datos seleccionados de los encabezados de los archivos ccf y e2ds. Se corrió el script de forma remota en el servidor de SOPHIE obteniendo las tablas de datos en crudo. En el Apéndice A.1 se detalla la lista de variables extraídas de los headers de las imágenes y una descripción breve de cada una. Luego de filtrar sólo las observaciones correspondientes a modo HR (alta resolución), la cantidad de datos por estrella que obtuvimos fueron: 383 de HD 185144, 174 de HD 89269 y 105 de HD 9407. Por tanto, el conjunto inicial contó con 662 observaciones.

5.1.2. Datos de sensores externos

SOPHIE cuenta con 30 sensores externos que toman medidas de manera continua cada seis minutos y se van almacenando en archivos de texto cada $24hs$. De estos, 26 son de temperaturas, 2 de presiones, 1 de humedad y 1 de corriente. En el Apéndice A se muestran todas las variables medidas por los sensores. Para obtener los datos correspondientes a cada una de las observaciones se escribieron dos scripts, el primero para unificar todos los archivos

[(i)] http://atlas.obs-hp.fr/sophie/

correspondientes al período del 1 de Enero de 2018 al 31 de Diciembre de 2020 en una sola tabla con una columna indicando la hora exacta de cada medida. Con el segundo se hizo la tarea de asignar a cada observación el conjunto de medidas de los sensores que estuviera más cercano al tiempo de comienzo de la exposición. En promedio, la diferencia de tiempo entre el inicio de la observación y las medidas de los sensores es de ~ 2.5 minutos.

5.2. Preprocesado

5.2.1. Valores atípicos

Al juntar los datos descriptos en las dos subsecciones anteriores, obtuvimos una tabla por cada una de las tres estrellas con un total de 107 columnas, donde una corresponde a la velocidad radial medida y las demás a todas las variables predictoras iniciales. Exploramos primero los valores de las RVs de las tres estrellas. En la Figura 5.1 se grafican las velocidades luego de sustraer la media para cada conjunto, se observa que hay algunos valores atípicos en las tres estrellas, probablemente debido a irregularidades en las observaciones o asociados a observaciones con muy baja SNR. Para remover los valores atípicos adoptamos un mismo criterio para los tres conjuntos, dejando afuera los valores con un valor absoluto mayor a 20 $m\ s^{-1}$. Con esto se removieron 17 puntos, quedando el conjunto con 645 observaciones. Una vez corregidas, las series temporales tienen una dispersión de 3.20 $m\ s^{-1}$ para HD 185144, 5.09 $m\ s^{-1}$ para HD 89269 y 4.02 $m\ s^{-1}$ en el caso de HD 9407.

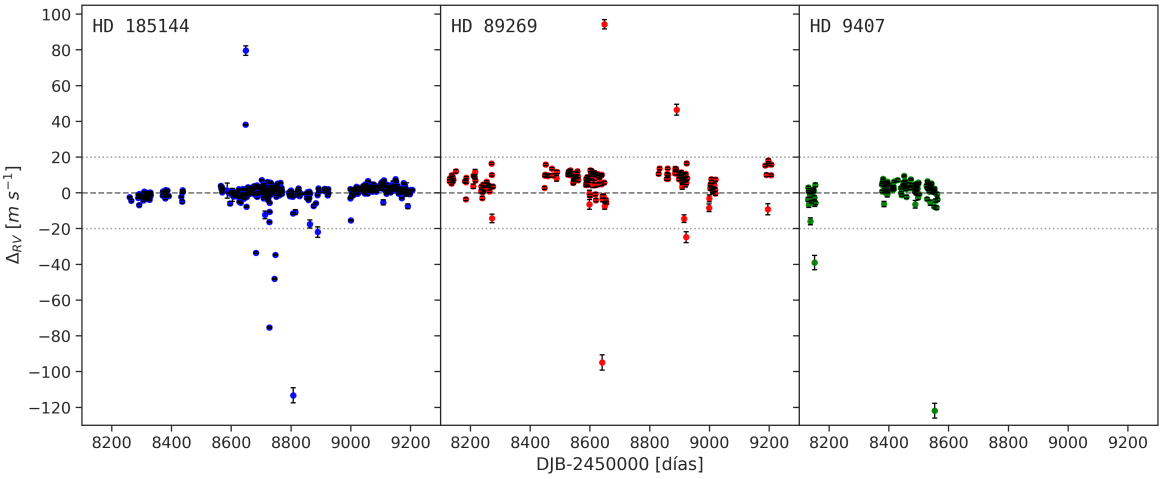

Figura 5.1. Datos iniciales de velocidades radiales, se aprecia la presencia de valores atípicos. Las líneas punteadas representan una desviación de $\pm$ 20 $m\ s^{-1}$ del cero.

5.2.2. Limpieza

Lo siguiente que se hizo fue inspeccionar la calidad de los datos recolectados en las variables predictoras. Es esperable que en un conjunto tan grande de datos producidos de forma heterogénea y con distintos objetivos haya numerosas irregularidades, como valores nulos,

faltantes, atípicos, o incluso variables constantes que no aportan nada. Todas estas características afectan negativamente a los algoritmos por lo que es necesario hacer una limpieza antes de proceder al entrenamiento. Con este objetivo graficamos histogramas para todas las variables. Inmediatamente tuvimos que descartar 10 variables que tenían un porcentaje muy alto de valores nulos ($> 20\,\%$), que se indican en el apéndice A.1. Para las variables que tenían menos de 5 valores faltantes, se completaron los registros faltantes con la mediana de las otras observaciones. En el Apéndice A.2 se muestran los histogramas luego de la limpieza de todas las variables que quedaron, para el caso de la estrella HD 185144. El mismo proceso se aplicó a las otras dos. Luego de la limpieza, el conjunto quedó con 96 variables predictoras.

5.2.3. Correlaciones

A continuación, estudiamos las correlaciones entre todos los pares de variables predictoras. La colinealidad (dependencia lineal entre dos variables) y multicolinealidad (entre muchas variables) provoca que los coeficientes en una regresión lineal múltiple cambien erráticamente frente a pequeños cambios en los datos de entrada. Esto hace que la determinación de los coeficientes de la regresión sea imprecisa, que a su vez se traduce en mayor incerteza en las predicciones de los modelos para datos nuevos. Se utiliza el coeficiente de Pearson para cuantificar la dependencia lineal entre dos variables x e y:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{5.1}$$

cuyos valores van entre -1 y 1. Un valor absoluto igual a 1 indica que una recta ajusta perfectamente la relación entre ambas variables, siendo igual a 1 si la relación es directa y -1 si es inversa. Mientras que un valor de 0 significa que no hay dependencia lineal entre ellas. En la Figura 5.2 se muestra una matriz de correlación entre todas las variables. Lo primero que salta a la vista en la esquina superior izquierda es que los SNR de los 39 órdenes de dispersión están altamente correlacionados. Algo similar pero no tan marcado se ve con las columnas de las variables de los sensores externos en la esquina inferior derecha. También los pares de variables que expresan el máximo y mínimo de una misma medida durante la observación están fuertemente correlacionados como es de esperarse, por ejemplo temp1_min y temp1_max.

Por otro lado, nos interesa indagar qué variables son las que están más correlacionadas con la velocidad radial y por lo tanto serán importantes para la predicción. De mayor a menor en valor absoluto, las primeras 15 son las siguientes (para esta lista dejamos una sola de las 39 SNR):

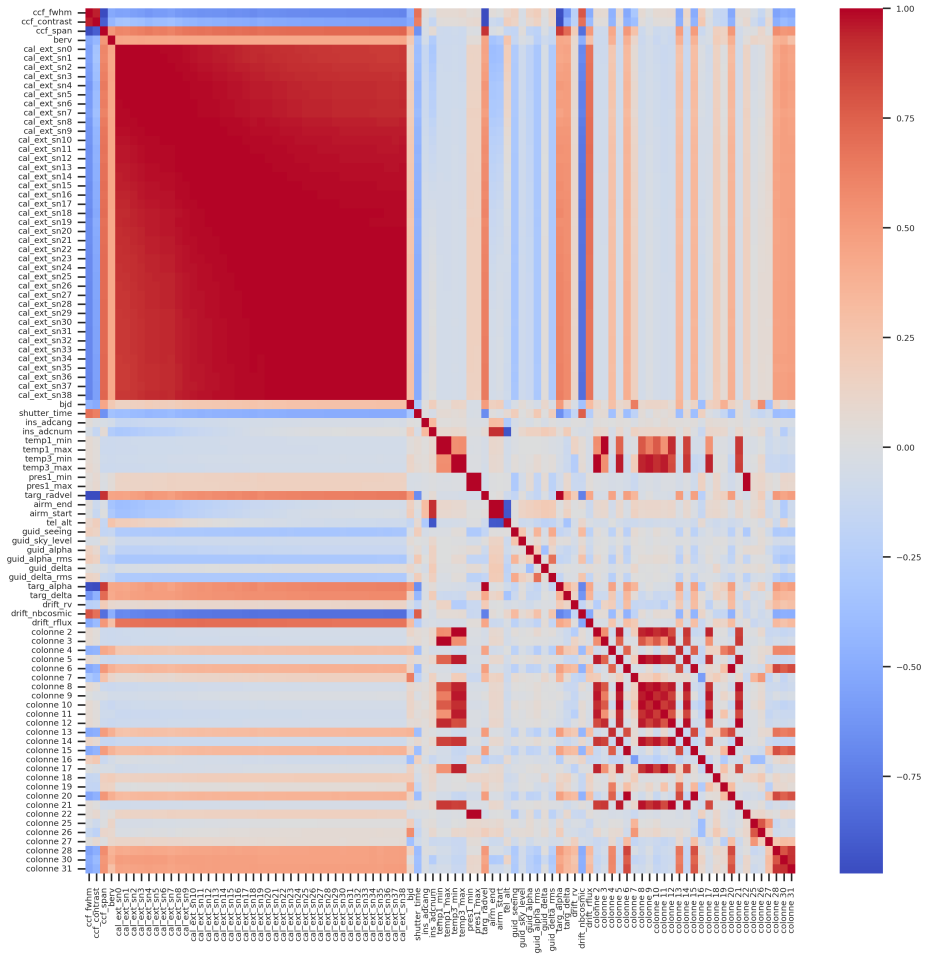

Figura 5.2. Matriz de correlación entre todas las variables predictoras iniciales. Para cada par de variables, el color representa el valor del coeficiente de correlación de Pearson.

Correlación con la velocidad radial

Variable	Coeficiente de Pearson
cal_ext_sn0	0.41
airm_end	0.38
airm_start	0.37
drift_rv	0.36
tel_alt	0.35
ins_adcnum	0.32
berv	0.26
drift_rflux	0.21
bjd	0.19
colonne 3	0.18
temp1_max	0.18
temp1_min	0.18
pres1_max	0.15
pres1_min	0.15
colonne 22	0.15

En general se habla de una correlación *fuerte* a partir de un valor de 0.5 del coeficiente de Pearson. Por lo que en principio en nuestras variables predictoras iniciales no tenemos ninguna fuertemente correlacionada con la velocidad radial.

5.2.4. Ingeniería de variables

La ingeniería de variables es el procedimiento de crear variables extras con el objetivo de mejorar el resultado del proceso de aprendizaje automático. Puede ser simplemente utilizando nuestro conocimiento del problema para inventar nuevas variables que puedan llegar a ser útiles o probar combinaciones distintas de las variables que ya tenemos.

Consideramos los pares de variables que expresan valores máximos y mínimos de una misma cantidad a lo largo de la observación, que por sí solas están muy correlacionadas, y las combinamos creando nuevas variables que contengan otra información que pueda llegar a ser relevante: por ejemplo, de "temp1_min" y "temp1_max", dejamos sólo la que expresa el mínimo y agregamos una nueva "temp1_dif" que expresa la diferencia entre el mínimo y el máximo durante la observación. Que no está correlacionada con las anteriores. Con este procedimiento sumamos 5 variables más.

La SNR de los órdenes de dispersión es la variable más correlacionada con la RV en el conjunto completo, como vimos en la subsección anterior. Sin embargo, son 39 variables que están extremadamente correlacionadas entre sí. Algo interesante que observamos cuando miramos estas correlaciones en cada estrella por separado, es que la correlación según el orden de dispersión con la RV es distinta para cada estrella (Figura 5.3). Es posible que este efecto esté relacionado con el tipo espectral de las estrellas. Para "capturar" esta dependencia en nuevas variables hicimos el siguiente procedimiento:

- De las 39 variables "cal_ext_sn**k**" con **k** de 0 a 38, definimos todas las combinaciones (sin orden): "cal_ext_snk/cal_ext_snj" y "cal_ext_snk-cal_ext_snj" para todo j≠k. Lo que da un total de 1482 variables adicionales.
- Ordenamos las variables resultantes según su coeficiente de correlación con la RV y agregamos la primera a nuestro conjunto de variables.
- Calculamos la correlación entre la nueva variable y todas las siguientes en la lista anterior, agregamos la primera que tenga un coeficiente menor a 0.8, de manera que no sea redundante.
- Repetir el paso anterior con la nueva variable agregada.

Así, se agregaron 19 nuevas variables. Se muestran en el apéndice A.1.

Un factor clave en la precisión de los instrumentos como SOPHIE es la estabilidad del perfil instrumental y por lo tanto la distribución de flujo a lo largo de los órdenes. En consecuencia, es esperable que estrellas de distintas características exhiban efectos diferentes. Por lo tanto, decidimos agregar variables que incorporen esta información al algoritmo. Sumamos entonces tres parámetros estelares como variables predictoras: "teff" (temperatura efectiva), "fe/h"

(metalicidad) y “logg” (gravedad superficial). Utilizamos los valores que se muestran en la Tabla 5.1 (obtenidos de Valenti & Fischer (2005)).

Estrella	T_{eff}[K]	Fe/H	log g
HD 185144	5246	-0.19	4.530
HD 89269	5586	-0.20	4.440
HD 9407	5657	0.02	4.480

Tabla 5.1. Valores de los parámetros estelares para las tres estrellas del conjunto de entrenamiento.

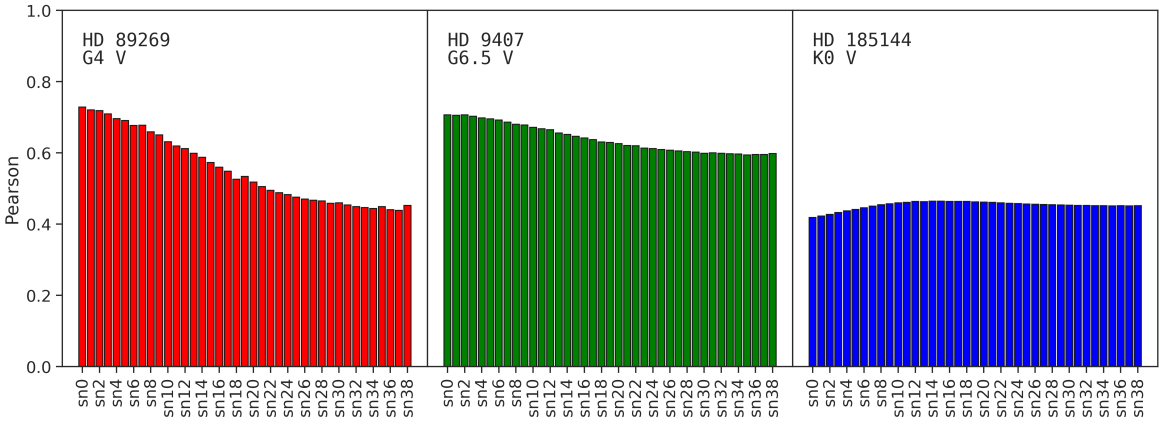

Figura 5.3. Correlaciones de las SNR de cada orden espectral con la velocidad radial para las tres estrellas de entrenamiento.

Después de agregar las nuevas variables, el conjunto quedó con un total de 119 variables predictoras. Si ahora volvemos a inspeccionar las 10 más correlacionadas con la velocidad radial, vemos que conseguimos varias con mejor correlación que las originales:

Correlación con la velocidad radial	
Variable	Coeficiente de Pearson
sn7-8	0.52
sn0/2	0.49
sn0/27	0.48
sn14/16	0.46
sn0-1	0.45
sn5/33	0.43
cal_ext_sn0	0.41
sn12/34	0.41
sn6-4	0.39
sn8/9	0.38
airm_start	0.37
sn23/12	0.37
drift_rv	0.36
tel_alt	0.35
ins_adcnum	0.32

5.2.5. Selección de variables

Hay varias razones por las cuales es deseable reducir la cantidad de variables predictoras. Tener muchas variables redundantes o irrelevantes sólo aumenta la complejidad del modelo y la posibilidad de encontrar patrones espúreos lo cuál lleva a mayor probabilidad de sobreajuste. Además, hay un efecto conocido como la *maldición de la dimensión:* cuanto mayor es la dimensionalidad del conjunto de entrenamiento mayor es el riesgo de que las instancias de entrenamiento sean demasiado dispersas, el aumento del volumen del espacio de características es exponencial con la cantidad de variables. Con lo cual las nuevas instancias que demos al modelo (datos de testeo) seguramente caerán lejos de cualquier dato de entrenamiento. Esto hace que las predicciones tengan mayor varianza que en un modelo más simple. Una solución en teoría sería incrementar el tamaño del conjunto de entrenamiento para tener una mayor densidad de datos, pero en nuestro caso estamos limitados a las 645 observaciones con las que contamos. Otras ventajas de reducir la cantidad de variables son que obtendremos un modelo más interpretable, y reducimos el tiempo de entrenamiento de los algoritmos, lo cual tiene un gran impacto en el proceso de búsqueda de hiperparámetros.

No existe una forma correcta o única de hacer selección de variables y en general depende de cada problema y conjunto de datos en específico. Las opciones son muchas pero nombraremos de forma general algunas maneras de abordar el problema. Se puede hacer de forma no-supervisada, es decir considerando sólo las variables predictoras y no su relación con la variable a predecir: por ejemplo, eliminando las variables redundantes. Y también hay métodos supervisados, donde se tiene en cuenta la interacción con la variable a predecir. Estos se dividen en varias categorías: por medio de estadísticas univariadas entre cada variable predictora y la variable objetivo, como puede ser el coeficiente de Pearson. O basados en modelos,

los cuales usan algoritmos que incorporan de forma intrínseca una forma de selección de las variables más importantes, con esta información se pueden entrenar sucesivos modelos e ir eliminando variables recursivamente. Alternativamente, se puede proceder de forma iterativa hacia adelante, empezando con una selección mínima y agregando de a una variable y computando si es beneficiosa para el algoritmo o no.

Probamos numerosas estrategias durante el trabajo, pero finalmente nos quedamos con una combinación de selección supervisada mediante modelos que hacen selección automática y a eso le sumamos eliminación de variables redundantes por coeficiente de correlación de Pearson entre ellas. El procedimiento de selección fue el siguiente:

- Se entrenó un modelo de GradientBoosting con las 119 variables iniciales, buscando los mejores hiperparámetros por validación cruzada, y se seleccionaron las 30 más importantes a través de sus coeficientes de *importancia*.
- Se entrenó un modelo lineal LassoLars también con las 119 variables iniciales, y se seleccionaron las 30 con coeficientes más altos (en módulo).
- Se procedió agregando en órden jerarquico de las dos listas: 1°_{GB}, 1°_{LASSO}, 2°_{GB}, ... Donde antes de agregar cada variable se computó su coeficiente de correlación de Pearson con todas las variables agregadas anteriormente. Si éste era mayor a 0.80 con alguna de ellas no se agregaba. Así hasta agotar ambas listas.

Al finalizar el procedimiento quedaron 27 variables. Luego hicimos un análisis de multicolinealidad entre estas, utilizamos el factor de inflación de varianza (VIF) que da una medida de qué tan bien se puede explicar una variable como combinación lineal de las otras. Se calcula como $VIF_i = 1/(1 - R_i^2)$ donde R_i^2 es el coeficiente de determinación obtenido al ajustar la variable x_i como combinación lineal de las variables restantes. Tomamos como límite un VIF de 10 (correspondiente a un R_i^2 de 0.9), y de esta forma descartamos tres variables con alta multicolinealidad de la lista. El conjunto final quedó con 24 variables que se muestran en la Figura 5.4.

5.2.6. Escalado de variables

Antes de pasar al entrenamiento de los algoritmos aplicamos una última transformación a las variables predictoras. En general los algoritmos tienen problemas cuando las variables de entrada tienen escalas de valores muy distintas, para evitar esto se normalizaron sustrayendo el valor medio y se dividieron por la desviación estándar, de manera que la distribución resultante tendrá media cero y varianza unitaria. En scikit-learn esta transformación se emplea con el método `StandarScaler`.

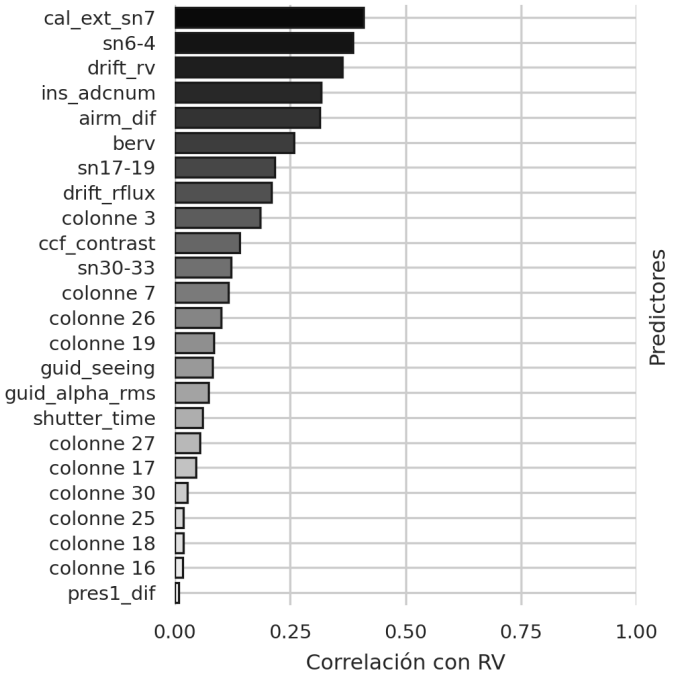

Figura 5.4. Variables seleccionadas y su coeficiente de correlación lineal con la velocidad radial.

Capítulo 6

Entrenamiento y resultados

En este capítulo se describe la etapa de entrenamiento de los algoritmos, se comparan sus desempeños y se elige el mejor modelo. Luego se aplica el mismo para corregir el error sistemático en observaciones de un grupo de estrellas fuera del conjunto de entrenamiento, y se comparan los resultados con la corrección mediante el método tradicional de la constante maestra.

Antes de comenzar, recapitulemos. Contamos con un conjunto de 645 puntos con observaciones de velocidad radial para las cuales preparamos y seleccionamos 24 variables predictoras para entrenamiento de los algoritmos. Como vimos en la Sección 4, es necesario contar con un conjunto de validación para selección de hiperparámetros y otro de testeo para evaluación final. En este caso, dado que no tenemos un conjunto tan grande de puntos y no podemos darnos el lujo de perder datos, entrenamos con el conjunto completo para aprovechar al máximo cada instancia de entrenamiento. Separamos una parte sólo para buscar los mejores hiperparámetros mediante validación cruzada (en 7 subconjuntos), pero luego reentrenamos el modelo elegido en el conjunto completo. El testeo lo realizamos en datos de otras estrellas que adquirimos luego. El conjunto de velocidades radiales sobre las cuales entrenamos se muestra en la Figura 6.1.

La lista de algoritmos que probamos es la siguiente:

1. Regresión Lineal
2. Regresión Lineal Ridge
3. Regresión Lineal LassoLARS
4. SVM (*support vector machines*)
5. Árboles de decisión
6. AdaBoost (*adaptive boosting*)
7. XGB (*extreme gradient boosting*)
8. Random Forest

9. Extra Trees (*extremely randomized trees*)

10. Gradient Boosting

Para medir y comparar el desempeño de los modelos elegimos el WRMSE (ver Sec. 4.1) como nuestra métrica principal para poder tener en cuenta las incertezas en las velocidades radiales y el hecho de que nuestro conjunto está *desbalanceado* en la proporción de datos que tenemos de cada estrella. Por tanto, definimos el peso de cada dato de velocidad radial como:

$$w_i = \frac{1}{\sigma_i^2 f_j}$$

donde σ_i es la incerteza en la RV y f_j es la fracción de datos del total del conjunto que corresponden a la estrella j.

Para la validación cruzada (ver Sec. 4.2), en todos los casos usamos el método `GridSearchCV` de scikit-learn que permite definir una grilla de hiperparámetros para un dado estimador, calcula la métrica de validación cruzada para cada uno de ellos y devuelve los parámetros del estimador óptimo. Y en algunos casos también usamos `RandomizedSearchCV`, donde en vez de una grilla de definen distribuciones para cada parámetro y se toman N combinaciones aleatorias a partir de las distribuciones definidas. La validación que hacen ambos es de tipo *K-fold*, es decir que a cada estimador se lo entrena K veces por cada combinación de hiperparámetros, por lo que es computacionalmente muy costoso para ciertos algoritmos. En nuestro caso usamos $K = 7$, con lo que un $\sim 15\,\%$ de puntos se usan para validar en cada iteración. La separación de datos en los 7 subconjuntos se realiza de manera aleatoria pero *estratificada*: respetando la proporción de puntos que tiene cada estrella en el total. Esto es para equilibrar la diferencia de datos que tenemos en cada tipo espectral.

Al principio hicimos numerosas pruebas con todos los algoritmos y encontramos que los de mejor rendimiento para nuestro conjunto de datos eran los métodos de ensamble de Random Forest y Gradient Boosting, con este último dando el mejor desempeño de todos.

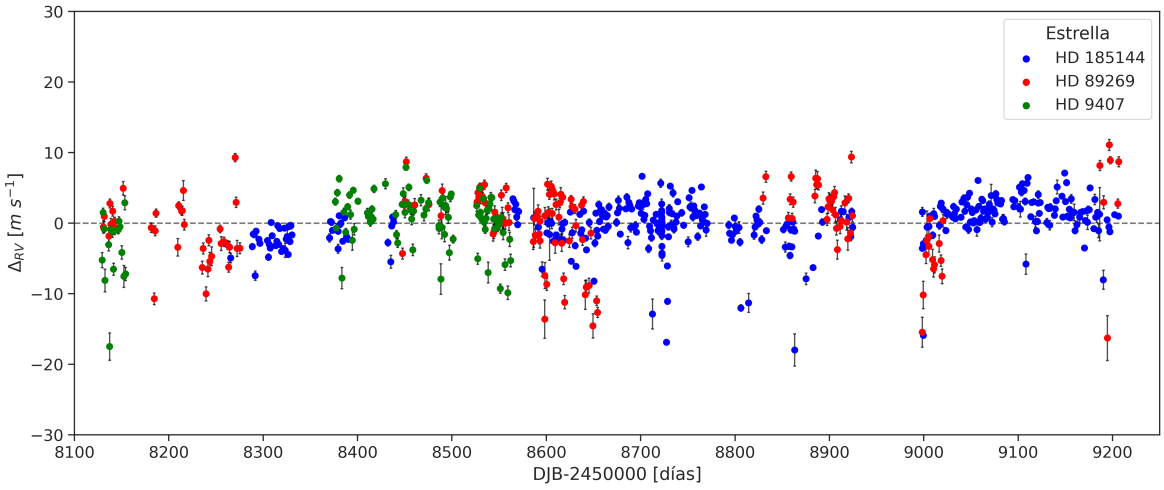

Figura 6.1. Las 645 velocidades radiales usadas en el entrenamiento. Se diferencian con colores las tres estrellas constantes.

Mientras que de los métodos lineales, LassoLARS tuvo un buen desempeño al usarlo con variables de interacción polinomiales, que aumentaron su poder predictivo significativamente. Utilizaremos los resultados de este último algoritmo como modelo base para comparar luego con el de Gradient Boosting.

6.1. Entrenamiento de LassoLARS

El algoritmo de LassoLARS es un modelo lineal Lasso (ver Subsec. 4.3.2) que utiliza el algoritmo LARS (ver Subsec. 4.3.3) para encontrar a cada paso los mejores coeficientes de la regresión. Para este algoritmo el único parámetro que debemos optimizar es la constante α (hiperparámetro `alpha`) que multiplica el término de penalización, recordemos que el objetivo del algoritmo Lasso es minimizar:

$$\min_{w} \quad \frac{1}{2N_{muestras}} \|Xw - y\|_2^2 + \alpha \|w\|_1 \tag{6.1}$$

de manera que un valor de `alpha` igual a cero da como resultado un ajuste normal de mínimos cuadrados. La métrica WRMSE que se reporta es el *score* obtenido en la validación cruzada por la mejor combinación de hiperparámetros encontrada, que se calcula como el promedio de los scores obtenidos en las 7 iteraciones, además mostramos el rango de valores que obtuvo en las iteraciones. Luego ese modelo se reentrena en el conjunto completo de entrenamiento y se reporta el coeficiente de determinación R^2 del modelo en los datos de entrenamiento. Se probó una grilla de valores para `alpha` entre 10^{-3} y 10^{-7}, el resultado obtenido fue:

Parámetros óptimos LassoLars	Resultado
`alpha` = 5.5405×10^{-5}	• `CV WRMSE: 2.759 m/s` `RANGO: [2.253, 3.363]` • `R2 SCORE: 0.520`

Aquí vemos que el algoritmo lineal simple de LassoLARS tiene un desempeño muy pobre, probablemente no haya una relación lineal entre las variables y el efecto sistemático. Para enriquecer el modelo de manera que incorpore no sólo las variables de forma independiente sino también interacciones entre las mismas, entrenamos nuevamente pero transformando las variables con el método `PolynomialFeatures` de scikit-learn. Esta función recibe como entrada un conjunto de variables y devuelve todas las combinaciones polinomiales posibles entre ellas hasta cierto grado que se define con el parámetro `degree`. Usamos esto para definir nuevas variables polinomiales hasta grado 2, y sólo de interacción, es decir sin los términos cuadráticos, esto es: si la entrada es (X_1, X_2, X_3) la función nos devuelve $(X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3)$. De esta forma, dado que tenemos 24 variables de entrada, la cantidad aumenta a 3700. Nuevamente probamos una grilla entre 10^{-3} y 10^{-7} para `alpha` y el resultado mejoró considerablemente respecto al caso anterior:

Parámetros óptimos LassoLars con variables de interacción	Resultado
`alpha =` 6.5085×10^{-5}	• `CV WRMSE: 2.331 m/s RANGO: [1.779, 2.789]` • `R2 SCORE: 0.741`

6.2. Entrenamiento de Gradient Boosting

Los algoritmos basados en árboles de decisión tienen la capacidad de capturar relaciones altamente no lineales entre las variables predictoras y la variable objetivo, por lo que no necesitamos incluir la transformación polinomial en principio. Al entrenar un algoritmo de Gradient Boosting, la cantidad de hiperparámetros a definir y optimizar es notablemente mayor. El número de estimadores base -dado por el parámetro `n_estimators`- lo vamos a obtener a través de la técnica de parada anticipada (*early stopping*), esto es: se entrena con un número grande de estimadores, pero se define una tolerancia (`tol`) y una fracción de validación (`validation_fraction`) de manera que para cada nuevo estimador el algoritmo calcula la mejora aportada en la función de pérdida. Si esta resulta menor a la tolerancia durante n iteraciones consecutivas (`n_iter_no_change`), entonces se frena el entrenamiento. Con esta técnica se reduce sustancialmente el costo computacional de la validación cruzada y se previene el sobreajuste. Los hiperparámetros que vamos a optimizar son:

- `learning_rate`: la tasa de aprendizaje.
- `max_depth`: la profundidad máxima de los árboles, limita el número de nodos y va a estar relacionado con las interacciones entre variables: a mayor profundidad más interacciones se pueden aprender. Menor profundidad previene el sobreajuste.
- `min_samples_split`: el número mínimo de muestras requeridas para abrir un nodo interno. Sirve para controlar el sobreajuste dado que un valor más alto previene que el algoritmo aprenda relaciones muy específicas para una dada muestra seleccionada en un árbol.
- `min_samples_leaf`: el número mínimo de muestras en un nodo terminal u *hoja*. Un número mayor restringe la flexibilidad del modelo y previene el sobreajuste.
- `max_features`: el número de variables que considera el algoritmo al buscar la mejor división interna o *split*.
- `subsample`: fracción de muestras que se usan para ajustar los estimadores base. Valores menores a 1 resultan en una disminución de la varianza y aumento del sesgo del modelo.

Mientras que los parámetros que dejamos fijos son:

- `loss`: “squared error” (error cuadrático medio)

- `n_estimators_`: 500
- `tol`: 1×10^{-10} (corresponde a una mejora de 0.01 $m\ s^{-1}$ en el RMSE)
- `n_iter_no_change`: 5
- `validation_fraction`: 0.1

La optimización se hizo en dos partes, primero optimizamos para la tasa de aprendizaje y la profundidad de los árboles:

```
learning_rate=[0.01, ... , 0.20]
max_depth=[2, 3, 4, 5]
```

dejando los demás parámetros con sus valores por defecto excepto por `max_features=‘‘sqrt’’`. Los valores óptimos resultaron ser 0.09 y 3 respectivamente. A continuación, fijamos estos valores y ajustamos los siguientes:

```
min_samples_split=[2, ... , 100]
min_samples_leaf=[1, ... , 100]
max_features=[5, ... , 24]
subsample=[0.75, ... , 1.0]
```

Para los cuales obtuvimos 5, 1, 11 y 0.83 como valores óptimos respectivamente. Al entrenar en el conjunto completo de entrenamiento con estos parámetros, la parada temprana se dio para `n_estimators_=77`. Las métricas fueron:

Parámetros óptimos
Gradient Boosting

```
learning_rate = 0.09
n_estimators = 77
max_depth = 3
min_samples_split = 5
min_samples_leaf = 1
max_features = 11
subsample = 0.83
```

Resultado

- **CV WRMSE: 2.265 m/s**
  RANGO: [1.844, 2.483]
- **R2 SCORE: 0.828**

De los tres modelos el que tuvo la mejor métrica en la validación cruzada fue Gradient Boosting (Tabla 6.1). En la Figura 6.2 mostramos el orden de importancia de las primeras 10 variables en los dos mejores modelos. La variable “drift_rv” resulta la de mayor importancia en ambos algoritmos de mejor desempeño, también había sido la de mayor ranking en el procedimiento de selección de variables por lo que no fue una sorpresa. Recordemos que esta variable mide justamente la deriva del punto cero (de *alta frecuencia*) al registrar cuánto se movieron las líneas de las lámparas desde la anterior calibración, lo cual da cuenta del error sistemático diario (las velocidades radiales ya son corregidas por este efecto en el pipeline)

y aparentemente también tiene relación con el error sistemático de largo plazo. La siguiente variable que aparece en ambos modelos está relacionada con la SNR (cal_ext_sn7) (también hay varias combinaciones entre SNRs de distintos órdenes en la lista), ya habíamos visto que las SNR tenían una buena correlación con las velocidades radiales por lo cual es esperable que tuvieran importancia. Lo otro destacable es que las dos temperaturas que aparecen entre las 10 variables más importantes para GB, que son "colonne 17" y "colonne 3", están relacionadas con el mismo contenedor refrigerado: la temperatura de la pared norte del contenedor y la temperatura del aire dentro del mismo, ambas son estables dentro de un rango de 0.1°C, pero posiblemente variaciones de menor amplitud afectan a las mediciones de velocidad radial.

En la sección siguiente aplicaremos la corrección de nuestro modelo en datos nuevos de otras estrellas.

Modelo	CV WRMSE $[m\ s^{-1}]$	R^2
LassoLARS lineal	2.759	0.520
LassoLARS polinomial	2.331	0.741
Gradient Boosting	2.265	0.828

Tabla 6.1. Métricas para cada modelo. CV WRMSE corresponde al puntaje obtenido en la validación cruzada por el modelo con los mejores parámetros. El coeficiente R^2 se calcula con el mejor modelo entrenado sobre todo el conjunto de entrenamiento.

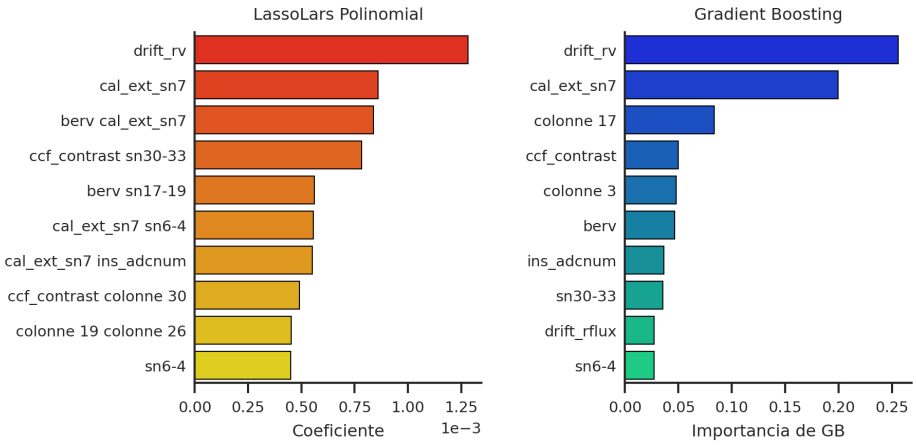

Figura 6.2. Las 10 variables de mayor importancia para los modelos de LassoLARS con variables polinomiales y para Gradient Boosting. Para el primero reportamos el valor absoluto del coeficiente de regresión, y en GB se muestra la importancia computada por el algoritmo.

6.3. Aplicación del modelo y comparación con la corrección maestra

Para usar nuestro modelo como método de predicción a la deriva del punto cero en las medidas de velocidad radial, hicimos una búsqueda en el archivo de SOPHIE de observaciones

dentro del rango temporal de entrenamiento (BJD 2458129.419 a BJD 2459206.623) con las siguientes restricciones: sólo tipos espectrales F, G y K, con observaciones en modo de alta resolución (HR), que hubieran sido observadas con el método de calibración simultanea, y que tuvieran más de 15 medidas de velocidad radial con una dispersión no mayor a 1 $km\ s^{-1}$ (para descartar binarias). Encontramos 6 estrellas, que se detallan en la siguiente tabla:

Estrella	Tipo espectral	$N°$ de medidas
HD 73344	G2V	320
HD 163183	G2V	144
HD 161284	K0V	94
HD 173701	G8V	62
TOI 1386	K0V	23
HD 207897	K0V	18

A diferencia de las estrellas de entrenamiento, estas no son constantes de velocidad radial. De hecho algunas tienen planetas confirmados. HD 73344 tiene un planeta de tipo sub-Neptuno con un período de 15.6 d detectado por tránsitos pero sin determinación de masa (de Leon et al., 2021). HD 161183 es una variable BY Draconis. TOI 1386 tiene un candidato planetario detectado por tránsitos con un radio estimado de 6.55 $R_\oplus$ pero sin determinación de período ni masa (Chontos et al., 2022). Mientras que HD 207897 cuenta con un planeta de tipo sub-Neptuno confirmado mediante tránsitos y velocidades radiales de 16.2 d de período, la masa reportada tiene dos valores posibles de 14.4 $M_\oplus$ o 15.9 $M_\oplus$ con igual probabilidad (Heidari et al., 2022). Las dos estrellas restantes, HD 161284 y HD 173701 no cuentan por el momento con reportes de candidatos planetarios. Debido a esto, no serviría de nada calcular las métricas que usamos para entrenamiento para estas estrellas, ya que la variación intrínseca domina sobre la instrumental.

Para evaluar el desempeño de nuestro modelo con respecto a la corrección tradicional definida por Courcol et al. en 2015 y adoptada como estándar, aplicamos ambas correcciones a todas las estrellas y comparamos el cambio en las dispersiones de los datos antes y después de cada corrección.

La corrección maestra que obtuvimos del equipo de SOPHIE consiste en una serie de 2881 medidas de velocidades radiales de estrellas constantes que abarcan desde BJD 2455930.117 a BJD 2459489.5452. Siguiendo el procedimiento descripto por Courcol et al. les aplicamos un filtro de media cada 15 medidas, dando un tiempo de escala típico de 9.3 días para la corrección. Luego interpolamos estos puntos para definir una función que da la corrección para cualquier fecha dentro del rango. Se muestra la serie de medidas que constituyen la corrección maestra en la Figura 6.3, donde también se marca cómo queda la función de corrección dada por la interpolación de los promedios cada 15 días. Hicimos lo mismo para los datos usados en entrenamiento como comparación y se muestran en la Figura 6.3.

6.3.1. Comparación de las correcciones

Para aplicar el modelo de Gradient Boosting que entrenamos, necesitamos obtener las 24 variables predictoras para las seis estrellas. Para esto realizamos la adquisición y el procesado

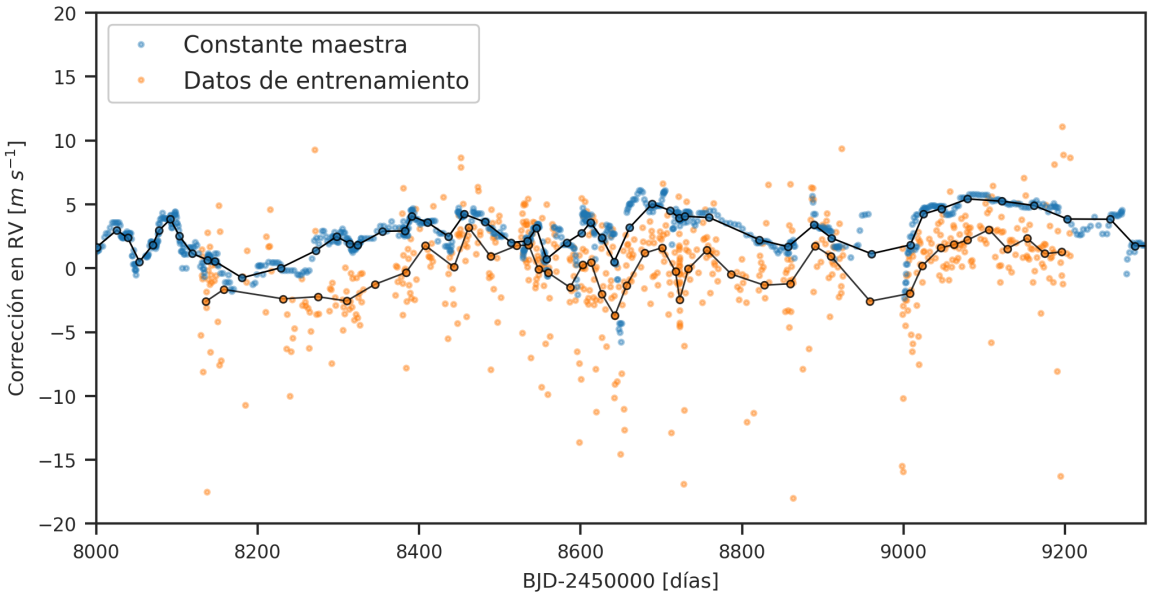

Figura 6.3. Los puntos azules corresponden a la constante maestra usada en la corrección tradicional. La línea sólida es el resultado de interpolar los promedios de los puntos cada 15 medidas. Los puntos naranja son las velocidades radiales usadas en el entrenamiento del algoritmo, para comparación se muestra también la interpolación del promedio cada 15 medidas.

de datos de manera idéntica a lo descripto para las estrellas de entrenamiento pero para las 6 nuevas estrellas. Luego escalamos las variables utilizando la misma normalización que aplicamos a estas, es decir, se les sustrae la media y se dividen por la varianza, siendo ambos parámetros los de la distribución de entrenamiento. En la Figura 6.4 se muestran las medidas sin correcciones de velocidades radiales de las seis estrellas (a las cuales se les sustrajo la media). Para cada una también se muestran los valores de la deriva en la velocidad radial predichos por el método tradicional y por nuestro modelo de Gradient Boosting. Lo primero que notamos es que las predicciones de nuestro modelo parece caer siempre por debajo de la corrección tradicional, con un *offset* negativo que va en promedio desde 1.9 $m\ s^{-1}$ en HD 73344 hasta 10 y 11 $m\ s^{-1}$ en HD 207897 y TOI-1386 respectivamente. Esto puede estar relacionado con el offset que se observa también entre las RVs de entrenamiento y la constante maestra que se ve en la Figura 6.3. Si bien ambos conjuntos siguen la misma tendencia en el tiempo, la diferencia en el punto cero seguramente está relacionada con que la constante maestra contiene el promedio de las velocidades de un número mayor de estrellas además de las 4 super-constantes iniciales, como mencionamos en la Sección 3.4. Por esto, no es extraño que las medias difieran.

Por otra parte, dado que estamos dando al modelo datos completamente nuevos, algo a considerar es que los datos de entrenamiento no sean suficientemente representativos para los datos nuevos. Si los valores de las variables predictoras en estas nuevas estrellas caen muy alejadas de las distribuciones "vistas" en el entrenamiento por el algoritmo, eso puede dar lugar a que el modelo tenga problemas para predecir bien estos nuevos puntos. Para examinar esto, graficamos las distribuciones conjuntas de a pares de las 5 variables predictoras

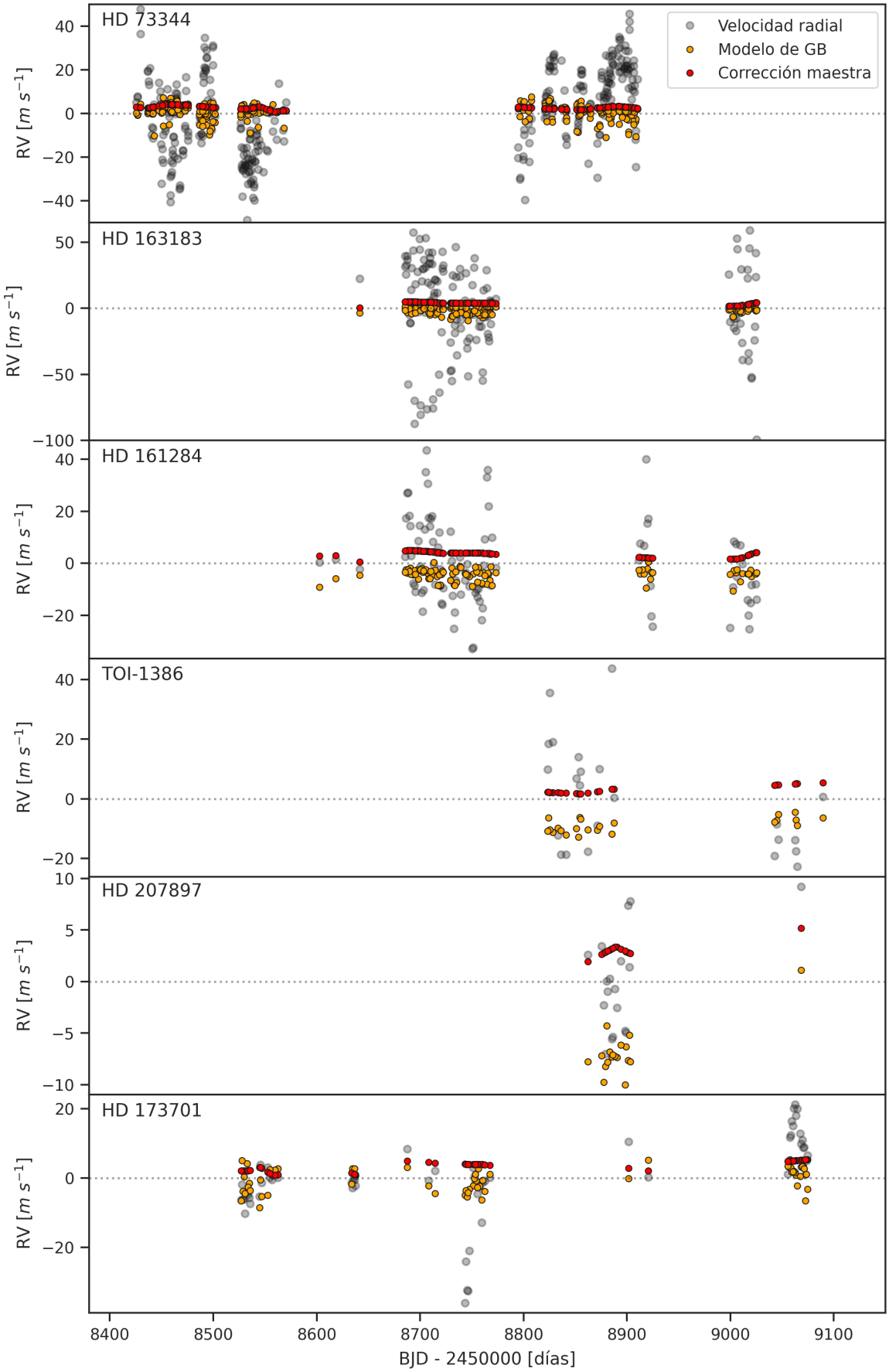

Figura 6.4. Velocidades radiales (gris) de las seis estrellas y las correspondientes predicciones de la deriva en velocidad radial dada por el modelo de GB (naranja) y por la constante maestra (rojo).

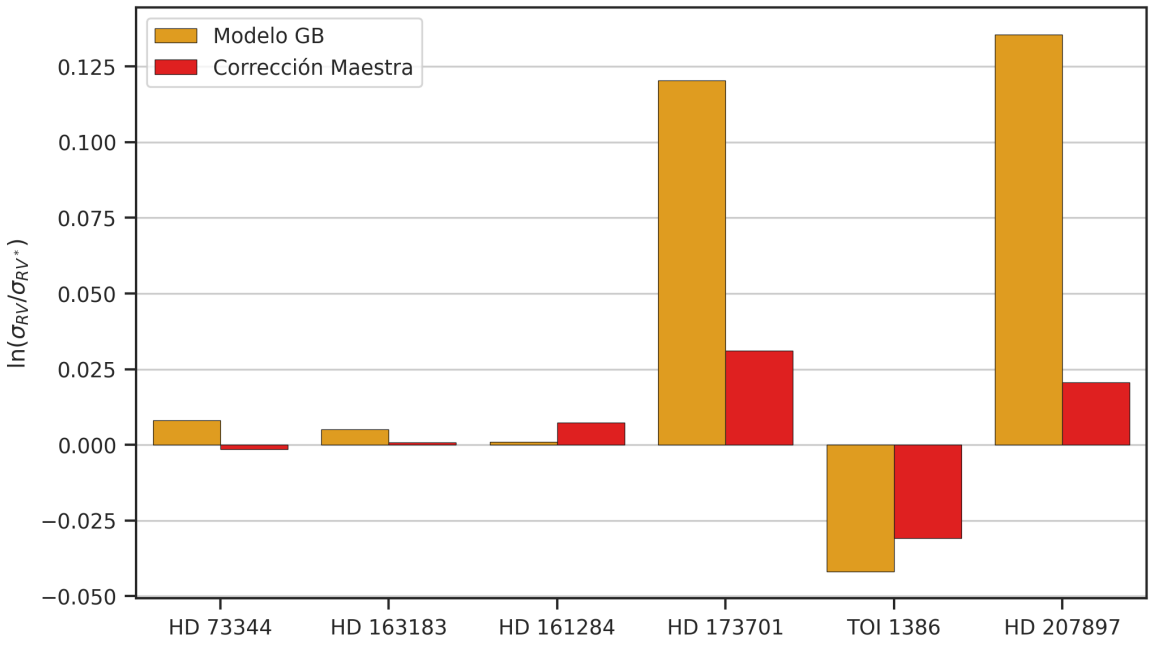

Figura 6.5. Cocientes de las dispersiones en las velocidades radiales antes (σ_{RV}) y después de aplicar la corrección dada por el modelo de GB (σ_{GB}) y la de la constante maestra (σ_{CM}). Los valores de los cocientes están en escala logaritmica.

más importantes, tanto para los datos de entrenamiento como para las seis estrellas nuevas (ver Apéndice C). Analizando estas distribuciones, en particular la columna correspondiente a la relación SNR del orden de dispersión 7 (cal_ext_sn7) -que es la segunda variable más importante- se observa que para la estrella en la que el offset es menor los valores de esta variable caen cerca del centro de la distribución de entrenamiento, mientras que para las dos que tienen los offsets más extremos (HD 207897 y TOI-1386) los nuevos puntos están en la cola negativa de la distribución. Esto podría explicar el offset en las predicciones para estas estrellas.

Más allá de este offset, podemos analizar el efecto que tiene la corrección en la dispersión del conjunto de velocidades radiales de cada estrella. Para esto, calculamos las dispersiones antes y después de cada corrección. Se muestran los resultados en la Tabla 6.2 y en la Figura 6.5. Salvo casos particulares, la tendencia general en un conjunto de estrellas al aplicar la corrección debería ser la de reducir la dispersión, por más que no sean estrellas constantes. Vemos que en dos estrellas nuestro modelo produce una reducción de la dispersión mucho mayor que la corrección tradicional, en HD 173701 que pasa de 11.637 a 10.381 $m\ s^{-1}$ y en HD 207897, donde se reduce de 4.816 a 4.206 $m\ s^{-1}$. En TOI-1386 con ambas correcciones la dispersión aumenta, y en las tres restantes no se aprecia un cambio significativo con ninguna de las dos correcciones.

Estrella	HD 73344	HD 163183	HD 161284	HD 173701	TOI 1386	HD 207897
σ_{RV} $[ms^{-1}]$	20.335	34.998	16.075	11.637	18.212	4.816
$\sigma^*_{RV_{CM}}$ $[ms^{-1}]$	20.363	34.972	15.958	11.282	18.781	4.718
$\sigma^*_{RV_{GB}}$ $[ms^{-1}]$	20.171	34.823	16.061	10.318	18.991	4.206

Tabla 6.2. Dispersiones en las velocidades radiales antes y después de las correcciones.

6.3.2. Discusión

En la primera iteración de este trabajo, el entrenamiento se había hecho sólo en la estrella HD 185144 obteniendo buenos resultados en el testeo sobre una porción de esos datos. Sin embargo, al usar el modelo en las estrellas constantes HD 89269 y HD 9407 comprobamos que no arrojaba buenas predicciones. Dada la importancia de la SNR en los modelos, supusimos que esto se debía a que las estrellas en las que testeamos eran de distintos tipos espectrales a la de entrenamiento. Por esta razón, en la siguiente etapa agregamos estas dos constantes al conjunto de entrenamiento para poder capturar en el modelado este efecto relacionado al tipo espectral. El modelo entrenado en esa iteración nuevamente resultó excelente cuando lo probamos en un conjunto de testeo con datos de las mismas estrellas de entrenamiento (Serrano Bell & Díaz, 2022). No obstante, faltaba testear todavía en datos *nuevos* para ver si tenía la capacidad de generalizar las predicciones a estrellas dentro del rango F, G y K.

Para testear en nuevos datos, surgió el problema de que no encontramos dentro del rango temporal observaciones de otras estrellas constantes. Así que testeamos el modelo en estrellas con altas dispersiones intrínsecas, para las cuales la predicción no se puede evaluar con la métrica de entrenamiento. Surgen dos consideraciones respecto al desempeño en estas estrellas: por un lado, vimos al comparar con la corrección tradicional que la dispersión de las velocidades radiales luego de la corrección en líneas generales no empeora (salvo por un caso donde ambas correcciones empeoran la dispersión) y en dos casos obtenemos una mejora en la dispersión respecto del método anterior.

Por otra parte, si bien nos resultó llamativo que las predicciones tengan una tendencia a irse hacia el negativo, el hecho de que difieran de la corrección maestra es esperable por la forma en que ámbos métodos están planteados. La corrección que se aplica hoy en día por medio de la serie maestra de RVs constantes supone que el efecto es el mismo para todas las estrellas F, G y K y les aplica la misma corrección, sin embargo, en nuestro abordaje hemos constatado en más de una oportunidad que el efecto varía con el tipo espectral. Nuestro modelo produce correcciones distintas para cada estrella, por lo tanto es normal que difieran en cierta medida.

Una forma ideal para testear las correcciones sería a partir de obtener periodogramas antes y después de la correción para una estrella que tenga un planeta de masa y periodo conocidos. El efecto de la corrección no debería afectar la señal periódica correspondiente al planeta compañero. Sin embargo, de las seis estrellas de testeo, la única que cumple estas condiciones es HD 207897, pero la cantidad de puntos con los que contamos (18) no es suficiente para realizar este test.

Capítulo 7

Conclusiones y trabajo a futuro

El objetivo de este trabajo fue realizar un estudio del efecto sistemático que se observa en las velocidades radiales de SOPHIE como una deriva a largo plazo de su punto cero. Esta variación sistemática de las velocidades se monitorea observando estrellas de velocidad constante, que se usan además para realizar una corrección de las observaciones. La finalidad de este trabajo fue usar algoritmos de aprendizaje automático para modelar este efecto entrenándolos sobre datos de estrellas de velocidad radial constante, para las cuales la deriva se hace evidente. Teniendo como meta la mejora de la precisión en las medidas de velocidad radial utilizadas para la detección y caracterización de exoplanetas.

Bajo la hipótesis de que existe una relación entre las variables observacionales, ambientales e instrumentales, como la posición del telescopio, la temperatura ambiente y en distintas partes del instrumento, etc., con el efecto observado en el punto cero de las velocidades radiales, compilamos un extenso conjunto de datos a partir de 645 observaciones de velocidad radial de estrellas constantes. A partir de técnicas de preprocesado, y de ingeniería y elección de variables seleccionamos un grupo óptimo de variables predictoras.

Entrenamos diversos algoritmos y evaluamos sus desempeños, para finalmente quedarnos con un modelo de Gradient Boosting. Mostramos la comparación con un modelo lineal regularizado como base.

Luego aplicamos la corrección predicha por nuestro modelo a observaciones de un grupo de seis estrellas y la comparamos con la corrección que debería aplicarse mediante el método tradicional que se usa en SOPHIE definido por Courcol et al. (2015). Surgió una dificultad para comparar los métodos debido a que las estrellas disponibles para testeo no son constantes de velocidad radial y pueden tener compañeros invisibles que generan una variación real (no instrumental) de sus velocidades radiales. Esto complica medir el impacto de las correcciones. Pero la comparación de las dispersiones de las RVs antes y después de efectuar las correcciones permitió concluir que el método implementado en este trabajo conduce a una mejor corrección de los efectos sistemáticos.

Otro de los objetivos del trabajo era obtener conocimiento del instrumento a partir de los datos. En ese sentido, remarcamos que la corrección que se aplica hoy en día por medio de la serie maestra de RVs constantes supone que el efecto es el mismo para todas las estrellas F,

G y K y les aplica la misma corrección, sin embargo, en nuestro abordaje hemos constatado en más de una oportunidad que el efecto varía con el tipo espectral.

El trabajo de tesis requirió conocer en detalle el espectrógrafo de alta resolución SOPHIE y las técnicas de obtención de velocidades radiales de alta precisión para la búsqueda y caracterización de exoplanetas. Además, fue necesario profundizar los conocimientos de Python y familiarizarse con el paquete scikit-learn de aprendizaje automático.

En el futuro, en el marco del trabajo de Doctorado, se espera utilizar la experiencia adquirida con SOPHIE para desarrollar una herramienta similar para otros instrumentos como HARPS o SPIRou. Además, se buscará extender la herramienta para poder predecir variables multivariadas. En su versión actual, el algoritmo predice solamente la velocidad radial, pero en el futuro, intentaremos predecir además las variaciones del ancho a mitad de altura de una línea media y su deformación (medida por el *span* del bisector; ver Queloz et al. (2001)). En una segunda instancia, se buscará predecir las observaciones en cada orden espectroscópico individualmente. Como última instancia, buscaremos directamente predecir el flujo en rangos de la longitud de onda, lo que producirá una corrección a nivel del espectro, en lugar de hacerlo posteriormente sobre las mediciones reducidas de velocidad radial. Así, en cada paso nos acercamos a una descripción de más bajo nivel de las observaciones.

Apéndice A

Datos

A.1. Lista de variables predictoras iniciales.

Feature	Descripción
extraídos de los headers	
ccf_rv	Velocidad radial medida (en $km\ s^{-1}$)
ccf_fwhm	FWHM de la CCF
ccf_contrast	Contraste de la CCF
ccf_span	Bisector span de la CCF
berv	Corrección por velocidad radial baricéntrica
cal_ext_sn**k** (con k de 0 a 38)	SNR del orden de dispersión **k**
bjd	Fecha juliana baricéntrica
ccf_rv_error	Error de la velocidad radial medida (en $m\ s^{-1}$)
shutter_time	Tiempo de exposición
ins_adcang	Ángulo de la lente ADC[(i)]
ins_adcnum	Número de lente usada en el ADC
temp1_min	Temperatura mínima de la red de difracción
temp1_max	Temperatura máxima de la red de difracción
temp3_min	Temperatura mínima del aire local
temp3_max	Temperatura máxima del aire local
pres1_min	Presión mínima del aire local
pres1_max	Presión máxima del aire local
pres2_min (descartada)	Presión mínima en el tanque presurizado
pres2_max (descartada)	Presión máxima en el tanque presurizado
targ_radvel	Velocidad radial conocida
airm_end	Masa de aire al finalizar la exposición
airm_start	Masa de aire al empezar la exposición

[(i)] ADC (*Atmospheric dispersion corrector*) es un sistema de corrección por dispersión atmosférica. Las lentes van rotando a medida que se mueve la estrella en el cielo. El sistema cuenta con distintas lentes según la masa de aire.

tel_alt	Altura del telescopio
guid_seeing	Estimación del seeing del sistema de guiado
guid_sky_level	Estimación del nivel de cielo del sistema de guiado
guid_alpha	Corrección media en alpha en arcsec
guid_alpha_rms	RMS de la corrección en alpha en arcsec
guid_delta	Corrección media en delta en arcsec
guid_delta_rms	RMS de la corrección en delta en arcsec
senti_text (descartada)	Temperatura exterior
senti_humi (descartada)	Humedad exterior
senti_seeing (descartada)	Estimación del seeing
senti_tciel (descartada)	Temperatura del cielo
senti_mag (descartada)	Magnitud del cielo
targ_alpha	Ascensión recta del objeto
targ_delta	Declinación del objeto
drift_rv	Deriva en la velocidad radial medida en las líneas de lámpara desde la última calibración
drift_nbcosmic	Número de rayos cósmicos detectados en los espectros de la lámpara
drift_rflux	Cociente de flujo en la lámpara de Th usada para el "drift"
extraídos de los sensores externos	
colonne 2	Temperatura de la red de difracción
colonne 3	Temperatura del aire en el tanque refrigerado
colonne 4	Temperatura del aire en el tanque a presurizado
colonne 5	Temperatura en el punto de contacto del soporte sureste
colonne 6	Temperatura en la extensión del soporte sureste
colonne 7	Temperatura de la superficie del criostato
colonne 8	Temperatura de la férula
colonne 9	Temperatura del obturador
colonne 10	Temperatura del mezclador (*scrambler*
colonne 11	Temperatura de la pared Sur
colonne 12	Temperatura del banco óptico (lado Este)
colonne 13	Temperatura del techo por fuera del contenedor
colonne 14	Temperatura del aire en el recinto aislado
colonne 15	Temperatura del aire en el soporte base
colonne 16	Temperatura del CCD
colonne 17	Temperatura de la pared Norte del contenedor regrigerado
colonne 18	Temperatura del aire en la sala de observación
colonne 19	Temperatura del aire en la sala de electrónica
colonne 20	Temperatura en el contacto del soporte base
colonne 21	Temperatura de contacto del aislante mecánico del pie sureste
colonne 22	Presión atmosférica

colonne 23 (descartada)	Presión en el tanque presurizado
colonne 24 (descartada)	Temperatura interna del sensor de presión
colonne 25	Temperatura ambiente en la unidad de calibración
colonne 26	Temperatura de la alimentación en la unidad de calibración
colonne 27	Temperatura de las lámparas en la unidad de calibración
colonne 28	Humedad en la habitación de la unidad de calibración
colonne 29 (descartada)	Corriente de la lámpara de Tungsteno
colonne 30	Temperatura interior del armario eléctrico
colonne 31	Temperatura interior del covertor (*bonnette*)
agregadas por ingeniería de features	
temp1_dif	Diferencia entre temp1_max y temp1_min
temp3_dif	Diferencia entre temp3_max y temp3_min
pres1_dif	Diferencia entre pres1_max y pres1_min
pres2_dif	Diferencia entre pres2_max y pres2_min
airm_dif	Diferencia entre airm_end y airm_start
airm_ratio	Cociente entre airm_end y airm_start
sn0/2	Cociente entre cal_ext_sn0 y cal_ext_sn2
sn7-8	Diferencia entre cal_ext_sn7 y cal_ext_sn8
sn14/16	Cociente entre cal_ext_sn14 y cal_ext_sn16
sn0-1	Diferencia entre cal_ext_sn0 y cal_ext_sn1
sn8/9	Cociente entre cal_ext_sn8 y cal_ext_sn9
sn6-4	Diferencia entre cal_ext_sn6 y cal_ext_sn4
sn0/27	Cociente entre cal_ext_sn0 y cal_ext_sn27
sn11-2	Diferencia entre cal_ext_sn11 y cal_ext_sn2
sn23/12	Cociente entre cal_ext_sn23 y cal_ext_sn12
sn26-37	Diferencia entre cal_ext_sn26 y cal_ext_sn37
sn22/34	Cociente entre cal_ext_sn22 y cal_ext_sn34
sn17-19	Diferencia entre cal_ext_sn17 y cal_ext_sn19
sn2-8	Diferencia entre cal_ext_sn2 y cal_ext_sn8
sn32/33	Cociente entre cal_ext_sn32 y cal_ext_sn33
sn12/34	Cociente entre cal_ext_sn12 y cal_ext_sn34
sn13-38	Diferencia entre cal_ext_sn13 y cal_ext_sn38
sn5/33	Cociente entre cal_ext_sn5 y cal_ext_sn33
sn30-33	Diferencia entre cal_ext_sn30 y cal_ext_sn33
sn22/33	Cociente entre cal_ext_sn22 y cal_ext_sn33

Tabla A.1. Nombre y descripción de las variables predictoras iniciales y las agregadas por ingeniería de variables. Se indica las que fueron descartadas por contener $> 20\,\%$ de valores malos o nulos.

A.2. Histogramas de variables predictoras iniciales

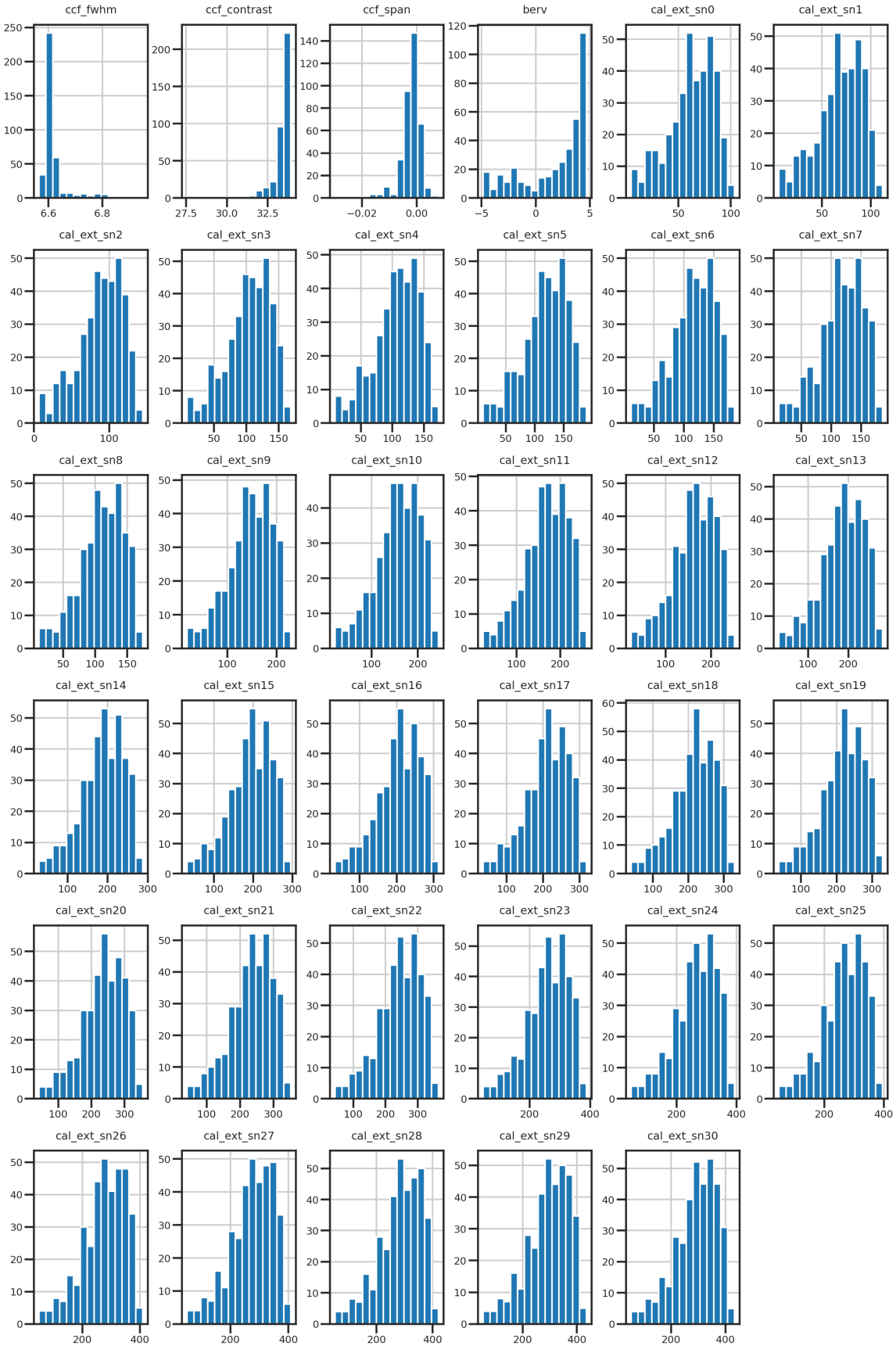

Figura A.1. Histogramas de variables predictoras iniciales.

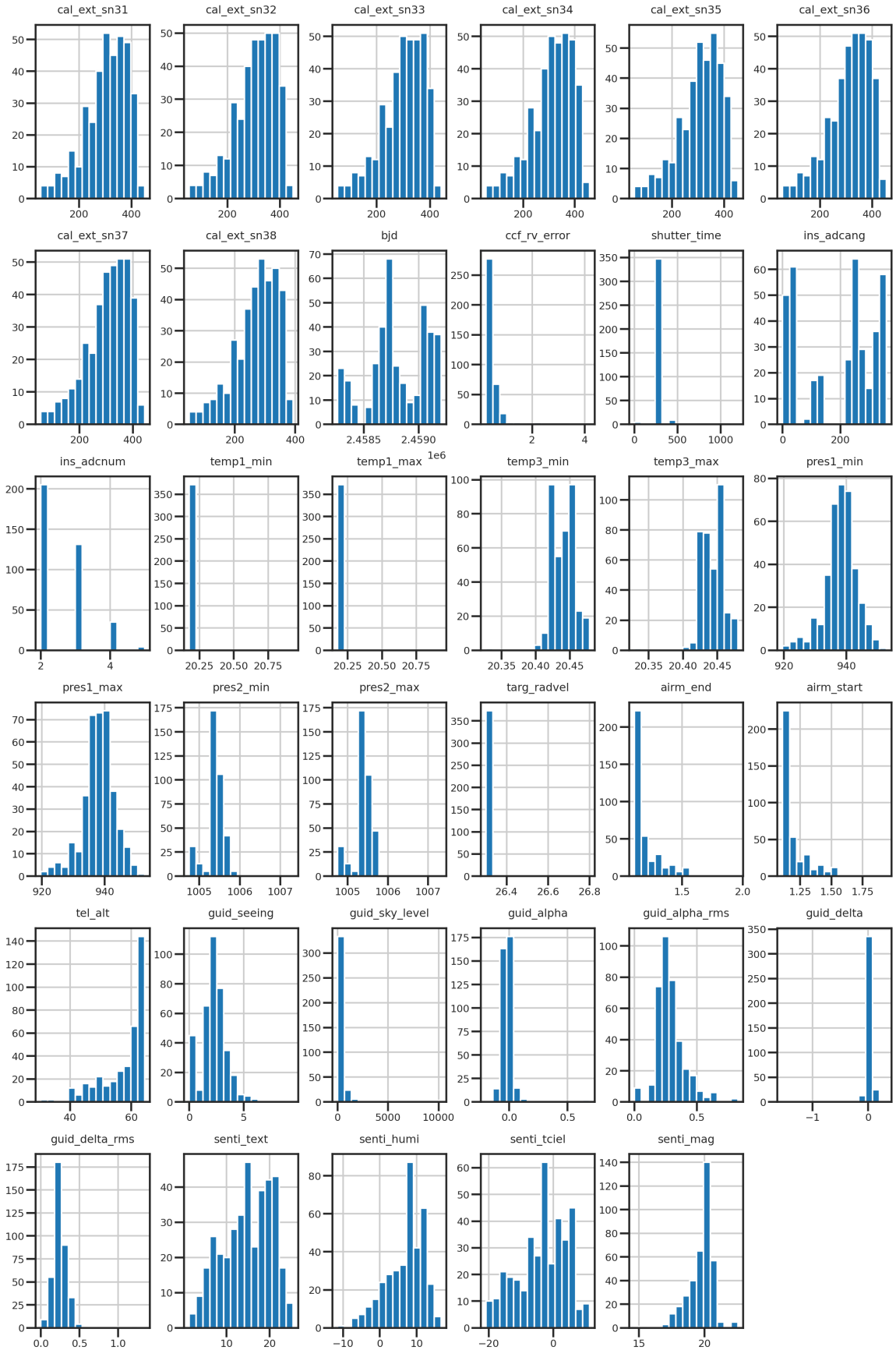

Figura A.2. Histogramas de variables predictoras iniciales.

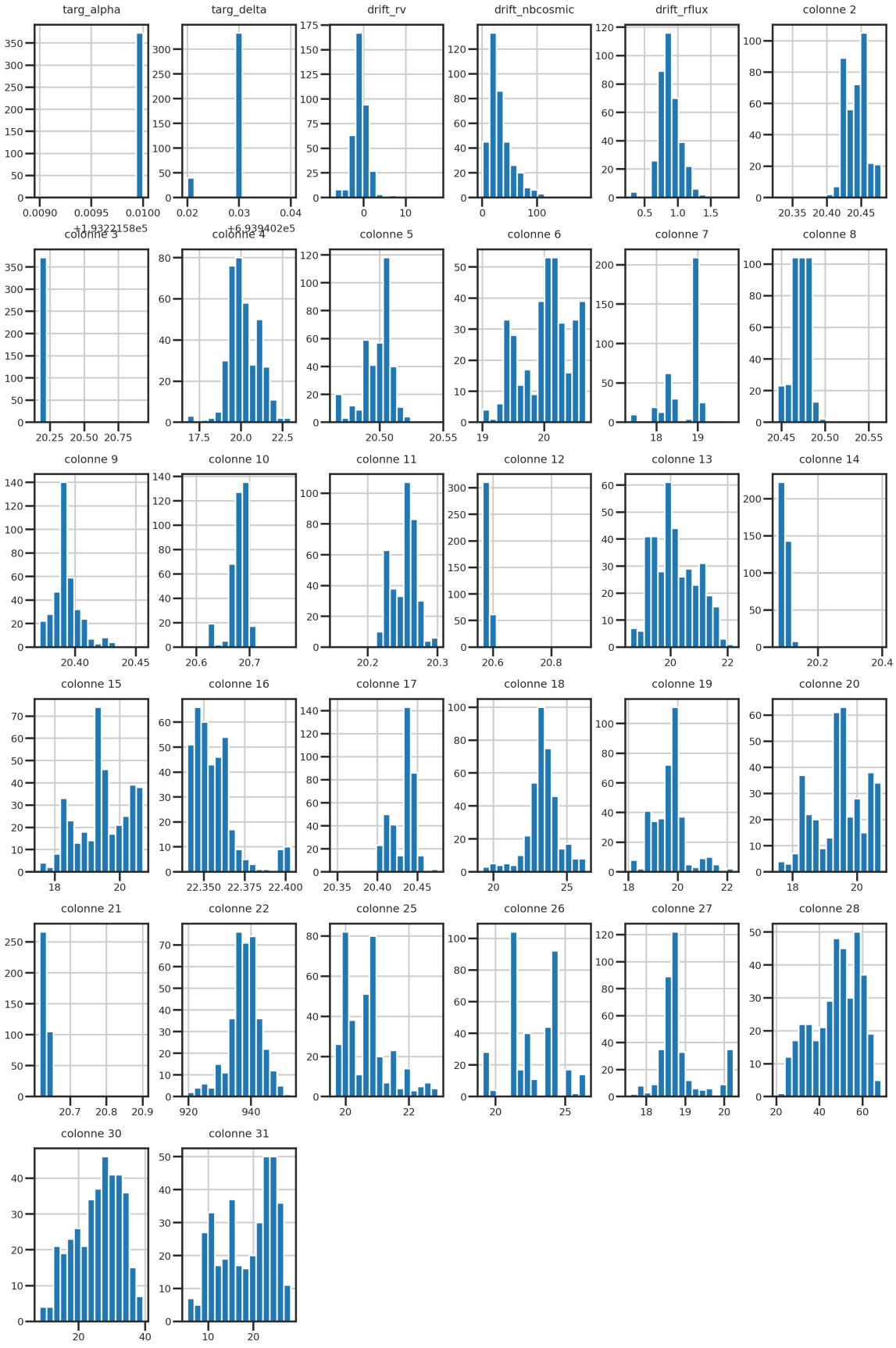

Figura A.3. Histogramas de variables predictoras iniciales.

Apéndice B

Árboles de decisión

B.1. Explicación breve del algoritmo de Árboles de decisión

Los árboles de decisión son modelos muy usados en aprendizaje automático tanto para regresión como en clasificación, y generalmente son el modelo ”base” de elección en los métodos de ensamble. El algoritmo tiene una estructura de ”árbol” con tres tipos de nodos. El **nodo raíz** es el que inicia la estructura y contiene todas las muestras que luego pueden ser separadas en nodos adicionales, los **nodos internos**, que son los que se derivan del nodo raíz y los **nodos hoja**, que son los que dan los valores de *salida* del algoritmo. Los nodos se unen a través de **ramas**.

La predicción para una dada instancia empieza en el nodo raíz, donde se hace una pregunta de Verdadero o Falso respecto al valor de la instancia para una de las variables predictoras (si es mayor o menor que un dado número de corte), y según la respuesta se sigue por una de las dos ramas salientes. De esta forma se dividió el conjunto en dos subconjuntos. Según la respuesta anterior, se llega a través de una de las ramas a un nodo interno, donde el algoritmo se hace una nueva pregunta que puede ser respecto a otra de las variables predictoras o la misma. Esto subdivide otra vez el número de muestras, abriendo nuevamente dos ramas. Si es un árbol de profundidad 2, ya no hay más preguntas, y la salida del nodo interno va hacia las hojas, que representan el valor predicho por el algoritmo para esa dada instancia. Dicho valor será el promedio de las muestras que quedaron en esa hoja. De esta forma, para todo punto, el algoritmo empieza en la raíz y termina con una predicción en una de las hojas.

Para entenderlo mejor veamos en un ejemplo práctico con nuestros datos. Vamos a construir un modelo que prediga la altura del telescopio teniendo como variable predictora a la masa de aire. Primero entrenamos un modelo simple primero, con una profundidad de 2. En la figura B.1 se muestra la estructura óptima del árbol encontrada por el algoritmo. La forma de encontrar el mejor valor numérico para hacer las divisiones en una dada variable la hace calculando la función de pérdida (el error cuadrático medio, en este caso) para una grilla de puntos de división y seleccionando el valor que minimiza el error. A través de este procedimiento el algoritmo va construyendo el árbol.

En la figura B.2 se muestran dos modelos con profundidad 2 y 4, a medida que aumentamos

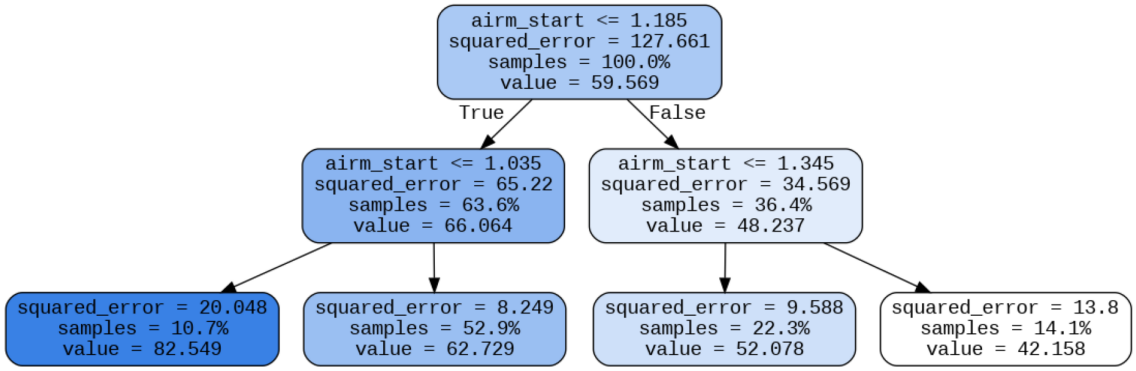

Figura B.1. Estructura del árbol entrenado con profundidad de 2. "airm_start" es la variable que indica la masa de aire al inicio de la observación. En este caso, el árbol solo hace 4 predicciones distintas que están representadas por la etiqueta "value" en los nodos hoja.

la profundidad mejora el ajuste. Sin embargo, este algoritmo es muy propenso a sobreajustar si no se controla el tamaño del árbol ya que tiene la flexibilidad necesaria para ajustarse a cualquier curva.

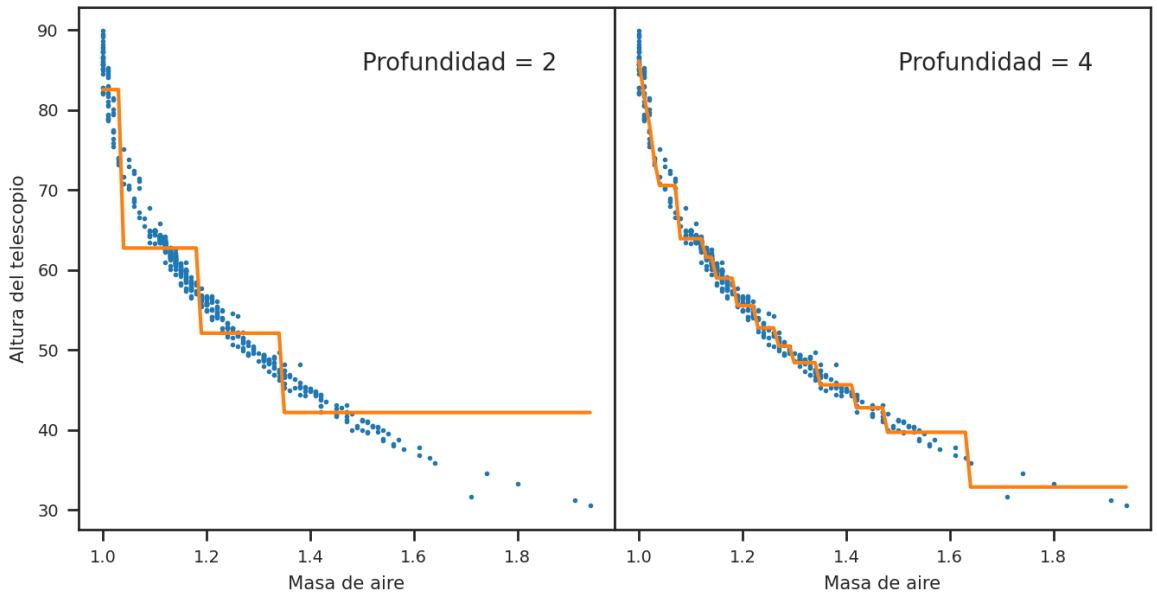

Figura B.2. Altura del telescopio (en grados) en función de la masa de aire. Los puntos azules son los datos y la línea naranja el ajuste del árbol de decisión.

Apéndice C

Gráficos de esquina

C.1. Distribuciones de las 5 variables predictoras de mayor importancia para las estrellas de testeo y entrenamiento.

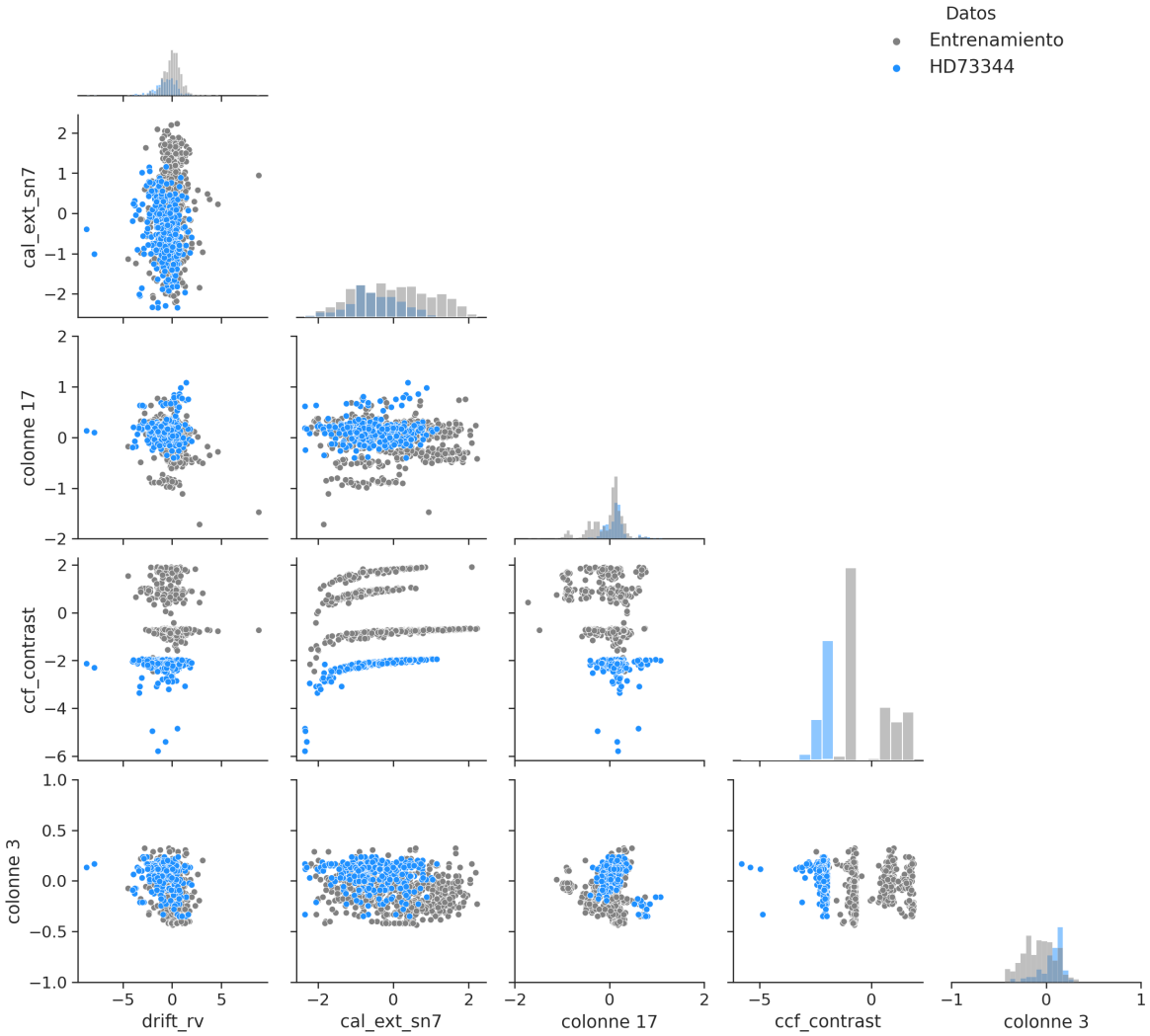

Figura C.1. Variables más importantes para el modelo de Gradient Boosting. En gris se representan los datos de entrenamiento y en celeste de la estrella HD 73344.

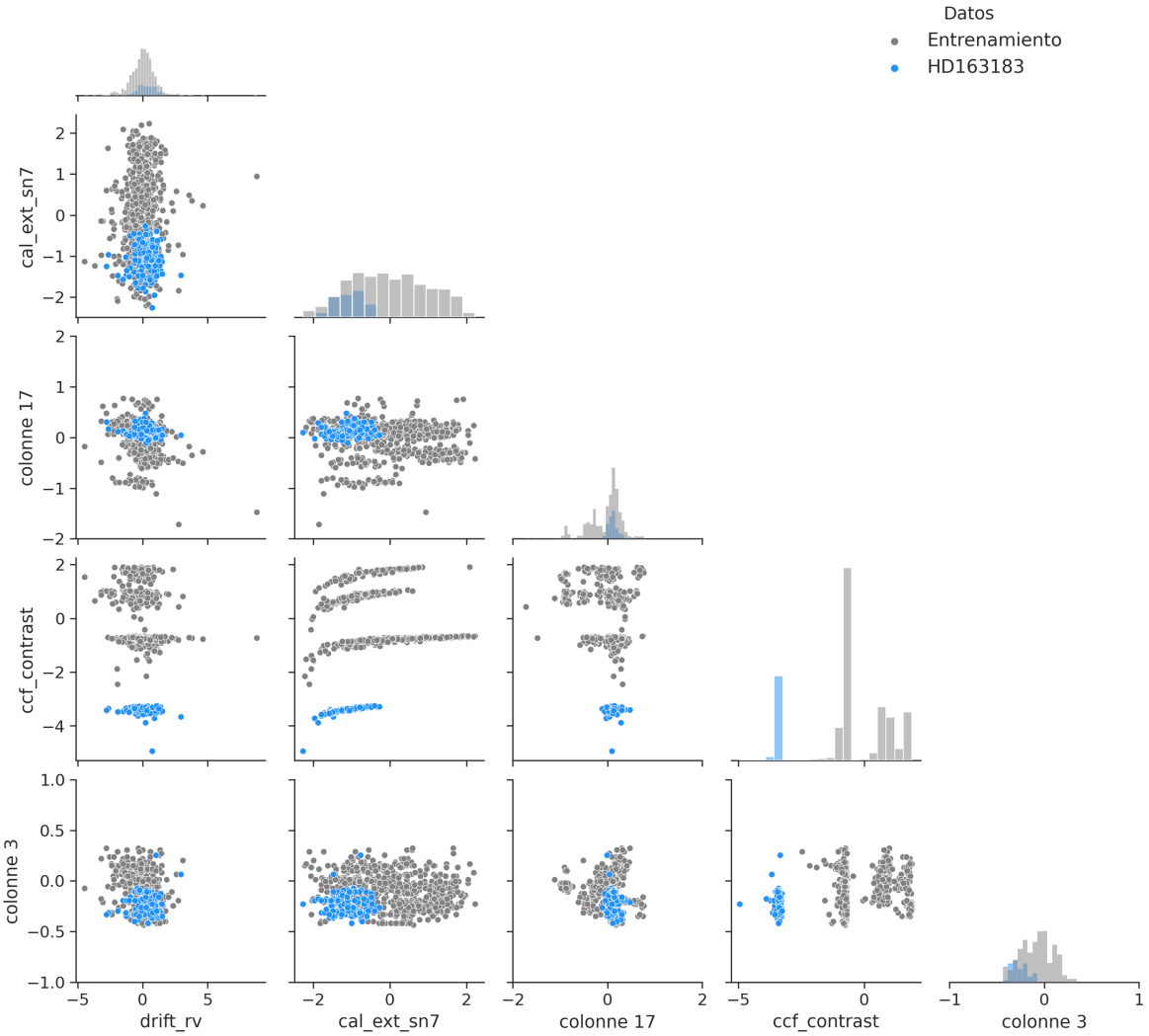

Figura C.2. Variables más importantes para el modelo de Gradient Boosting. En gris se representan los datos de entrenamiento y en celeste de la estrella HD 161183.

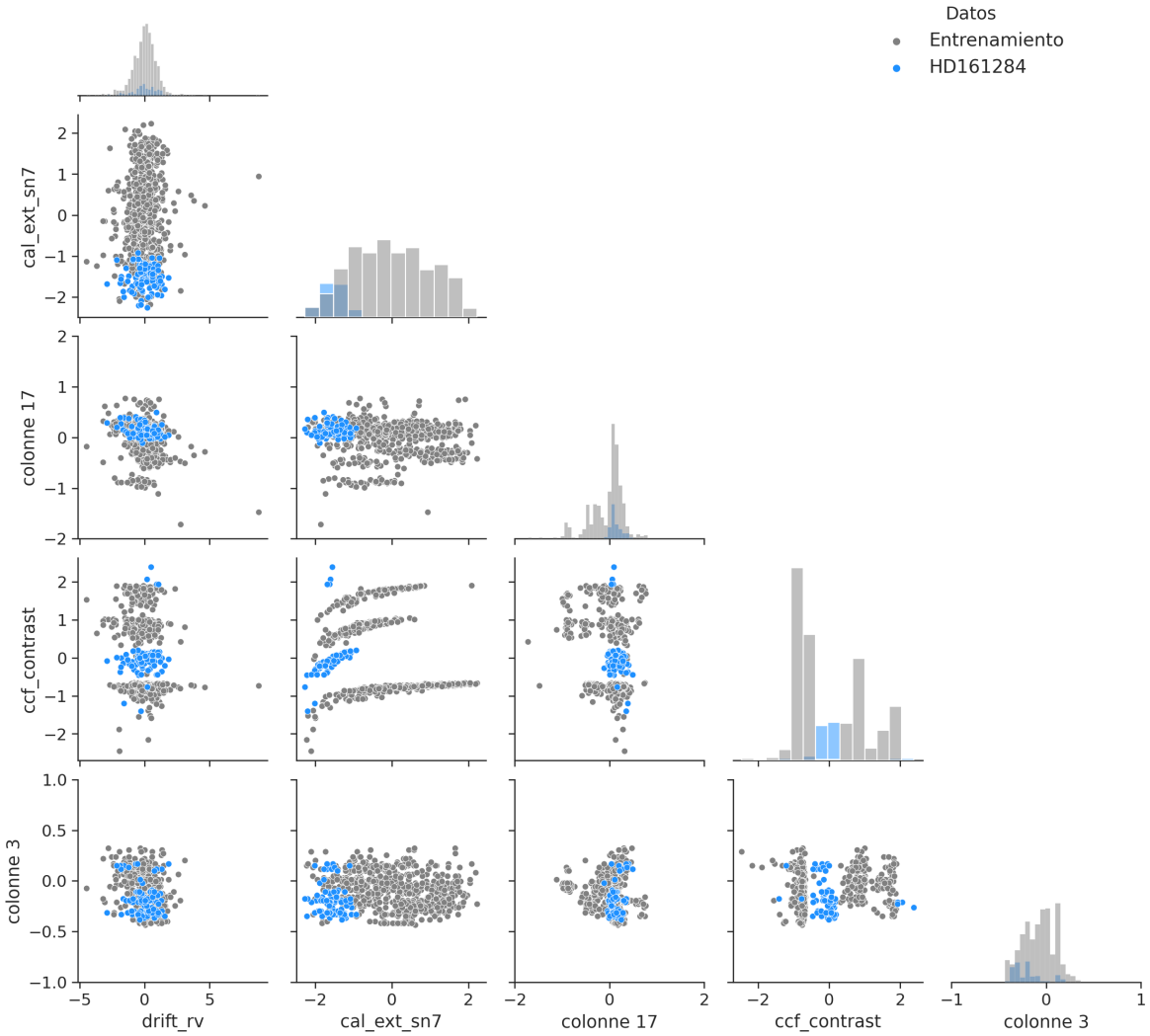

Figura C.3. Variables más importantes para el modelo de Gradient Boosting. En gris se representan los datos de entrenamiento y en celeste de la estrella HD 161284.

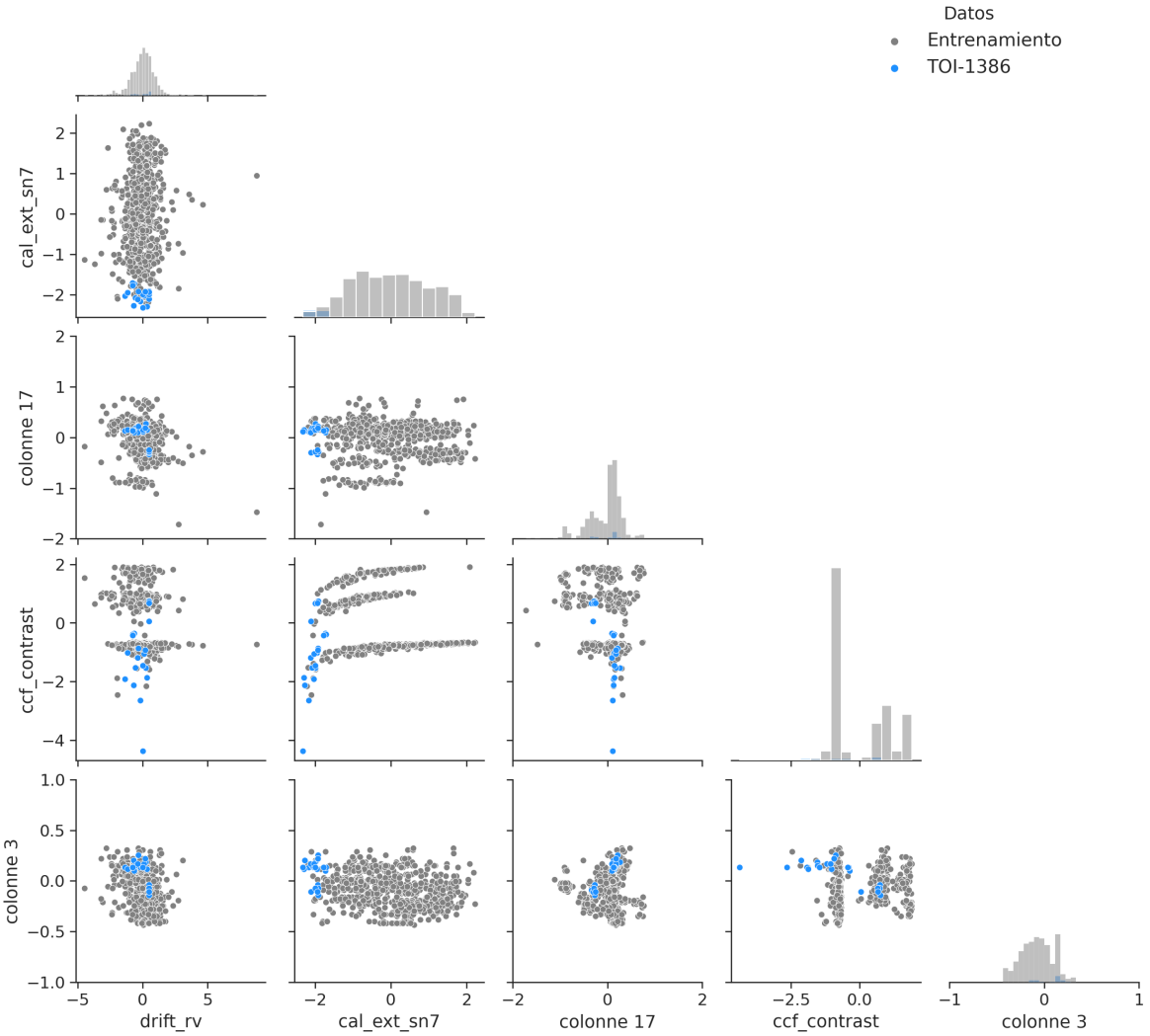

Figura C.4. Variables más importantes para el modelo de Gradient Boosting. En gris se representan los datos de entrenamiento y en celeste de la estrella TOI-1386.

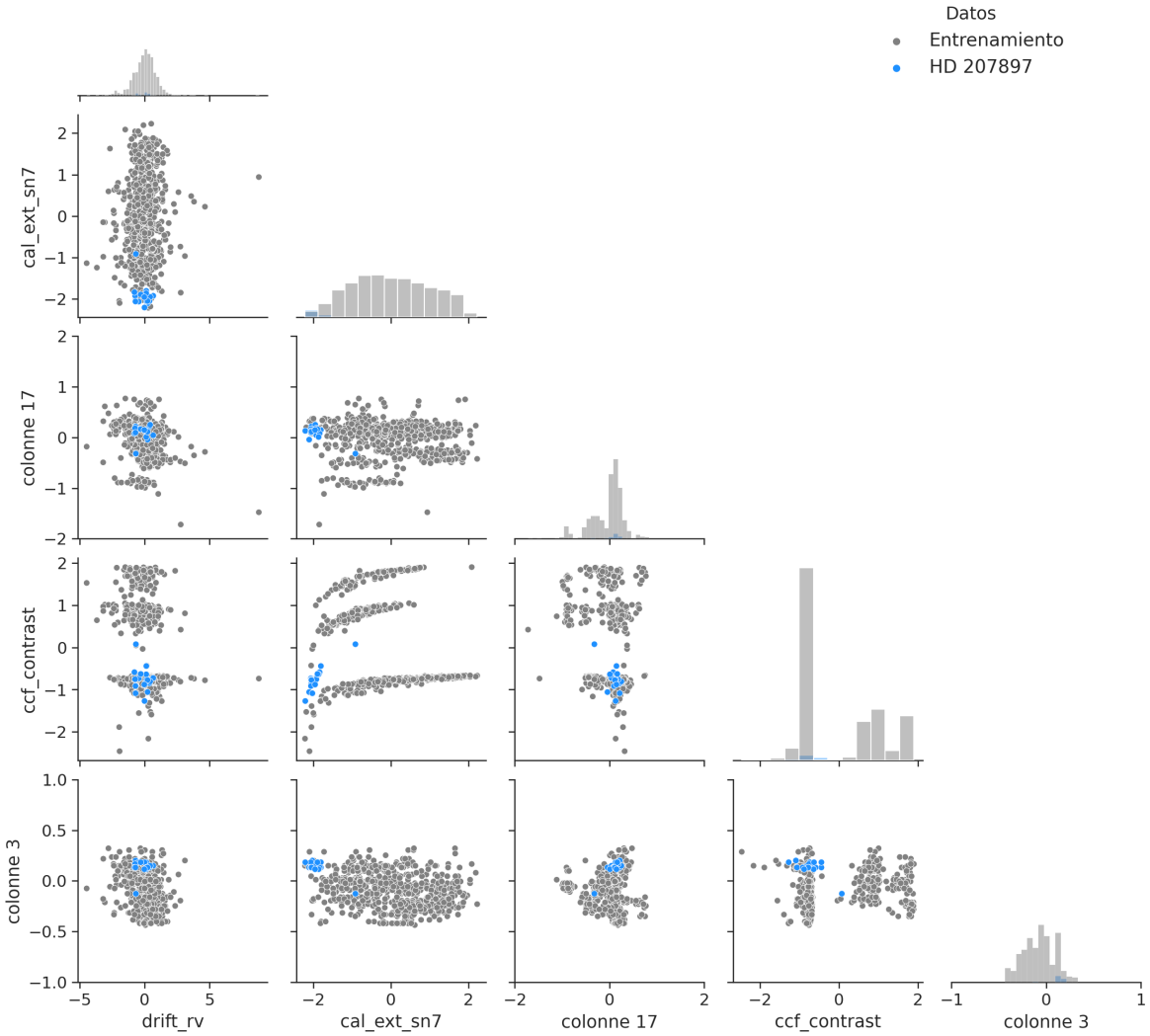

Figura C.5. Variables más importantes para el modelo de Gradient Boosting. En gris se representan los datos de entrenamiento y en celeste de la estrella HD 207897.

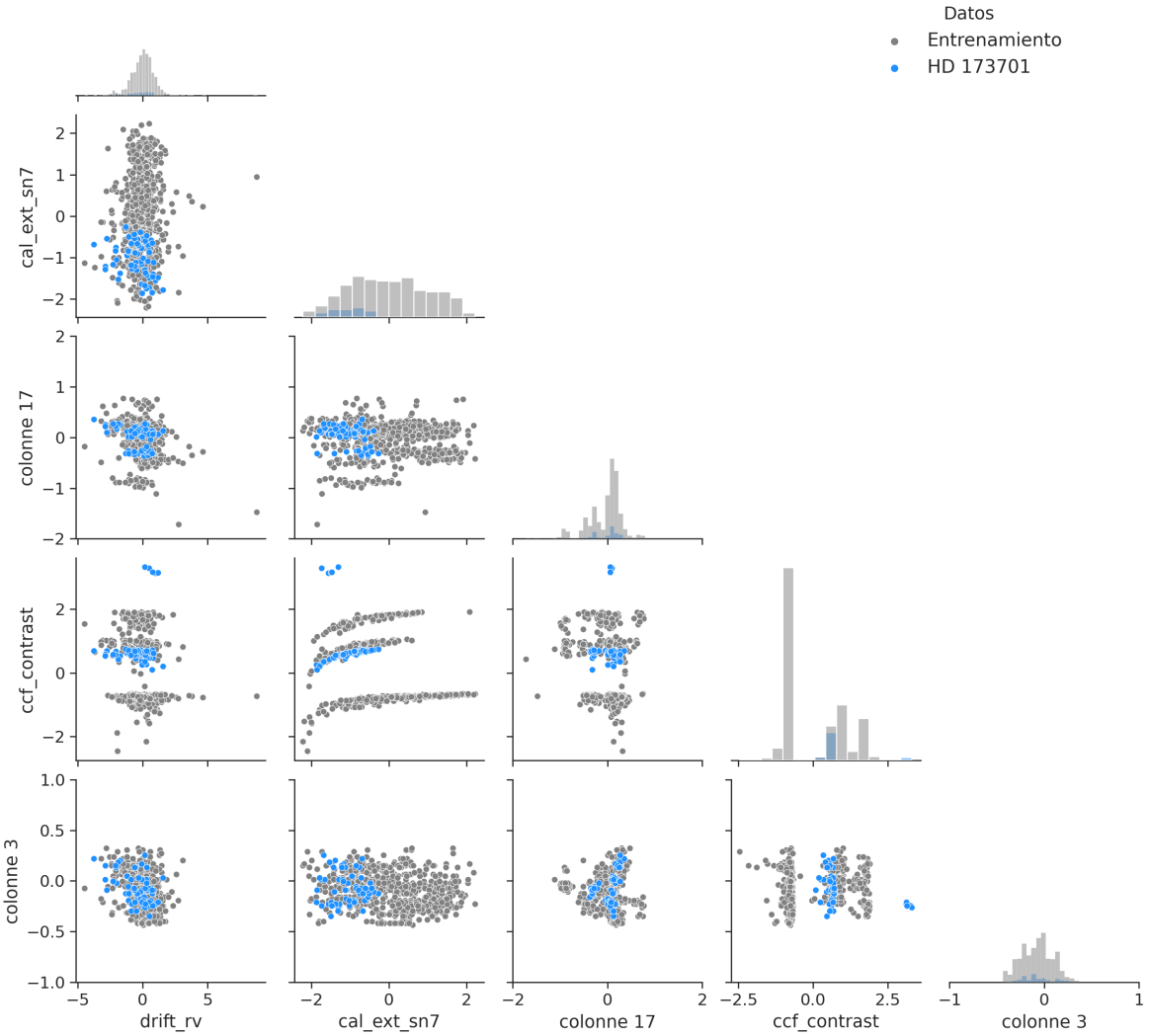

Figura C.6. Variables más importantes para el modelo de Gradient Boosting. En gris se representan los datos de entrenamiento y en celeste de la estrella HD 173701.

Bibliografía

Baranne, A., Mayor, M., & Poncet, J. L. 1979, Vistas in Astronomy, 23, 279

Baranne, A., Queloz, D., Mayor, M., et al. 1996, A&AS, 119, 373

Barge, P., Baglin, A., Auvergne, M., et al. 2008, A&A, 482, L17

Beaulieu, J. P., Bennett, D. P., Fouqué, P., et al. 2006, Nature, 439, 437

Bishop, C. 2006, Pattern Recognition and Machine Learning, Information Science and Statistics (Springer)

Boisse, I., Bouchy, F., Hébrard, G., et al. 2011, A&A, 528, A4

Boisse, I., Eggenberger, A., Santos, N. C., et al. 2010, A&A, 523, A88

Boisse, I., Pepe, F., Perrier, C., et al. 2012, A&A, 545, A55

Boss, A. P. 1997, Science, 276, 1836

Bouchy, F., Hébrard, G., Udry, S., et al. 2009, A&A, 505, 853

Bouchy, F., Pepe, F., & Queloz, D. 2001, A&A, 374, 733

Bouchy, F., Ségransan, D., Díaz, R. F., et al. 2016, A&A, 585, A46

Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, ApJ, 529, L45

Chauvin, G., Lagrange, A. M., Dumas, C., et al. 2004, A&A, 425, L29

Chontos, A., Murphy, J. M. A., MacDougall, M. G., et al. 2022, AJ, 163, 297

Christiansen, J. L., Bhure, S., Zink, J. K., et al. 2022, AJ, 163, 244

Collier Cameron, A., Bouchy, F., Hébrard, G., et al. 2007, MNRAS, 375, 951

Courcol, B., Bouchy, F., Pepe, F., et al. 2015, A&A, 581, A38

Crass, J., Gaudi, B. S., Leifer, S., et al. 2021, arXiv e-prints, arXiv:2107.14291

Cumming, A., Butler, R. P., Marcy, G. W., et al. 2008, PASP, 120, 531

Dalal, S., Kiefer, F., Hébrard, G., et al. 2021, A&A, 651, A11

de Leon, J. P., Livingston, J., Endl, M., et al. 2021, MNRAS, 508, 195

Demangeon, O. D. S., Dalal, S., Hébrard, G., et al. 2021, A&A, 653, A78

Díaz, R. F. 2018, in Astrophysics and Space Science Proceedings, Vol. 49, Asteroseismology and Exoplanets: Listening to the Stars and Searching for New Worlds, ed. T. L. Campante, N. C. Santos, & M. J. P. F. G. Monteiro, 199

Díaz, R. F., Delfosse, X., Hobson, M. J., et al. 2019, A&A, 625, A17

Díaz, R. F., Rey, J., Demangeon, O., et al. 2016a, A&A, 591, A146

Díaz, R. F., Santerne, A., Sahlmann, J., et al. 2012, A&A, 538, A113

Díaz, R. F., Ségransan, D., Udry, S., et al. 2016b, A&A, 585, A134

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. 2004, The Annals of Statistics, 32, 407

Figueira, P. 2018, in Astrophysics and Space Science Proceedings, Vol. 49, Asteroseismology and Exoplanets: Listening to the Stars and Searching for New Worlds, ed. T. L. Campante, N. C. Santos, & M. J. P. F. G. Monteiro, 181

Géron, A. 2017, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (O'Reilly Media)

Hara, N. C., Bouchy, F., Stalport, M., et al. 2020, A&A, 636, L6

Hébrard, G., Arnold, L., Forveille, T., et al. 2016, A&A, 588, A145

Hébrard, G., Bonfils, X., Ségransan, D., et al. 2010, A&A, 513, A69

Heidari, N., Boisse, I., Orell-Miquel, J., et al. 2022, A&A, 658, A176

Hobson, M. J., Delfosse, X., Astudillo-Defru, N., et al. 2019, A&A, 625, A18

Hobson, M. J., Díaz, R. F., Delfosse, X., et al. 2018, A&A, 618, A103

Hortensius, M. & Gassendi, P. 1633, Martini Hortensi ... Dissertatio de Mercurio in sole viso et Venere invisa : instituta cum ... D.Pedro Gassendo

Koch, D. G., Borucki, W. J., Basri, G., et al. 2010, ApJ, 713, L79

Krist, J., Nemati, B., & Mennesson, B. 2016, Journal of Astronomical Telescopes, Instruments, and Systems, 2, 011003

Macintosh, B., Graham, J. R., Barman, T., et al. 2015, Science, 350, 64

Mayor, M., Marmier, M., Lovis, C., et al. 2011, arXiv e-prints, arXiv:1109.2497

Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20

Mayor, M. & Queloz, D. 1995, Nature, 378, 355

Mazeh, T., Holczer, T., & Faigler, S. 2016, A&A, 589, A75

Miguel, Y., Guilera, O. M., & Brunini, A. 2011, MNRAS, 417, 314

Mistry, P., Pathak, K., Prasad, A., et al. 2023, VaTEST II: Statistical Validation of 16 Exoplanets of TESS

Moutou, C., Hébrard, G., Bouchy, F., et al. 2014, A&A, 563, A22

Müller, A. & Guido, S. 2016, Introduction to Machine Learning with Python: A Guide for Data Scientists (O'Reilly Media)

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

Penny, M. T., Gaudi, B. S., Kerins, E., et al. 2019, ApJS, 241, 3

Pepe, F., Cristiani, S., Rebolo, R., et al. 2021, A&A, 645, A96

Perruchot, S., Bouchy, F., Chazelas, B., et al. 2011, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8151, Techniques and Instrumentation for Detection of Exoplanets V, ed. S. Shaklan, 815115

Perruchot, S., Kohler, D., Bouchy, F., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7014, Ground-based and Airborne Instrumentation for Astronomy II, ed. I. S. McLean & M. M. Casali, 70140J

Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, , 124, 62

Pueyo, L. 2018, in Handbook of Exoplanets, ed. H. J. Deeg & J. A. Belmonte, 10

Queloz, D., Henry, G. W., Sivan, J. P., et al. 2001, A&A, 379, 279

Rey, J., Hébrard, G., Bouchy, F., et al. 2017, A&A, 601, A9

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Santerne, A., Díaz, R. F., Bouchy, F., et al. 2011, A&A, 528, A63

Santos, N. C., Mayor, M., Bonfils, X., et al. 2011, A&A, 526, A112

Serrano Bell, J. & Díaz, R. F. 2022, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 63, 45

Serrano Bell, J., Hebrard, G., & Díaz, R. F. in prep

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757

Valenti, J. A. & Fischer, D. A. 2005, ApJS, 159, 141

Wang, J. & Fischer, D. A. 2015, AJ, 149, 14

Wilson, P. A., Hébrard, G., Santos, N. C., et al. 2016, A&A, 588, A144

Winn, J. N. 2018, in Handbook of Exoplanets, ed. H. J. Deeg & J. A. Belmonte, 195

Wolszczan, A. 1994, Science, 264, 538

Wolszczan, A. & Frail, D. A. 1992, Nature, 355, 145